141 Topics in Current Chemistry

Chemometrics and Species Identification

With Contributions by
C. Armanino, M. Forina, P. H. E. Gardiner,
E. J. v. d. Heuvel, G. Kateman, S. Lanteri,
H. C. Smit, B. G. M. Vandeginste

With 84 Figures and 19 Tables

Springer-Verlag Berlin Heidelberg NewYork
London Paris Tokyo

This series presents critical reviews of the present position and future trends in modern chemical research. It is addressed to all research and industrial chemists who wish to keep abreast of advances in their subject.

As a rule, contributions are specially commissioned. The editors and publishers will, however, always be pleased to receive suggestions and supplementary information. Papers are accepted for "Topics in Current Chemistry" in English.

ISBN 3-540-17308-0 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-17308-0 Springer-Verlag New York Heidelberg Berlin

Library of Congress Cataloging-in-Publication Data
Chemometrics and species identification.
(Topics in current chemistry; 141)
Contents: Chemometrics/B. G. M. Vandeginste —
Chemometrics: sampling strategies/G. Kateman — Signal and data analysis in chromatography/H. C. Smit, E. J. v. d. Heuvel — [etc.]
1. Chemistry — Mathematics. 2. Chemistry — Statistical methods.
I. Armanino, C. II. Series.
QD1.F58 Vol. 141 540 s 86-31648
[QD39.3.M3] [543'.00151]
ISBN 0-387-17308-0 (U.S.)

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically those of translation, reprinting, re-use of illustrations, broadcasting, reproduction by photocopying machine or similar means, and storage in data banks. Under § 54 of the German Copyright Law where copies are made for other than private use, a fee is payable to "Verwertungsgesellschaft Wort", Munich.

© Springer-Verlag Berlin Heidelberg 1987
Printed in GDR

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting and Offsetprinting: Th. Müntzer, GDR;
Bookbinding: Lüderitz & Bauer, Berlin
2152/3020-543210

Editorial Board

Prof. Dr. *Michael J. S. Dewar*	Department of Chemistry, The University of Texas Austin, TX 78712, USA
Prof. Dr. *Jack D. Dunitz*	Laboratorium für Organische Chemie der Eidgenössischen Hochschule Universitätsstraße 6/8, CH-8006 Zürich
Prof. Dr. *Klaus Hafner*	Institut für Organische Chemie der TH Petersenstraße 15, D-6100 Darmstadt
Prof. Dr. *Edgar Heilbronner*	Physikalisch-Chemisches Institut der Universität Klingelbergstraße 80, CH-4000 Basel
Prof. Dr. *Śhô Itô*	Department of Chemistry, Tohoku University, Sendai, Japan 980
Prof. Dr. *Jean-Marie Lehn*	Institut de Chimie, Université de Strasbourg, 1, rue Blaise Pascal, B. P. Z 296/R8, F-67008 Strasbourg-Cedex
Prof. Dr. *Kurt Niedenzu*	University of Kentucky, College of Arts and Sciences Department of Chemistry, Lexington, KY 40506, USA
Prof. Dr. *Kenneth N. Raymond*	Department of Chemistry, University of California, Berkeley, California 94720, USA
Prof. Dr. *Charles W. Rees*	Hofmann Professor of Organic Chemistry, Department of Chemistry, Imperial College of Science and Technology, South Kensington, London SW7 2AY, England
Prof. Dr. *Fritz Vögtle*	Institut für Organische Chemie und Biochemie der Universität, Gerhard-Domagk-Str. 1, D-5300 Bonn 1
Prof. Dr. *Georg Wittig*	Institut für Organische Chemie der Universität Im Neuenheimer Feld 270, D-6900 Heidelberg 1

Table of Contents

Chemometrics — General Introduction and Historical Development

Bernard G. M. Vandeginste

Laboratory for Analytical Chemistry, Faculty of Sciences,
Catholic University of Nijmegen, Toernooiveld
6525 ED Nijmegen, The Netherlands

Table of Contents

Analytical chemists deal with a wide range of decision making problems such as: the selection and design of an analytical method, and processing and interpretation of the measured data. For this purpose, formal strategies and algorithms based on statistical and mathematical techniques are needed. The development of such strategies is the area of research of Chemometrics, a discipline of analytical chemistry. In this paper the role of chemometrics in the analytical process is discussed and a historical survey is given of the development of Chemometrics during the past two decades. A selection of standard Chemometric tools available to the analytical chemist is discussed in some more detail: multivariate optimization, data processing and calibration. The paper is closed with a few remarks on future directions of Chemometrics.

Topics in Current Chemistry, Vol. 141
© Springer-Verlag, Berlin Heidelberg 1987

Bernard G. M. Vandeginste

1 The Role of Chemometrics in the Analytical Process

1.1 Economical Aspects of Analytical Information

In its narrow sense, chemical analysis is an activity of obtaining information on the identity or on the quantitative composition of a sample. By chemical analysis an analytical result is produced, which may be one or more numbers, or one or more compound names. Why do analysts, or in general analytical laboratories, produce these numbers and names? This question has been addressed by several analytical chemists [1,2,3].

The proposed answers vary from "because everyone does" to "because we think that the analytical results contain relevant information for the customer who asked for the analysis". Another often mentioned reason is simply "because the customer asked for it". As Massart [1] pointed out, it is to be hoped that your own answer is not this last one, but is instead "because we think the value of the information present in the analytical result is more worth than the cost of obtaining it". This means that analytical information has an economical value. This fact confronts us with three problems, namely: how can we quantify the amount of information, or the quality of information present in the analytical data? What are the cost of chemical analysis? How to quantify the economical value of analytical information?

Intuitively we can feel that the economical value of the analytical result is related to its quality. The quality of an analytical result depends upon two factors: first of all we should know how confident we are about the produced result. In fact, an analytical result without an explicit or implicit (by the number of significant figures) indication of its precision has no quality at all. Second, the quality of the analytical result depends on how well the sample represents the system of its origin. The sample may be contaminated or may be modified because of inappropriate storage and aging. In other instances, when the sample is taken from a chemical reactor in which a chemical reaction is occurring, the constitution of the reactor content is usually time varying. Because of inevitable time delays in the analytical laboratory, the constitution of the sample will not anymore represent the actual constitution in the reactor at the moment when the analytical result is available. Therefore, both the precision of the analytical method and the analysis time are important indicators for the quality of an analytical result [4].

This requirement of being able to attach a quality label to our analytical results, made that statistics and the statistical treatment of our data have become of a tremendous importance to us. This is reflected by the fact that in 1972 ANALYTICAL CHEMISTRY started with the publication of a section on "Statistical and Mathematical Methods in Analytical Chemistry" [5,6] in its bi-annual reviews. Although we feel us quite confident on how to express our uncertainty (or certainty) in the produced numbers, we are less sure on how to quantify our uncertainty in produced compound names or qualitative results.

The economical value of the analytical result depends upon the amount of information the customer actually receives, and upon whether the customer indeed uses that information. The amount of received information can be defined as the difference between the initial uncertainty (H_0) of the customer before receiving the analytical result(s) and the remainder uncertainty (H_1) after having received the result(s). The

net yield of information is thus: $\Delta H = H_0 - H_1$. If we apply this definition to the case of process control, then H_0 is related to the variance of the uncontrolled process and H_1 is related to the variance of the controlled process. When considering the cost-effectiveness of analytical effort, we should therefore, weight the cost (C) of producing analytical information (H_1), against the return or profit (P), earned by applying the net amount of received information (ΔH) by the customer. Decision making in the analytical laboratory is, therefore, focussed on maximizing the net profit (P—C). This obliges the manager of the analytical laboratory to keep evaluating the analytical methods and equipment in use in the laboratory, in relation to the changing demands for information by his customers and to the new technologies introduced on the market place. Todays equipment is of an increasing sophistication, with capabilities to determine more analytes in a shorter time and with better precision and contains software to treat the complex data structures it generates.

Two examples are given below, which demonstrate the economical principles mentioned above.

The first example is from Leemans [4] and applies to process control. When monitoring a process, the uncertainty about the actual state of an uncontrolled process is the variance of the parameter of interest: s_0^2. From information theory [7] it follows that the initial information (I_0) available on the process is inversely proportional to the uncertainty, namely: $I_0 = \log_2 1/s_0^2$ ($\log_2$ is the logarithm on a base 2).

Equally, the uncertainty about the process value after analysis is the variance of

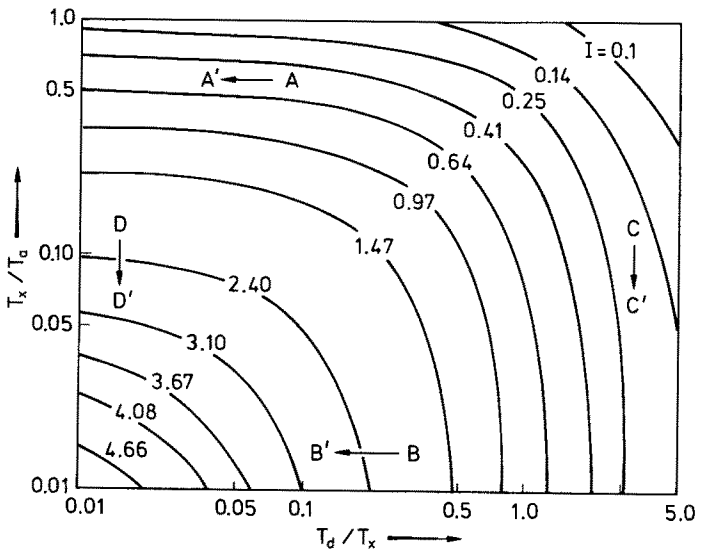


Fig. 1. The net yield of information (I), in bits, obtained by analysis with an analytical method with an analysis time, Td, and analysis frequency, 1/Ta, for the control of a process with a time constant, Tx. From G. Kateman and F. W. Pijpers, "Quality control in analytical chemistry" p. 87 (1981). Copyright © 1981, John Wiley & Sons Inc. New York. Adapted and reproduced by permission of Wiley & Sons, Inc., New York

the controlled process: s_1^2. The information after analysis is therefore, $I_1 = \log_2 1/s_1^2$. The net yield of information $\Delta I = I_1 - I_0 = \log_2 s_0^2/s_1^2$.

Leemans [4] and Müskens [8] derived a relationship between the net yield of information and the quality of the analytical procedure, expressed in terms of analysis time (T_d), precision (σ_a) and the sampling frequency ($1/T_a$). Their results are graphically displayed in Fig. 1. The conclusion on one hand is predictable but is on the other hand very striking. Equal investments in the laboratory may have different effects on the net yield of information, depending on the particular situation, e.g. an increase of the workload by a factor of 2 may have a very minor effect (point C → C′ in Fig. 1), or may have a very pronounced effect (point D → D′ in Fig. 1). The diagram also shows that a replacement of a method by a twice as fast one (e.g. as a result of optimization) may have no effect at all (point A → A′ in Fig. 1) or may have a significant effect (point B → B′ in Fig. 1). This proves that equal marginal cost may yield different marginal returns.

The second example is from Massart [1], who derived a relationship between the quality of an analytical result and its utility for medical diagnosis. As an indicator for the utility of a clinical test, one can use its specificity. This is the percentage of "normal" patients which are recognized as such. The less analytical errors are made, the better the specificity will be, which is shown in table 1. This table demonstrates

Table 1. Percentage (u) of "normal" patients recognized as such and amount of produced analytical information (I). Data from Acland and Lipton [9] and adapted by Massart [1]. Reprinted by permission of Elsevier Science Publishers, Amsterdam

S_A/S_N	u	$I = \log_2 (S_N/S_A)$
0.1	99	3.32
0.2	99	2.32
0.3	98	1.73
0.4	97	1.32
0.5	95	1.00
0.6	94	0.73
0.8	90	0.32
1.0	86	0

that the law of marginal utility applies in analytical chemistry. For the same amount of extra information, the obtained marginal utility decreases. Both examples demonstrate that although decision making in the analytical laboratory is very complex, it could be made easier when formal knowledge is available on basic relationships between the amount of generated information and the characteristics of the analytical method. In most of the cases, these relationships are expressed in mathematical or statistical formulas or models. It is, therefore, necessary to try to formalize the different parts of the analytical process.

1.2 Stages in the Analytical Process

Despite the large amount of different analytical procedures, a regularily returning pattern of activities can be discovered in chemical analysis. This pattern of activities can be considered to define the analytical process. In the previous section it was explained that the analytical system, which consists of a sample input and result output (Fig. 2) represents only a part of the total analytical process. The sample is the result of several actions after the receipt of the customers' chemical problem. The obtained analytical result, however, still needs to be converted into information. Therefore, the analytical process is better described as a cycle, which begins with the formulation of the problem that needs a solution by chemical analysis and is finished after

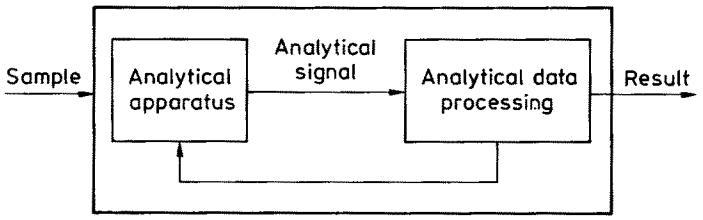


Fig. 2. The analytical method. From K. Eckschlager, V. Stepanek, Anal. Chem. *54*, 1115A (1982). Reproduced by permission of the American Chemical Society, Washington DC.

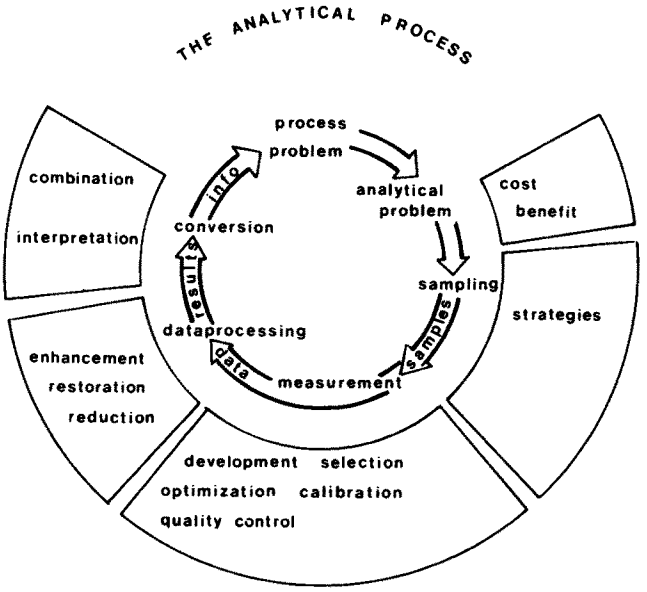


Fig. 3. The analytical process. From B. G. M. Vandeginste, Anal. Chim. Acta *150*, 201 (1983). Reproduced by permission of Elsevier Science Publishers, Amsterdam.

the analytical information has been actually supplied to solve that problem (Fig. 3). By a number of consecutive actions (Fig. 3) the initially formulated problem is translated into an analytical problem, (a) sample(s), rough data, refined data, analytical results, information and finally knowledge. The activities necessary for these conversions are: method selection, sampling, sample preparation and measurement, data processing and finally results processing. These steps define the principal stages of the analytical process. Associated with these steps are the decisions we have to make in order to obtain a maximal net profit. Without aiming to be exhaustive, a number of these decisions are listed below:

- method selection:
  make a list of candidate methods, evaluate cost, precision, analysis time, expected workload in view of required sampling scheme.
- sampling:
  derive the best sampling strategy.
- measurement:
  tune the initial measurement conditions for optimal performance (e.g. resolution in chromatography).
  select a proper calibration method: univariate or multivariate.
  design a system for quality control
- data processing:
  enhance the data if necessary; select the proper filter or smoothing technique
  restore the data if necessary; select the proper method for deconvolution.
  reduce the data to concentrations; select the proper univariate or multivariate method result processing:
  combine, classify and interpret the results — select the proper multivariate method.

Because the analytical process is a cycle or chain, each link or operation defines the ultimate quality of the analytical information. The effect of a poor sampling strategy will be very difficult to be compensated by a very good calibration method and vice versa. It is, therefore, the uneasy task of the analytical chemist to make the right or best decision at every stage of the analytical process. A large part of the decision process was believed being impossible to be formalized. Many have put up with the apparent fact that a successful analyst has an inexplicable sense of the right decision. This would reduce analytical chemistry to an art, which is not. It is likely that the above mentioned decisions cannot be made without the support of applied mathematics and statistics. Our possibilities to apply these techniques depend strongly on the availability of modern computer technology and on the imagination of the analytical chemist to follow closely on heels the advances in computer science, mathematics and statistics. The necessity to apply these techniques becomes the more and more urgent when analytical equipment produces the more complex data. A typical example is a new class of analytical methods, which consists of two linked methods such as gas chromatography-mass spectrometry.

In the present time with almost unlimited computer facilities in the analytical laboratory, analytical chemists should be able to obtain substantial benefits from the application of time series, information theory, multivariate statistics, a.o. factor analysis and pattern recognition, operations research, numerical analysis, linear algebra, computer science, artificial intelligence, etc. This is in fact what chemometricians have been doing for the past decades.

1.3 Chemometrics in the Analytical Process

Chemometrics is not a mathematical discipline and should not be confounded with one or more of the disciplines from mathematics. It is, however, a chemical discipline, as is schematically shown in Fig. 4. The inner circle represents the chemical analysis. The decisions mentioned in previous sections are supported by chemometric tools. Chemical analysis together with the chemometric tools belong to the area of analytical chemistry, which is schematically represented by the outer circle. The mathematical techniques surrounding this outer circle are auxilary techniques to the analytical chemist. In this picture, Chemometrics is the interface between chemistry and mathematics. As Kowalski [10] clearly stated "Chemometric tools are vehicles that can aid chemists to move more efficiently on the path from measurements to information to knowledge". This brings us to the formal definition of chemometrics [11] "Chemometrics is the chemical discipline that uses mathematical and statistical methods (a) to design or select optimal mesurement procedures and experiments and (b) to provide maximum chemical information by analyzing chemical data. In the field of Analytical Chemistry, Chemometrics is the chemical discipline that uses mathematical and statistical methods for the obtention in the optimal way of relevant information on material systems".

Key words in the definition are "optimal" and "material systems". These express the fact that chemical analysis is related to a problem and not to "a sample" and that economical aspects of chemical analysis prevail. The result of chemometric research is chemometric software, which enables a large scale implementation and application of chemometric tools in practical chemical analysis.

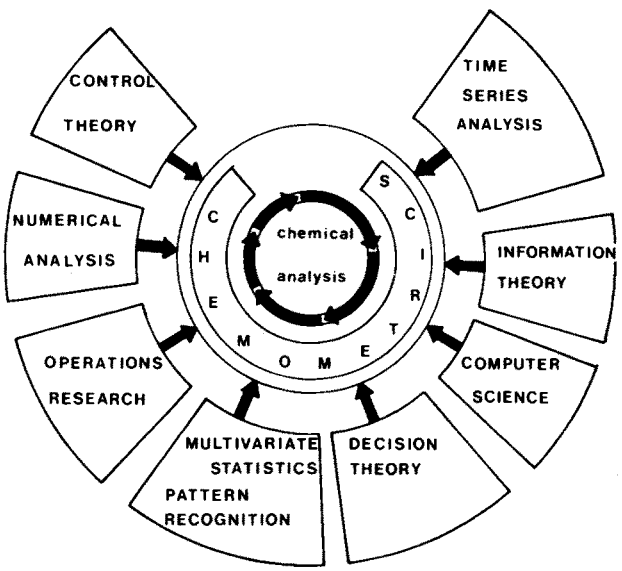


Fig. 4. Chemometrics, the interface between the analysis and mathematics. From B. G. M. Vandeginste, Anal. Chim. Acta *150*, 203 (1983). Reproduced by permission of Elsevier Science Publishers, Amsterdam.

2 Chemometrics in a Historical Context

It happens that this paper is published a year after the 10-th anniversary of the Chemometric Society, which was founded by Wold en Kowalski, who also coined at the time the word Chemometrics as a title for a new subdivision of analytical chemistry using the methods described in previous sections.

The current interest of analytical chemists may be read from the Chemometric Society Newsletter distribution list [12] (table 2).

Table 2. Chemometrics Society newsletter distribution list 1985

Country	Number	Country	Number
Australia	5	Indonesia	1
Austria	13	Iraq	1
Belgium	15	Italy	49
Brazil	3	Japan	8
Canada	12	Kenya	1
China	1	Netherlands	59
Czechoslovakia	7	Norway	29
Denmark	8	Portugal	5
Finland	5	Poland	6
France	17	Romania	4
Germany (East)	2	South Africa	1
Germany (West)	28	Spain	15
Great Britain	50	Sweden	9
Greece	2	Switzerland	4
Hungary	7	Turkey	4
India	2	U.S.S.R.	1
Yugoslavia	5	United States	262
Iceland	1		

Apparent centers with major interest in Chemometrics are the USA and the Netherlands in Europe, with respectively 262 and 59 members. Therefore, it is interesting to report on a survey made by Tuinstra et al. [13] of 300 Dutch analytical laboratories on their knowledge of, and familiarity with modern Chemometric optimization strategies. These optimization strategies require a relatively low level of abstraction and mathematics and can be expected to be fairly well known: 36% have heard or read about these strategies. Only 5% know and use optimization techniques and another 6% of the respondents have seriously considered using optimization techniques but finally decided against it. There was no difference observed between the private sector, clinical laboratories and governmental laboratories. In every case but one there was at least one university graduate present. The investigators of this survey found these figures disappointingly low. One should, however, take into account the facts that there is no real textbook available on Chemometrics and there is a considerable time lag between research and education. Another point which hindered a large scale application is the availability of certified and thoroughly tested Chemometric software at the time of this survey.

Two major developments in the past decade increased the impact of Chemometrics: the development of computers and micro-electronics and the advancement of analytical instrumentation.

In a lecture given to the Analytical Division of the Royal Chemical Society, Betteridge [14] summarized the impact of these developments on analytical chemistry as providing solutions to barriers to analytical information, giving rise to new problems (table 3). I will follow his lines to discuss the evolution of Chemometrics in its historical context.

Table 3. The generation of analytical information. From J. Betteridge, lecture presented at the Symposium "The integrated approach to laboratory automation", RCS, Analytical Division, Dorset, October 1985

	Barrier to analytical information	Solution	New problem
pre 1960	Data generation (burette)	Electronic control	Miles of chart paper
1960's	Data acquisition	Digitisers	Masses of data and results
1970's	Data/result reduction	Mini/microcomputers	Masses of information
1980's	Information management	Workstations, LIMS	Complexity of decisions
1990's	Intelligence	AI, expert systems	Fundamentals of analytical chemistry

The pre-sixties was a period just before the second phase of the electronics revolution that took place in 1960. During and before the fifties, most of our analytical equipment had to be controlled manually, making the data collection slow and laborous. The measurement of a spectrum with a single beam spectrometer, for instance, had to be carried out point by point, with in between a manual adjustment of the monochromator and baseline. The principal barrier to the production of analytical information was, therefore, the data generation. It increased the desire for having recording instruments. The first spectrometers with automatic wavelength scanning and baseline correction became widespread available in the fifties. As a consequence data generation became relatively easy and fast, causing, however, the production of miles of chart paper. The lack of the possibility to transform recorder traces into useful analytical information became the next barrier to the production of analytical information.

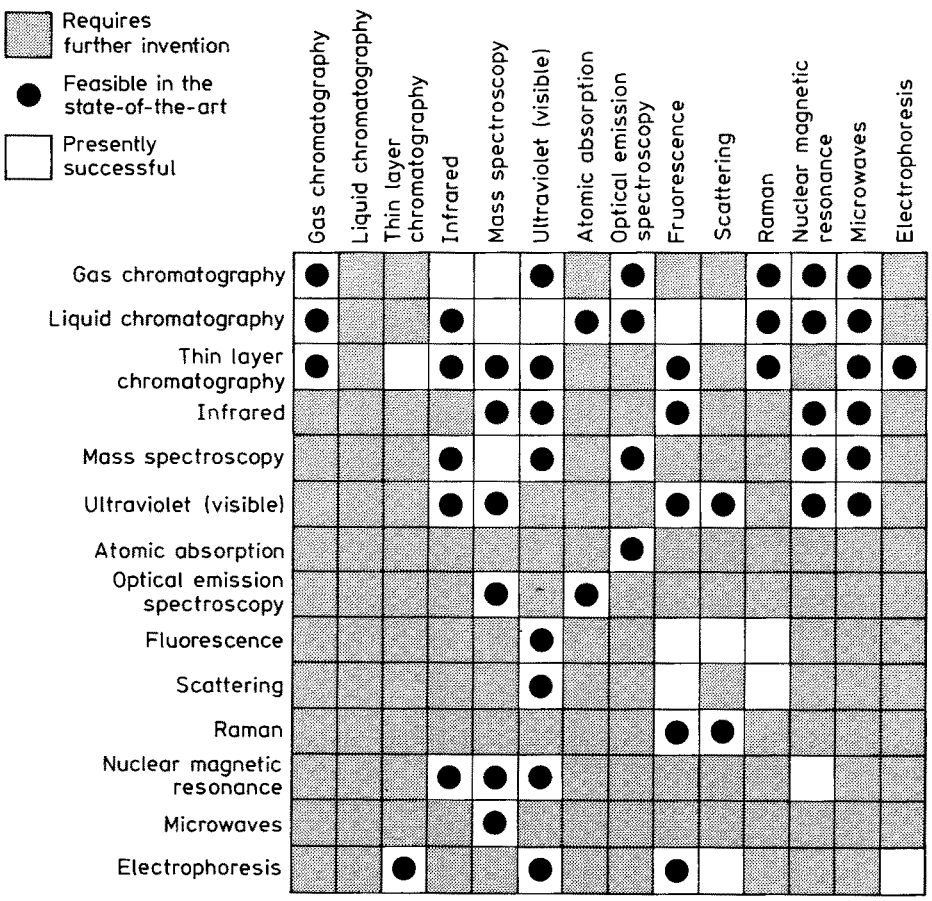
In the sixties semiconductor devices were introduced, changing the design of analytical instruments and dropping the price of computers by a factor 10 by the late sixties. The dedicated computer (by now called minicomputer) appeared in the bigger analytical research laboratories. Although access to such a computer was not very convenient and interfacing was not standardized and painful, the analytical laboratory could generate masses of data and results, such as digitized mass spectra, infrared spectra etc. The introduction of these dedicated (interfaced) computers was going to change the face of chemical analysis, if adapted to our needs. Before a computer can do anything with signals, they need to be converted from an analogue into a digital form. Digital data processing required the development of discrete, numerical al-

gorithms and their translation into a computer readable code, called software. It is, therefore, not surprising that at that time much attention was given on the treatment of signals. In 1968 the Fast Fourier Transform algorithm was published [15] for the calculation of a discrete Fourier Transform. Its counterpart in the time domain, digital smoothing, was developed by Savitzky and Golay [16] and published in 1964. The application of advanced algorithms on analytical signals became easier with the publication of complete program listings. For the processing of UV-Vis and IR-spectrometry data, for example R. N. Jones [17, 18, 19], published quite a complete and useful package, which contained programs for curve fitting, the calculation of derivative spectra, spectrum subtraction etc. The underlying algorithms were selected after a careful investigation of performance.

The seventies are marked by tumultuous developments. Computers evolved from an expensive tool for the few to a cheap, everywhere present tool for everybody. In 1978 the first complete microcomputersystem was introduced, first with relatively modest capabilities but later (80-ies) getting more calculating power than the obsolete mainframe from the sixties. Of course, the advances in digital electronics would also influence analytical instruments. Analogue meters were first replaced by digital BCD-displays. Later switches and buttons were replaced by a keyboard. At the 1975 Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, the first analytical instruments appeared which where operated under complete control of a microprocessor. These included an electrochemical system, a number of X-ray fluorescence systems, a mass spectrometer, and a programmable graphite furnace atomic absorption instrument. Besides the process of further sophistication of existing measuring principles, new types of devices for chemical analysis were introduced. For example, a chromatograph coupled to a mass spectrometer. This new type of equipment, by now called hyphenated method, generates no longer one spectrum or one chromatogram, but is capable to measure several spectra at short time intervals. The data form a data matrix. When operating such an instrument during 15 minutes, with one spectrum per second, digitized over 200 mass units, this data matrix contains 180.000 datapoints! If one data point occupies 4 byte, 720 Kb information has been collected. The impact of these hyphenated systems in analytical chemistry can be read from Table 4, which shows the state of the art by the beginning of the 80-ties 52 different hyphenated methods are available and 16 new methods are expected to appear in the eighties [20].

An eversince lasting development in analytical chemistry, of course, is the introduction of new techniques. Inductively coupled plasma atomic emission in the seventies, is an example. Obviously a new barrier to the production of analytical information was the problem of how to transform these masses of data into accurate and reliable information. Equally, a growing need was felt to evaluate and optimize analytical methods and procedures for obtaining a cost-effective laboratory operation. This was the ground, fertilized by spectacular developments in micro-electronics, on which a new discipline in analytical chemistry, Chemometrics, was born. Early Chemometric research (sometimes without using that name) was concentrated in the U.S.A., Sweden, Belgium and the Netherlands. The first paper mentioning the name "Chemometrics" was from Wold and was published in 1972 in the Journal of the Swedish Chemical Society. At the same time, the "Arbeitskreis Automation in der Analyse" initiated a discussion on a systems approach of analytical chemistry, in West-Germany. Chemo-

Table 4. The state of the art of the Hy-phen-ated methods. From T. Hirschfeld, Anal. Chem. *52*, 299A (1980). Reprinted by permission of the American Chemical Society, Washington DC



metrics research in the U.S.A. and Sweden was from the early beginning focussed on the application of pattern recognition and related techniques as factor analysis and on the optimization of analytical procedures.

In 1969 Jurs, Kowalski and Isenhour published a first series of papers in ANALYTICAL CHEMISTRY, reporting the results of applying the linear learning machine to low resolution mass spectral data [21,22,23]. The goal of these studies was to extract molecular structural information directly from spectral data. The first application of the SIMPLEX method for a sequential optimization, which was developed in the early 1960's [24], dates from 1969 [25] and was picked up soon by several others. Deming [26] investigated in detail the applicability of the SIMPLEX method for the optimization of a wide variety of analytical procedures. The Dutch-German-Belgian research was more focussed on a systems approach of the information production in the analytical laboratory.