Russian Chemical Reviews **75** (4) 271 – 287 (2006) © 2006 Russian Academy of Sciences and Turpion Ltd

DOI 10.1070/RC2006v075n04ABEH003599

Chemometrics: achievements and prospects

O Ye Rodionova, A L Pomerantsev

Contents

I.	Introduction	271
II.	Data and models used in chemical analysis	274
III.	Qualitative analysis methods. Exploration, classification and discrimination	276
IV.	Quantitative analysis methods. Calibration	280
V.	Data preprocessing and signal processing	282
VI.	Conclusion	283

Abstract. The key chemometric methods and models used to solve the problems of qualitative and quantitative analysis and for process analytical technology are considered. The achievements in the field of chemometrics made in the last 20 years are surveyed. The trends and prospects for its development are discussed. The bibliography includes 228 references.

I. Introduction

1. The history of chemometrics and its position in the system of knowledge

Twenty years have passed since the publication of the Russian translation of the only (until recently) book on chemometrics;[1] much has changed during this period. Currently, chemometric methods are used in various fields of science and engineering. This review is mainly devoted to analytical chemistry where three fields of application of chemometrics can be distinguished: qualitative and quantitative analysis, process analytical chemistry and design of experiments.[2] The attention is focused on the first application, the second one receives less attention and the third is barely considered. This choice of key points is due to the fact that the awareness of Russian chemists of the chemometric methods increases in exactly this order. Numerous papers dealing with design of experiments [3] and metrology [4] are published in Russian scientific journals.

The number of publications dealing with chemometrics rapidly grows: 15 years ago ~ 100 papers per year were published, while now their number is more than 5000 a year. Therefore, when preparing the review we reasonably restricted its scope. Analysis of chemical data is the most important direction in chemometrics. In recent years, it has been rapidly and fruitfully developing; analytical chemists have proposed not only new methods for data processing, but also new approaches to experiment setting.

Chemometrics is a synthetic discipline at the boundary of chemistry † and mathematics. As is often the case with boundary disciplines, it still lacks a generally recognised definition. The most popular definition was proposed by Massart,[6] namely, 'Chemometrics is the chemical discipline that uses mathematical, statistical and other methods employing formal logic to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical information by analyzing chemical data'. Probably, many people would accept this definition. However, the scope of science should be determined by the objects and goals it pursues rather than by the methods and instruments used. Certainly, the problem of information retrieval from source data is very important for both practice and the development of the theory; however, the experiment setup that would give results containing the required information is equally important. These two equivalent aspects, *i.e.*, retrieval of information from the data and collection of data that contain the desired information, have been reflected in the modern definition of chemometrics proposed by Wold.[7] Chemometrics solve the following problems in chemistry:

— how to get chemically relevant information out of measured chemical data;

— how to represent and display this information;

— how to get such information into data.

The vigorous development of chemometrics in the late 1970s is correlated with the advent, in the same period, of high-speed computer facilities, which have become universally available to scientists and engineers. This allowed implementation of many complicated algorithms, especially for analysis of data obtained in multiresponse and multivariate experiments. As a consequence, more complex equipment capable of performing a much higher number of measurements appeared. However, it turned out that a large amount of data does not necessarily mean that there is enough information. Therefore, analytical chemists have started to use chemometric methods to retrieve this information and to confirm that the conclusions drawn are reliable. This led to the first obvious success. It was found that traditional labour-consuming analytical methods that require unique equipment and expensive chemicals can be replaced by much faster and less expensive indirect methods. This trend is manifested most clearly in the use of IR spectroscopy, especially in the near region, which has previously considered to be of low utility due to high noise level difficult to eliminate, caused by intense absorption of water

O Ye Rodionova, A L Pomerantsev N N Semenov Institute of Chemical Physics, Russian Academy of Sciences, ul. Kosygina 4, 119991 Moscow, Russian Federation. Fax (7-495) 939 74 83, tel. (7-495) 939 74 83, e-mail: oksana@chph.ras.ru (O Ye Rodionova), forecast@chph.ras.ru (A L Pomerantsev)

Received 23 August 2005
Uspekhi Khimii **75** (4) 302 – 321 (2006); translated by Z P Bobkova

† Chemometrics appeared as a separate subdiscipline within analytical chemistry in 1974.[5] B Kowalski (USA) and S Wold (Sweden) can be considered its founders.

and by the scattering effect in reflectance spectra.[8] Therefore, the first works in chemometrics were devoted to methods of analysis of spectroscopic data,[9–11] construction of calibration models (calibrations) by the principal component analysis [12] and projections to latent structures.[13]

When speaking about the history of chemometrics, one cannot but mention the scientists who laid the grounds of the chemometric approach well before the 1970. Apparently, K Gauss, who proposed the least-squares method in 1795, is the first to be mentioned. Gosset (known under the name Student),[14] who worked as the analyst at a brewer and used methods for analysis of chemical data back in the late 19th century, should also be regarded a first chemometrician. In the early 20th century, Pearson's study was published [15] in which he proposed the principal component analysis; the works of Fisher,[16] the author of numerous statistical methods such as the maximum likelihood method and factor analysis, and the pioneering studies on experiment planning [17] were published somewhat later. Among Russian scientists, one should mention, first of all, Nalimov,[18] who greatly contributed to the theory of design of chemical experiment.

The chemometrics appeared and has developed for a long period within the framework of analytical chemistry; specialists in this field still remain the main users of chemometric methods. However, a tendency appeared with time regarded by some researchers as the departure of chemometrics from 'under the wing' of analytical chemistry to become an independent discipline. Two circumstances provided grounds for this conclusion. The first one is complication of the mathematical tools used in chemometrics. Ten years ago analytical chemists could learn and accept the multivariate approach to data analysis, *i.e.*, methods such as the projection to latent structures [19] or singular value decomposition.[20] However, subsequently, in the period of general enthusiasm of chemometricians about new methods of data analysis (multiway approach,[21] wavelet analysis,[22] support vector machines,[23] and so on), some gap between chemists and chemometricians started to form: chemists did not understand what and why chemometricians did, while the latter in turn did not realise why their new methods are not in demand in analytical chemistry. Second, numerous applications appeared in which the chemometric approach was successfully used in the fields far from analytical chemistry, for example in multivariate statistical process control,[24] image analysis [25] and in biology.[26] This lack of understanding is also obvious from the fact that at the last conference 'Chemometrics in Analytical Chemistry' (CAC-2004, Lisbon),[27] many participants argued whether chemometrics is still a part of analytical chemistry.

It can be seen from the foregoing that chemometrics is closely related to mathematics, especially mathematical statistics. Most analytical chemists understand the necessity of using statistical methods in chemical analysis and apply them to calculate average values, deviations or detection limits, to verify hypotheses and so on. They believe that these simple operations form the basis of chemometric approach in analytical chemistry. However, only some of researchers realise that this is not true and can use all diversity of chemometric methods for the analysis of chemical data.

It should be noted that for efficient application of chemometrics, it is not necessary to know, for example, the statistical theory of the principal component analysis; understanding of the fundamentals and basic ideas of this approach is sufficient. However, one should indeed know the methods of data preprocessing and variable selection principles, and, what is most important, know how to interpret correctly the data projections (loadings and scores) in the principal component space. As shown by long-term experience, this skill can be applied without in-depth mathematical knowledge. The idea of this review is to describe the main principles, methods and achievements of chemometrics using as little mathematics as possible and with geometric interpretation prevailing over the algebraic one.

Mathematicians [28] consider with every reason that many methods and algorithms used in chemometrics are poorly substantiated. Chemometric specialists regard their activity as a compromise between the possibility and necessity, believing that a practical result is more important than a theoretical substantiation of its impossibility. Being faced with practical problems of interpretation of very large and intricately organised data,[29] they create new methods of analysis so quickly that mathematicians, according to American statistician Friedman,‡ have no time not only to criticise them, but even to merely understand what happens in chemometrics. This approach is in contrast with the situation existing in biometrics,[30] which can be figuratively called the 'elder sister' of chemometrics. Since Fisher's time, only approved classical methods of mathematical statistics such as factor analysis or linear discriminant analysis have been traditionally used in biometrics. Meanwhile, specialists engaged in another closely related field, psychometrics,[31] are actively developing new approaches to data analysis. For example, the method of projection to latent structures, most popular in chemometrics, has been developed by Wold [32] for application in this field.§

Owing to this vigorous approach to data analysis, chemometrics has found numerous applications in various fields of chemical science (for example, to study kinetics in physical chemistry,[34] to predict the activity of compounds from their structure (QSAR) in organic chemistry,[35] in polymer chemistry,[36] and in theoretical and quantum chemistry [37]) and in related and other fields (for example, in brewing,[38] astronomy,[39] in forensic science [40] and quality control of the manufacture of superconductors [41]).¶

Some directions of chemometrics were developed in the USSR and later in Russia. Back in the 1950s, studies dealing with mathematical description of equilibria were carried out under Komar's [43] direction at the Kharkov State University. More recent relevant publications include the studies by Gribov [44] and Elyashberg [45] on spectroscopic methods, Mar'yanov [46] on titrimetric analysis, Derendyaev and Vershinin [47] on computer identification of organic compounds and Zenkevich [48] on chromatography. The active use of the chemometric approach is characteristic [49] of the scientific school of Academician Zolotov.[2] QSAR studies related to chemometrics headed by Academician Zefirov are underway.[50] The metrological aspects and control of the quality of chemical analysis are investigated in Dvorkin's works.[51] A research group headed by Vlasov [52] at the St Petersburg State University is working on sensor systems known as the 'electronic tongue', and analogous systems called 'electronic nose' are developed at the Voronezh Technological Academy.[53] All of these fields actively utilise chemometric methods. Razumov [54, 55] and his colleagues from the Institute of Chemical Physics of the RAS (Chernogolovka) employ multivariate methods of data analysis to solve problems of chemical kinetics. In recent years, new research groups that develop and utilise chemometric methods appeared in Russia, namely, the groups of Rodionova,[56] Pomerantsev,[57] Bogomolov [58, 59] (in Moscow); Kucheryavski,[60] Zhilin [61] (in Barnaul); Romanenko [62] (in Tomsk) and Shabanova and Vasil'ev [63] (in Irkutsk).

2. Information and software provision

We have already noted the monograph well-known in Russia,[1] which reflects the state-of-the-art in the chemometrics by the mid-

‡ J Friedman *Boosting and Bagging*. Available at http://www.amstat.org/sections/spes/GRC2001.htm

§ It is of interest that in the early 1970s the prevailing opinion was that 'the method seems to have few applications in the physical, engineering and biological sciences. It can sometimes be useful in the social sciences as a way of finding effective combination variables' (see Ref. 33).

¶ Detailed analysis of the use of chemometric methods in various fields is given in the monograph by Brereton,[42] to which an interested reader can refer.

1980s. At present, the chemometric methods are described most comprehensively in a two-volume edition [64, 65] written by a group of authors headed by Massart. Apart from the detailed description of the key chemometric methods and techniques, this edition gives numerous examples of their practical application. In addition, there are lots of editions meant for different types of readers. The students and specialists in analytical chemistry who start to become familiar with chemometrics are advised to resort to the monograph by Brereton;[42] other books [66, 67] would be useful for the researchers engaged in spectral analysis. A lot of useful information can be found in the publication by Beebe *et al.*[68] One cannot but mention the monograph by Malinowski,[69] which is still considered to be the best relevant handbook by many analytical chemists. The theoretical grounds of chemometrics can be found in other studies.[70, 71] Recently, a handbook [72] containing a brief description of chemometric techniques has been translated into Russian. An interesting introduction to chemometrics was written by Maryanov.[73] An abridged translation of the handbook written by Esbensen [74] was published in Russia in a short run (for the participants of three scientific schools in chemometrics).

Two specialised journals, *Journal of Chemometrics* and *Chemometrics and Intelligent Laboratory Systems*, are devoted to the problems of chemometrics. Papers describing the results of application of chemometric methods for solving applied problems are routinely published by more than 50 scientific journals, for example, *Analytical Chemistry*, *Analytica Chimica Acta*, *Analyst*, *Talanta*, *Trends in Analytical Chemistry*, *Journal of Chromatography*, *Computers and Chemical Engineering*, *Vibrational Spectroscopy*, *etc*. The number of publications the authors of which use chemometric methods as the main tool for analysis and processing of experimental data increases every year (Fig. 1).

Problems of chemometrics are considered at both small regional conferences and seminars and at regular international conferences. The conference 'Chemometrics in Analytical Chemistry' [27] and 'The Scandinavian Symposium on Chemometrics' are most prestigious.† In Russia, international workshops symposia 'Modern Methods of Multivariate Data Analysis' are held annually starting from 2002.[75–77] The theoretical and applied aspects of chemometrics are widely represented as Internet resources, mainly, English-language ones;‡ however, some Russian resources are also available.§

The software used in chemometrics includes specialised program packages,¶ which allow fast and vivid data processing in the interactive mode. General-purpose statistical packages are also widely used.† Often, researchers compose the procedures themselves, for example, in MATLAB codes,‡ and publish them for free access on the Internet or in books.[78]

3. Designations and terms

The following designations are used in the review. The scalar variables are marked by italic, for example, *s*. The vectors (columns) are designated by Roman bold lower-case letters, for example **x**, while matrices are shown by upper-case letters, for example, **W**; the multiway matrices are marked by italic, for example, ***G***. The array elements are designated by the same but lower-case letter. For example, w_{ij} stands for an element of matrix **W**, the subscript *i* indicating the matrix row and varies from 1 to *I*; the subscript *j* corresponds to the column number and varies from 1 to *J*. Similar designations have been used for other subscripts, for example, $a = 1, \ldots, A$. The transposition operation is denoted by superscript t, for example, $\mathbf{X}^t$.

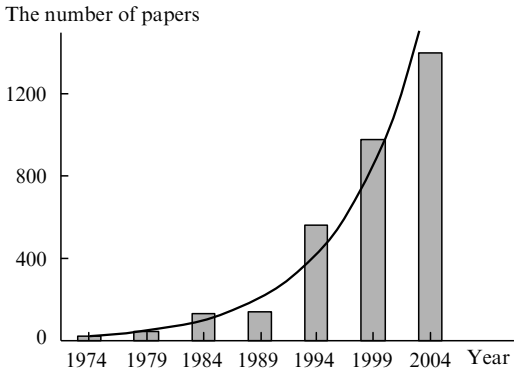


Figure 1. Diagram illustrating the growth of the number of papers in chemometrics among the publications in Elsevier periodicals.

No generally accepted set of chemometric terms has been formed in the Russian literature as yet. Some concepts were translated incorrectly or inaccurately. In many cases, translators simply avoided giving Russian names to key concepts of chemometrics, such as scores and loadings and used complex euphemisms instead. In our opinion, chemometrics cannot do without such notions or their analogues.

As in any other field of knowledge, specialists in chemometrics, often use abbreviations, *i.e*., abridged names of methods, algorithms and special terms. Below we present the list of the abbreviations used.

ALS, Alternating Least–Squares; ANN, Artificial Neural Network; DASCO, Discriminant Analysis with Shrunk COvariance matrices; EFA, Evolving Factor Analysis; GA, Genetic Algorithm; IA, Immune Algorithm; INLR, Implicit Non-linear Latent Variable Regression; ITTFA, Iterative Target Transformation Factor Analysis; KNN, *K*-Nearest Neighbours; LOO, Leave One Out; MIA, Multivariate Image Analysis; MSC, Multiplicative Signal Correction or Multiplicative Scatter Correction; MSPC, Multivariate Statistical Process Control; NAS, Net Analyte Signal; NIPALS, Non-linear Iterative Projections by Alternating Least-Squares; OSC, Orthogonal Signal Correction; PARAFAC, PARAllel FACtor Analysis; PAT, Process Analytical Technology; PC, Principal Component; PCA, Principal Component Analysis; PCR, Principal Component Regression; PLS, Projection on Latent Structures; PLS–DA, PLS Discriminant Analysis; PMN, Penalized Minimum Norm projection; QPLS, Quadratic PLS; QSAR, Quantitative Structure–Activity Relationship; RMSEC, Root-Mean Square Error of Calibration; RMSEP, Root-Mean Square Error of Prediction; SIMCA, Soft Independent Modeling of Class Analogy; SIMPLISMA, SIMPLe-to-use Interactive Self-modeling Mixture Analysis; SIMPLS, SIMple Partial Least Squares regression; SMCR, Self-Modeling Curve Resolution; SPC, Statistical Process Control; SVD, Singular Value Decomposition; SVM, Support Vector Machine; WFA, Window Factor Analysis.

† *The 9th Scandinavian Symposium on Chemometrics* (*SSC9*). Available at http://www.conference.is/ssc9

‡ *Home of Chemometry Consultancy*. Available at http://www.chemometry.com; *Chemometrics Literature Database*. Available at http://www.models.kvl.dk/ris/risweb.isa (May 1, 2005); *Chemometrics World*. Available at http://www.wiley.co.uk/wileychi/chemometrics/Home.html; *The Alchemist*. Available at http://www.chemweb.com/alchemist

§ *Russian Chemometric Society*. Available at http://rcs.chph.ras.ru; *Chemometrics in Russia*. Available at http://www.chemometrics.ru

¶ *The Unscrambler*. Available at http://www.camo.no; *Eigenvector Research* Inc. Available at http://www.eigenvector.com; *Umetrics*. Available at http://www.umetrics.com

† *SPSS*. Available at http://www.spss.com; *STATISTICA*. Available at http://www.statsoftinc.com

‡ *MATLAB*. Available at http://www.mathworks.com

II. Data and models used in chemical analysis

1. Chemical data and information

Experimental data form the key object of chemometrics. Let us consider a typical structure of chemical data according to a classification reported previously [79] (Fig. 2).

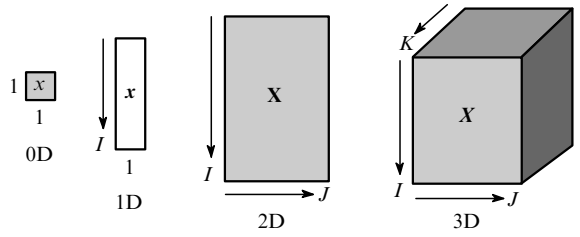


Figure 2. Graphical view of *N*-way data.

The simplest case is the case of one-dimensional data (0D), *i.e.*, merely a number, for example, the optical density, which could be obtained using a monochromatic photometer. A more complicated case is represented by multivariate, one-way data data, *i.e.*, a set of results from several measurements referring to the same sample. A spectrum and a chromatogram represent examples of this type of data. From the mathematical standpoint, they can be interpreted as a 1D vector (a column or a row), each element of which corresponds to some variable (wavelength, retention time). The number of variables stipulates the data dimensionality.

The next frequently encountered type of chemical data are two-way data. They are represented as a 2D matrix. *i.e.*, a table of numbers comprising I rows and J columns. A typical example is a set of spectra recorded for I samples at J wavelengths. Each row in the 2D matrix represents an object (in this case, a sample), while each column corresponds to a variable (wavelength). The assignment of data to either objects (samples) or variables (channels) is highly important for the interpretation. However, this classification is not always obvious. For example, when analysing the data obtained by high performance liquid chromatography with diode matrix detection, for 30 time points at 28 wavelengths, we can compose a matrix comprising 28 rows and 30 columns; but we can conversely consider the wavelengths to be variables and the retention times to be the objects. In most cases, in evolution (developing with time) experiments, the objects correspond to time instants, *i.e.*, a sample that varies with time is considered as a series of samples.[42]

With the progress in the hyphenated methods,[2] three-way (and more) data have started to receive considerable attention.[80] They can be represented by a 3D matrix, the geometric image of which is a parallelepiped with each edge corresponding to a particular type of variable. Examples of four- and even eight-way data have been reported.[80]

The data can be combined into blocks. In the simplest case, this is a single block **X**. This case is encountered most often in qualitative analysis, for example, in curve resolution. Quantitative analysis based on regression dependences makes use of data combined in two (or more) blocks. A block of predictors (for example, a 2D matrix of spectra **X**) and a block of responses (for example, a 1D vector of concentrations **y**) comprise a set of data, which are used to construct a calibration model

$$\mathbf{y} = \mathbf{Xb}.$$

More complicated structures of the data including three (or more) blocks can also be found.[81] They are analysed using special methods of path modelling.[82]

It may seem that this data systematisation — dimensionality, modality and block structure — is formal and is of interest only for mathematicians but not for chemists. However, this is not the case. During the last years, the criteria of what data can be regarded large-scale have dramatically changed. Whereas in the early 1970s, a matrix of data (for example, spectra) has been considered large if it contained 20 columns (variables, for example, wavelengths) and 100 rows (objects, for example, samples), currently a large matrix is that comprising 1 000 000 columns and 400 000 rows.[29] When these arrays are processed, they have to be divided into blocks and interpreted in turns. A formal division can be avoided by an experienced chemist who understands the point. The *N*-way concept has also been introduced not by mathematicians. This is a natural response to the demand for analysis of the data of hyphenated and evolution techniques, the number of which increases following the progress of the instrumentation. With the advent of new analytical methods such as hyperspectral measurements [83] and the use of microarrays,[84] the structure of data would become more complicated.

The main task of chemometrics is to retrieve the required chemical information from data. The notion of information is the key notion of chemometrics. What is to be treated as information depends on the problem to be solved. In some cases, it is sufficient to know that some substance is present in the system, but often quantitative measurements are also required. The data may contain the desired information; they can even be redundant, but they may also contain no information. However, in all cases, the data include an undesirable component, *i.e.*, noise (for example, errors), which hides the desired information.

For illustration, let us consider the following idealised experiment. Let there be a system representing a mixture of three substances A, B and C without admixtures, such that pure (without errors) spectra $s_A(\lambda)$, $s_B(\lambda)$ and $s_C(\lambda)$ of each component are known. The world 'spectra' is used here in the general sense; these may be any multivariate data, for example, chromatograms in which λ is the retention time. The task is to determine the component concentrations from the spectrum of the mixture $x(\lambda)$, which can also be recorded without errors. If each spectrum contains values for 30 wavelengths (times) λ, for solving this problem one can compose 30 equations with respect to three unknown concentrations c_A, c_B and c_C:

$$x(\lambda_1) = c_A s_A(\lambda_1) + c_B s_B(\lambda_1) + c_C s_C(\lambda_1),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x(\lambda_{30}) = c_A s_A(\lambda_{30}) + c_B s_B(\lambda_{30}) + c_C s_C(\lambda_{30}).$$

Evidently, for retrieving the desired information, one does not need so many equations; it is possible to retain only three of them corresponding to any[§] three wavelengths. Thus, the initial data (30-variate 1D vector) are excessive with respect to the desired information: using any three points of the spectrum gives the same concentrations.

Let us consider a more realistic example by assuming that all spectra have some random error. In this case, the concentration determined from different groups of three wavelengths would differ. These estimates can be averaged to obtain more precise concentrations. Note that the same result can be attained by repeated experiments. However, this way is inefficient, as it requires large effort and time. It is much easier to reduce the uncertainty of the quantitative analysis by increasing the number of variables (channels, wavelengths) in one experiment. This conclusion reflects the first important principle of chemometrics, namely, the use of multivariate approach in the design of experiments and in the analysis of results.

It has been noted above that data always (or nearly always) contain noise of different nature. This may be random errors, which accompany the experiment: shift of the baselines, errors in signal detection and inaccuracy in experiment preparation and conduction. In many cases, noise is represented by data that bear no valuable information. If the concentration of only two sub-

§ Strictly speaking, not any: a necessary condition is that the set has only one solution.

stances A and B was to be determined in the above experiment, then substance C would be an undesirable impurity and its contribution would be regarded as noise. What is to be attributed to noise and what should be considered information? This question is always solved with allowance for the goal and the methods used to attain the goal. This is the second principle of the chemometric approach to data analysis.

The noise and data redundancy are manifested through the correlation relationships between variables. Turning back to the idealised example, one can notice that in the matrix of 'pure' spectra **S** with the dimension 3×30 [3 rows (samples) by 30 columns (wavelengths)], only three columns would be linearly independent. Having fixed this triple, any fourth column can be represented as their linear combination. Naturally, the number equal to three is not occasional, as the system contains exactly three components. This number is called the rank of the matrix **S** and plays an important role in the chemometric analysis. When considering the same example in a more realistic version (with allowance for the errors), additional correlations between the data can be noted. This would be the case, for example, if the concentration of the third substance C is much lower than the error (noise). The data are now inadequate for reliable determination of all three concentrations and the effective rank of the matrix is equal to two. Thus, the errors in the data may give rise to random rather than systematic relationships between variables. Evidently, the former case deals with correlation relationship, while the latter case is a causal relationship.¶

The notion of effective (chemical) rank and hidden (latent) variables the number of which is equal to this rank underlies the third most important principle of chemometrics.[69] We shall illustrate this by the following example. Let there be several (I) mixtures of substances A, B and C, their pure spectra $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ being unknown. One can obtain the spectra of these samples as two-way data and to construct matrix **X** with the dimension $I \times 30$. By usual mathematical analysis, one can determine the rank of this matrix. This number gives information of how many components are present in the system or, at least, how many components can be distinguished.

Thus, chemical data involve most often inner latent relationships between the variables, resulting in numerous correlations, *i.e.*, multicollinearity. This feature can be manifested as data redundancy, which increases the quality of estimates. However, in the case of a faulty method for data processing, the multicollinearity can have an adverse effect on the quality of analysis. For example, multiple linear regression cannot be used with multicollinearity.[74] For regression analysis of this sort of data, special methods are required, for example, ridge regression[86] or projection approaches.[71]

Sampling can be an essential source of noise. The sampling theory, a substantial contribution to which was made by Jy,[87] has become very popular in recent years.[88] Its numerous applications can be found in a publication[89] completely devoted to this subject. Yet another problem that may be faced by an analytical chemist is the gaps in the data,[90] caused by various reasons such as instrument failure, going beyond the detection limits, sample deficiency and so on. Most of chemometric methods to not work with missing data; therefore, special methods are used to fill the gaps, a method based on an iteration algorithm being used most often. Each iteration consists of two steps. In the first step, the model parameters are estimated as if the data were known completely. To this end, the gaps are filled by some *a priori* permissible values, for example, average over the surrounding elements of the data array. In the second step, the model obtained is used to find the most probable values for the missing data, and the next iteration is carried out. An approach based on the likelihood maximum is also used to fill the gaps.[91] Details of these algorithms largely depend on the data description model used.

2. General strategy of data analysis. Models and methods

The chemometric methods of data analysis can be divided into two groups corresponding to two principal tasks: (i) exploration of the data, for example, classification and discrimination; (ii) prediction of new values, for example for calibration. The first-group methods usually operate with one block of data, while the second one, with at least two blocks (predictors and responses). Depending on the goals, the methods can be directed to prediction either within the range of experimental conditions (interpolation) or beyond this range (extrapolation). The methods are classified into soft also called 'black' and hard or 'white'. When formal models are used,[92] the data are described by an empirical dependence (most often, linear), which is valid in a limited range of conditions. In this case, one need not know the mechanism of the process under study. However, this method does not allow solving extrapolation problems. The parameters of soft models are devoid of a physical meaning and are to be interpreted using appropriate mathematical methods. The hard modelling[93] is based on physicochemical principles and permits extrapolating the system behaviour under new conditions. The parameters of a 'white' model have a physical meaning, and their values can help in interpretation of the elucidated dependence. However, this method is applicable only in the case where the model is known *a priori*. Each of the approaches has both advantages and drawbacks,[36] and both adherents and opponents. Historically, the hard method was intensively developed in Russia, while the soft method was developed more actively in other countries. The authors of many recent publications consider so-called 'grey' models,[94] which combine the advantages of both methods. Below we illustrate different approaches to modelling by examples from analytical chemistry.

Titrimetric processes, distinguished by a diversity of chemical reactions and signals registered, often serve as objects of mathematical modelling in analytical chemistry. The equations for titration curves are often fairly sophisticated and cannot be written in an explicit form with respect to the signal registered. This hampers the use of hard models for solving the inverse problem, namely, estimation of the parameters from the measured points of the curve. Nevertheless, by using modern computing systems, this problem can still be solved within the framework of the 'white' modelling.[95] It has been noted[96] that titration curves resemble in their shape the plots for reciprocal hyperbolic and trigonometric functions. Therefore, it has been proposed to use soft ('black') dependences composed of trigonometric functions arcsin, arccos, *etc.* According to the trade-off ('grey') approach proposed by Mar'yanov *et al.*,[46] a change in variables can transform a hard model into a piecewise-linear one. Subsequently, the parameters are estimated by the ALS method,[97] which implies progressive approximation of the model to the data: first, linear parameters are estimated by linear regression methods with non-linear parameters being fixed, and after that, non-linear parameters are estimated in a quickest descent procedure with the fixed linear parameters found previously. The procedures alternate until the results converge.

The interest in 'black' and 'grey' modelling methods is due to the difficulty of the selection and validation of a hard model. In many cases, this is reduced to mere enumeration of a small number of competing dependences, as a result of which a simple model with the smallest discrepancy is chosen. However, this does not validate the chosen method and may give rise to gross errors. Researchers often use models that have been reasonably called 'pink'† based on idealised dependences which poorly comply with artifacts present in real data such as baseline drifts, anomalous

¶ The difference between the causality and correlation is amusingly illustrated in a book,[85] which cites an example of high positive correlation between the numbers of citizens and storks in Oldenburg (Germany) in the period from 1930 to 1936. Of course, these two variables are related by a correlation dependence caused by the fact that the system includes a third latent variable to which they are both related by a causal relationship.

errors, *etc*. The formal multivariate linear models and appropriate methods for their analysis are much better 'adapted' to handling artifacts. These models operate in those cases where the hard physicochemical approach is by no means applicable. The grounds for using linear models are provided by the fact that any, even a very complicated but continuous dependence in a rather narrow range can be represented as a linear function. In this case, a fundamental question is what range can be considered acceptable, in other words, the question of the scope of applicability of the constructed soft model. This question can be answered by using model validation techniques.

When the model has been properly constructed, the initial data consist of two rather representative sets obtained independently of each other. The first set, called training set, is used for model identification, *i.e*., for estimation of its parameters. The second set, called test set, serves only for model validation. The constructed model is applied to the test set, and the results obtained are compared with the test data. The results of the comparison are considered to make decision about the model validity and accuracy. In some cases, the data array is too small for such validation. In this case, cross-validation is used.[98] According to this method, the test values are computed by the following procedure. Some invariable fraction (for example, the first 10%) of samples are excluded from the initial set of data. Then the model is constructed using only the remaining 90% of the data and then applied to the excluded set. In the next cycle, the excluded data are returned and another part of the data (the next 10%) are removed. A model is constructed once again and applied to the excluded data. This procedure is repeated until all data have functioned as excluded (in our cases, this requires 10 cycles). The leave-one-out (LOO) cross-validation procedure is used most often (without reason). Validation by the leverage correction algorithm is also employed in the regression analysis.[74] It is noteworthy that one or another validation procedure should be applied not only in quantitative, but also in qualitative analysis in solving discrimination and classification problems.

The results obtained in the analysis and modelling of experimental data always involve an uncertainty. The quantitative estimate or qualitative judgement may be changed after a repeated experiment as a result of diverse random or systematic errors either present in the initial data from the very beginning or introduced during the modelling.[99] The uncertainty in the quantitative analysis is described by either a value, *i.e*., standard deviation,[100] or an interval, *i.e*., a confidence [101] or prediction interval.[56] In qualitative analysis, verification of statistical hypotheses is used [102] in which the uncertainty is characterised by the probability of making a wrong decision.[103] The methods for estimation of the uncertainty during modelling of multivariate [104] and multiway [105] data arouse considerable interest in chemometricians. Various aspects of reliability of an analytical method are described using special characteristics: specificity, selectivity, detection limit, signal-to-noise ratio.[73] A topical method for their determination is the approach based on the NAS concept.[106] A multivariate NAS vector is defined as a part of the full signal (spectrum), which is used for modelling and prediction.[107] The remaining part of the signal, which includes errors and contributions from foreign components, is considered as noise. The NAS concept was applied to the problem of determining the detection limit in the analysis of two [108] and three-way [109] data. The obtained results have found numerous practical applications, one being considered below.

† See O N Karpukhin *Global (strategic) problems of the practical use of complex mathematical statistics methods (chemometrics)* Report at the fourth International Symposium 'Modern Methods of Analysis of Multivariate Data' (WSC-4). Chernoholovka, February 14 – 18, 2005. Available at http://www.chemometrics.ru/articles/karpukhin

The reliability of an analytical method largely depends on the data that have been used to construct and validate the model. The presence of outliers [110] or spurious data decreases the accuracy of the model and, conversely, the presence of representative (significant) samples in the experiment [111] substantially improves the model quality. The data significance can be estimated by classical regression methods [112] or by non-statistical procedures.[56] When the constructed model is used to determine desired parameters, similar problems arise. The method may prove inapplicable to some samples (an outlier in the prediction [113]) or give inaccurate results. Estimation of the method uncertainty for particular samples rather than on average [114] is a complicated task, which is tackled by a number of research groups.[115] Their effort determines the successful solution of practical problems such as calibration transfer,[116] variable selection [117] and the construction of robust models for data analysis.[118]

III. Qualitative analysis methods. Exploration, classification and discrimination

1. Principal component analysis

Modern instruments easily do a multitude of measurements per unit time. For example, if a spectroscopic sensor is used *in situ* to measure a spectrum at 300 wavelengths every 15 s, then after 1 h it will produce a 240×300 matrix of data, *i.e*., 72 000 values. However, due to the multicollinearity, the fraction of useful information in this array may be relatively low. To isolate useful information, data compression methods are used in chemometrics (unlike the traditional approach in which only the results of some, especially significant measurements are selected from the data). For representing the initial data, new latent variables are used in these methods. Two conditions must be fulfilled. First, the number of new variables (chemical rank) should be much lower than the number of the initial variables, and, second, the loss caused by data compression should be commensurable with the noise. Data compression allows one to represent useful information in a more compact form convenient for visualisation and interpretation.

Data are compressed most often using the PCA technique,[19] which underlies other similar chemometric methods including EFA,[119] WFA,[120] ITTFA [121] and numerous classification methods, for example, SIMCA.[122] The principal component analysis implies decomposition of the original 2D matrix **X**, *i.e*., representing it as a product of two 2D matrices **T** and **P**,[74]

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{t}} + \mathbf{E}. \qquad (1)$$

In this equation, **T** is called the matrix of scores, **P** is the matrix of loadings and **E** is the matrix of residuals (Fig. 3). The numbers of columns, $\mathbf{t}_a$ in the matrix **T** and $\mathbf{p}_a$ in the matrix **P**, are equal to the effective (chemical) rank of the matrix **X**. This value is designated by A and is called the number of principal components; naturally, it is smaller than the number of columns in the matrix **X**.

To illustrate the PCA method, let us turn back to the example considered in Section II.1. The matrix of the mixture spectra **X** can be represented as the product of the concentration matrix **C** and the spectrum matrix of pure components **S**

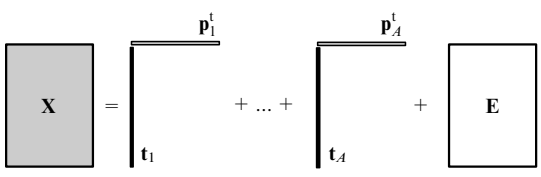


Figure 3. Graphical view of the principal component analysis.

$$\mathbf{X} = \mathbf{CS}^{t} + \mathbf{E}. \quad (2)$$

The number of rows in matrix **X** is equal to the number of samples (I), each row corresponding to the spectrum of a single sample recorded for J wavelengths. The number of rows in matrix **C** is also equal to I and the number of columns corresponds to the number of components in the mixture ($A = 3$). The matrix of pure spectra is present in expansion (2) in the transposed form, as the number of its rows is equal to the number of wavelengths (J), while the number of columns is equal to A. As noted above, in the analysis of real data complicated by errors represented by matrix **E**, the effective rank A does not necessarily coincide with the real number of components in the mixture. More often, it is greater due to the influence of non-concentration factors, for example, temperature.

The problem of resolution of the experimental matrix **X** into 'pure' components corresponding to the concentrations **C** and spectra **S** (understood in the generalised sense) is a special field in chemometrics called curve resolution.[123] Two directions can be distinguished in this field. The first one uses self-model curve resolution (SMCR)[124] and is mainly applied as a supplement to hyphenated chromatography.[125] The self-model approach is implemented using soft modelling method, for example, PCA or EFA, which do not utilise the conceptual knowledge about the system. Within the framework of this approach, the SIMPLISMA method can be distinguished,[126] which makes use of a simple but fairly effective procedure based on variable selection.[127] Conversely, the second direction employs *a priori* information about the processes and utilise 'grey' models.[128] This direction is applied in the studies of kinetics[34] and thermodynamics.[129] The key point in these problems is determining the chemical rank of the system, *i.e.*, the number of principal components A.[130] In the ideal case, the predicted spectra **S** and concentrations **C** should be close to the true values, although they cannot be exactly recovered. The reason is not only the experimental error, but also the fact that the spectra can partially overlap. When PCA is used to resolve the data into the chemically meaningful components, as in Eqn (2), it is also called factor analysis, unlike the formal principal component analysis.[131]

The principal component analysis is efficient not only in the problems of resolution. It is used to analyse any chemical data. In this case, the score **T** and loading **P** matrices can no longer be interpreted as the spectra and the concentrations, and the number of principal components A, as the number of chemical components present in the system. Nevertheless, even formal analysis of the scores and loadings is very useful for understanding the data structure. Below we present a simple two-dimensional illustration of PCA.

Data consisting only of two highly correlated variables x_1 and x_2 are presented in Fig. 4*a*. The same data in new coordinates are shown in Fig. 4*b*. Loading vector $\mathbf{p}_1$ of the first principal component (PC1) determines the direction of the new axis along which the data change more appreciably. The projections of all initial points on this axis form vector $\mathbf{t}_1$. The second principal component $\mathbf{p}_2$ is orthogonal to the first one, its direction (PC2) corresponding to the largest variation in the residuals (shown by segments perpendicular to the axis $\mathbf{p}_1$).

This trivial example shows that principal component analysis is executed successively, step by step. In each step, the residuals $\mathbf{E}_a$ are studied, the direction of the most pronounced change is chosen among them, the data are projected onto this axis, new residuals are calculated and so on (NIPALS algorithm).[74] According to another popular data compression algorithm, SVD,[132] the same decomposition (1) is constructed without iterations. The number of principal components A is chosen (in other words, the iteration procedure is terminated) based on criteria that show the accuracy of the decomposition attained. Assume that matrix **X** has I rows

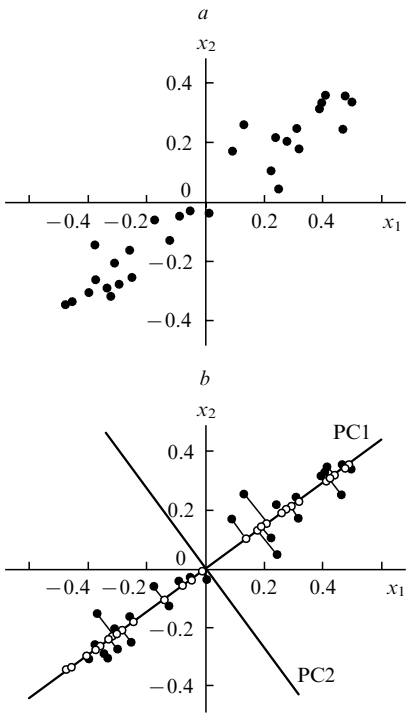


Figure 4. Graphical illustration of the principal component analysis. (*a*) data in the initial coordinates, (*b*) data in the principal component coordinates.

and J columns and A principal components participate in expansion (1). The values

$$\mu_a = 100\sum_{i=1}^{I} \mathbf{t}_{ia}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2 , \quad (3)$$

$$E_a = 100\left(1 - \sum_{i=1}^{I}\sum_{j=1}^{J} \mathrm{e}_{ij}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2\right), a = 1, \ldots, A$$

are called the normalised eigenvalue and explained variance, respectively. They are usually plotted *vs.* value a. A sharp change in these values attests to the required number of principal components. For a correct choice of A, test-validation or cross-validation is required.

Equations (1) contain no absolute term; therefore, prior to data decomposition, data centring (*i.e.*, subtraction of the average over the column) is often required.

The principal component analysis can be interpreted as data projection onto a subspace with a lower dimensionality. The residuals **E** thus arising are regarded as noise containing no significant chemical information. In this subspace, one can introduce a sample dissimilarity measure called Mahalanobis distance,[133] which helps to solve many problems of qualitative analysis. Yet another potent method for data analysis in a projection subspace is the Procrustean rotation method.[134]

When exploring the data by PCA, the attention is focused on score and loading plots. They bear information on the data structure. In the score plot, each sample is depicted in the $(\mathbf{t}_i, \mathbf{t}_j)$ coordinates, most often, $(\mathbf{t}_1, \mathbf{t}_2)$. The proximity of two points implies their similarity, *i.e.*, positive correlation. The points located at a right angle are uncorrelated, while those located in the opposite positions have a negative correlation. By using this approach in chromatographic analysis,[42] one can find out, for example, that the linear section in the score plot corresponds to the

regions of pure components in the chromatogram, the curved sections are regions of peak overlap, and the number of such sections corresponds to the number of components in the system. Whereas the score plot is used for analysis of sample relationships, the loading plot is used to study the role of variables. In the loading plot, each variable is reflected by a point in the $(\mathbf{p}_i, \mathbf{p}_j)$ coordinates, for example, $(\mathbf{p}_1, \mathbf{p}_2)$. By analysing this plot similarly to the scope plot, one can understand which variables are interconnected and which are independent. A joint investigation of pair score and loading plots also helps to retrieve useful information from the data.[74]

Consider an example of practical use of PCA in chemical analysis. The possibility of using near-IR spectroscopy for detection of counterfeit drugs has been considered.[135] Samples of true (N1, 10 items) and fake (N2, 10 items) tablets of a popular antispasmodic agent have been studied. Twenty diffuse reflectance spectra $R(\lambda)$ were recorded on a Bomem MB160 instrument with a Powder Sampler attachment in the $3800-10\,000\ \mathrm{cm}^{-1}$ range (1069 wavelengths) without special sample preparation. The initial data were converted to $-\log R$, centred and prepared by the MSC procedure (Fig. 5).[74] The negative values of the signal are due to the use of different gain settings for the background and the sample spectra.

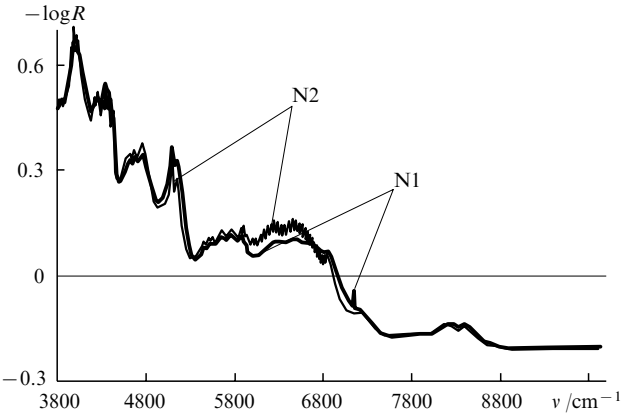


Figure 5. Spectra transformed by the MSC procedure.[135]
Here and in Fig. 6 the following designations are used: N1 are true tablets, N 2 are fake tablets.

In the PCA score plot $(\mathbf{t}_1, \mathbf{t}_2)$ for these spectra (Fig. 6 *a*), one can clearly see two groups of points corresponding to the true and counterfeit tablets. The scatter of points in group N2 (counterfeit) is much greater than in group N1 (original). This may be attributed to a better quality control in the legal production. In this example, it is sufficient to use only two principal components for which $\mu_1 = 94\%$, $\mu_2 = 4.9\%$, $E_2 = 99\%$.

2. Classification and discrimination

The example considered above refers to classification problems. This is a rather extensive class of problems of qualitative chemical analysis with the goal of attributing a sample to a particular class. Classification problems can be subdivided into two groups. One group includes so-called unsupervised problems that use no training set; they can be regarded as a sort of explorative analysis. This approach has been used in the above example with counterfeit tablets. The problems of the second group referred to as supervised classification are also called discrimination problems. They are solved using a training set of samples, which are known *a priori* to belong to particular classes. The methods for unsupervised classification are mainly based on PCA decomposition followed by analysis of distances between classes,[136] construction of dendrograms, the use of the fuzzy set,[137] *etc*. Procrustean rotation [138] and Mahalanobis distance [139–141] have been used for

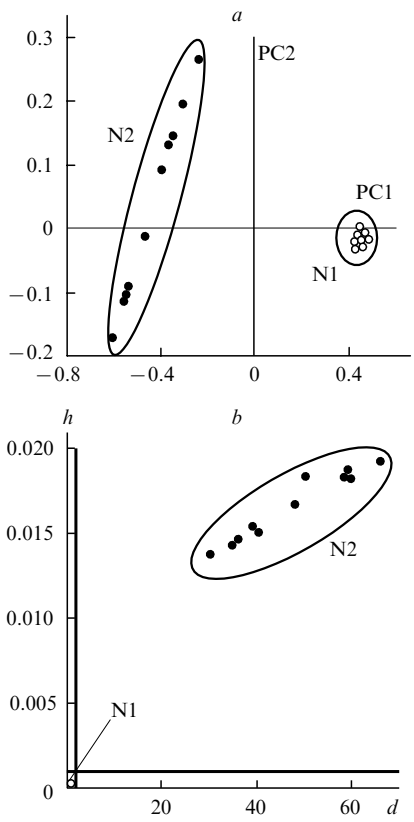


Figure 6. Determination of counterfeit drugs using PCA scores (*a*) and SIMCA (*b*).

these purposes. However, if discrimination is possible, these methods should be preferred.

A training set of samples is used to construct a classification, *i.e*., a set of rules that can be used to assign a new sample to one or another class. When the model (or models) has been constructed, it has to be test- or cross-validated to determine the degree of its precision. If the validation is successful, the model is ready for practical use. In analytical chemistry, multicollinear data (spectra, chromatograms) are usually classified; therefore, the discrimination model is nearly always multivariate and is based on projection approaches, PCA and PLS. Note the use of linear discriminant analysis in the near-IR spectroscopy [142] and canonical discriminant analysis.[143] The SIMCA method [144] developed by Wold [122] is very popular.

The SIMCA method is underlain by the assumption that all objects that belong to the same class have both similar and distinctive features. When constructing a discrimination model, one should take into account only the similarity, while the distinctive features should be rejected as noise. To this end, every class from the training set is modelled independently using PCA with different numbers of principal components A. After that, distances between the classes and the distances between each class and the new object are calculated. Two values are used as distance functions. The distance from the object to a class (d) is found as the root-mean-square value of the residuals $\mathbf{e}$ arising upon projecting an object onto the class

$$d = \sqrt{\frac{1}{J-A}\sum_{j=1} \mathrm{e}_j^2}.$$

This value is compared with the root-mean-square residue within the class

$$d_0 = \sqrt{\frac{1}{(I-A-1)(J-A)}\sum_{ij} \mathrm{e}_{ij}^2}.$$

The second value, the distance from the object to the centre of the class (h), is found as the leverage (the squared Mahalanobis distance)

$$h = \frac{1}{I} + \sum_{a=1}^{A} \frac{\tau_a^2}{\mathbf{t}_a^{\mathrm{t}} \mathbf{t}_a},$$

where τ_a is the projection of a new sample (score) on the principal component a, and $\mathbf{t}_a$ is the vector containing scores for all training samples in the class.

The use of the SIMCA method for tablet discrimination is illustrated in Fig. 6 *b*. As the class, true tablets were used; the plot shows the distances d and h from counterfeit samples. The vertical and horizontal lines dictate the rules that can be applied to assign a new item to the class of true tablets. It can be seen than all samples of fake tablets are located far away from the true class; therefore, they can be readily discriminated. On the given scale, the points that correspond to true samples converge into one point, which almost coincides with the origin of coordinates.

Apart from SIMCA, a similar DASCO method,[145] as well as KNN,[146] SVM [147, 148] and many other methods are used for discrimination of chemical data. The PLS–DA method is a powerful tool.[149] The basic idea is that the discrimination rules for K classes are specified by linear regression of the form

$$\mathbf{XB} = \mathbf{D},$$

where **X** is the full matrix of all initial data ($I \times J$), **B** is the matrix of unknown coefficients ($J \times K$), and **D** is a special matrix ($I \times K$) composed of zeros and ones. During the construction of matrix **D**, ones should be placed only in those rows (samples) that belong to the class corresponding to the number of the column. The regression problem is solved by the PLS method (see below), which allows one to use subsequently the constructed regression to predict the assignment of new samples. For this purpose, the response of a new sample is predicted and the result is compared with zero or one.

3. Three-way methods

Principal component analysis has been developed for the analysis of data that can be represented as a two-way 2D matrix. However, in recent years analytical chemists deal more and more often with three (and higher) way data with a more complex structure. These data are gained, for example, by using hyphenated [150, 151] and evolution methods.[152] The data are compressed using special approaches three of which (used most often) are briefly considered in this Section. The most comprehensive and systematic description of these methods with numerous examples of their application to chemical analysis can be found in a monograph.[80] A brief review of methods and algorithms used for analysis of three-way data has been reported.[153] The same algorithms are used to process the data obtained by hyperspectral measurements [83] and for image analysis.[21]

The unfolding method [154] is the simplest method of analysis of three-way data used to unfold 3D matrix $\boldsymbol{X}$ with dimension $I \times J \times K$ into a usual 2D matrix $^{\mathrm{u}}\mathbf{X}$ with the dimension $I \times JK$ (Fig. 7). The value I is called the basic mode. After unfolding, the principal component analysis can be applied (see Section III.1). This approach is often efficient (see, for example, a study [36]), although it suffers from a number of drawbacks; first, any of three directions can be chosen as the basic mode, *i.e.*, unfolding involves uncertainty; second, the connection between neighbouring points is lost, since on passing from a 3D to a 2D matrix, it is no longer taken into account that the measurements x_{ikj} and x_{ik+lj} are neighbouring, which may be important.

The Tucker3 algorithm [155] allows processing three-way data, while maintaining their initial structure and, hence, the sequence of measurements, for example, the order of wavelengths in the spectrum or the time sequence of points in a chromatogram. The initial 3D data $\boldsymbol{X}$ are presented as three conventional 2D loading matrices (**A**, **B**, **C**) and a three-way core array $\boldsymbol{G}$. The pattern of such data expansion is shown in Fig. 8. Each element of the initial 3D matrix $\boldsymbol{X}$ can be written as the sum

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + \mathrm{e}_{ijk}, \qquad (4)$$

where a, b and c are elements of the loading matrices, each of them corresponding to a particular way; g are elements of the core array $\boldsymbol{G}$. The number of principal components along each direction (P, Q, R) can be different.

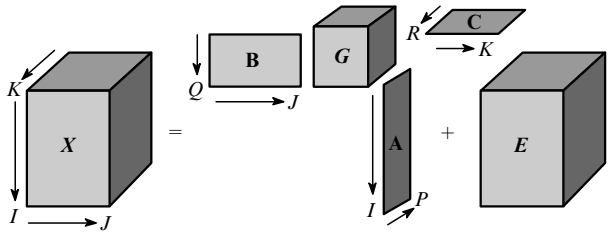


Figure 8. Graphical view of the Tucker3 model.

The PARAFAC method [153] differs from the Tucker3 model in the fact that each way is represented by the same number of principal components R. The expansion is constructed to minimise the sum of squares of the residuals e_{ijk}

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + \mathrm{e}_{ijk}. \qquad (5)$$

The major advantage of this method is the uniqueness of expansion. If a mixture of several chemical substances was studied, then with a correct choice of the number of principal components, the loading matrices are pure spectra of the initial compounds. The graphical scheme of the PARAFAC method is shown in Fig. 9.

The MATLAB codes of the PARAFAC algorithm can be found in a publication.[150] Since the loading matrices in expansion (5) are determined by an iteration procedure, this method requires a very large amount of computation. Studies aimed at acceleration of computation procedures are now in progress. A critical analysis of the most recent achievements in this field has been reported.[155] The algorithms of all the considered methods for decomposition of three-way data are documented.[156]

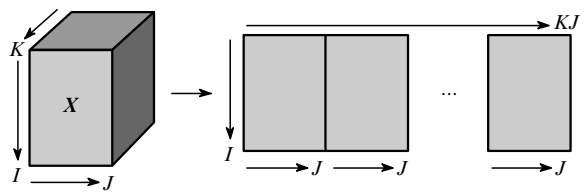


Figure 7. Graphical view of the unfolding method into a 2D matrix.

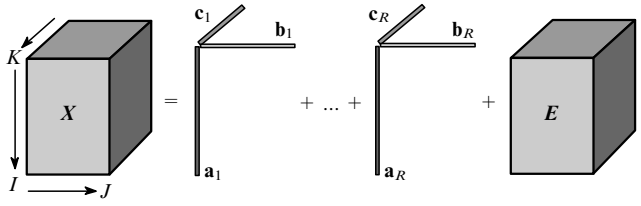


Figure 9. Graphical view of the PARAFAC model with R components.

IV. Quantitative analysis methods. Calibration

1. Linear calibration

Two blocks of data are used to solve problems of quantitative analysis.[2] The first block **X** is the matrix of analytical signals (for example, spectra, chromatograms and so on); the second block **Y** is the matrix of chemical parameters (for example, concentrations). The number of rows (I) in these matrices is equal to the number of samples, the number of columns (J) in the matrix **X** corresponds to the number of channels (wavelengths) in which the signal is recorded, and the number of columns (K) in matrix **Y** is equal to the number of chemical parameters, *i.e.*, responses. The purpose of calibration is to construct a mathematical model that would relate blocks **X** and **Y** and could be subsequently used to predict parameters **y** over a new row of analytical signals **x**.[13]

A simplest calibration model is one-dimensional regression ($J = 1, K = 1$)[157]

$$y = a + bx,$$

which corresponds to a single channel of the analytical signal. Using classical regression analysis, it is possible to construct a more complex multiple regression ($I > J, K = 1$) involving several channels,[33]

$$\mathbf{y} = \mathbf{Xb}.$$

Using these models normally implies that factors x_{ij} are known exactly, errors being present only in block **y**. Therefore, two approaches to the model construction can be distinguished: the first is called direct calibration and the other is inverse calibration.[158] In the first approach, chemical parameters ($\mathbf{X} = \mathbf{C}$) are used as independent parameters, while spectral measurements ($\mathbf{Y} = \mathbf{S}$) are used as responses. Previously, it has been believed that the direct model fits better to the assumption of the lack of errors in block **X** and, in addition, it is in agreement with the Bouger – Lambert – Beer law.[72] In the second approach, $\mathbf{Y} = \mathbf{C}, \mathbf{X} = \mathbf{S}$. This approach currently prevails in chemometrics, because it is more practically convenient, as it directly predicts the required analytical parameter (for example, the concentration **C**) from the measured signal (spectrum **S**). In addition, modern regression methods (PCR, PLS) make it poosible to handle data with errors present in both blocks.

To illustrate various calibration methods, let us turn back to the example considered in Section II.1. Now we fill this with a particular content by simulating data **X** and **Y**. Let there be a mixture of two substances A and B ($K = 2$) and an instrument that measures an analytical signal s (spectrum) at 101 channels ($J = 101$). The spectra of 'neat' substances ($c_A = 1$, $c_B = 1$) are presented in Fig. 10 *a* (curves *A* and *B*). The spectra significantly overlap; therefore, it is impossible to distinguish 'selective' channels for estimating the concentrations. Figure 10 *b* shows nine simulated spectra ($I = 9$) of various mixtures of substances A and B in which a random error has been introduced with a standard deviation of 0.05. They will be used as the training set.

To construct a one-dimensional calibration, we took the intensities $s(\lambda_{50})$ of nine signals for channel 50 and plotted them in Fig. 10 *c* as functions of concentrations c_A and c_B of A (points *1*) and B (points *2*). The calibration dependence $s = bc$ is shown by a straight line.

The calibration accuracy is usually characterised by the RMSEC value

$$\mathrm{RMSEC} = \sqrt{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2 / F}, \qquad (6)$$

where y_i and $\hat{y}_i$ are the measured and predicted values of a chemical parameter (concentration) for the training samples $i = 1, \ldots, I$; F is the number of degrees of freedom:[42] $F = I - 1$ for one-dimensional regression without an absolute term. Evidently, the lower the RMSEC, the more accurate the description of the training data. In addition, the quality of calibration is

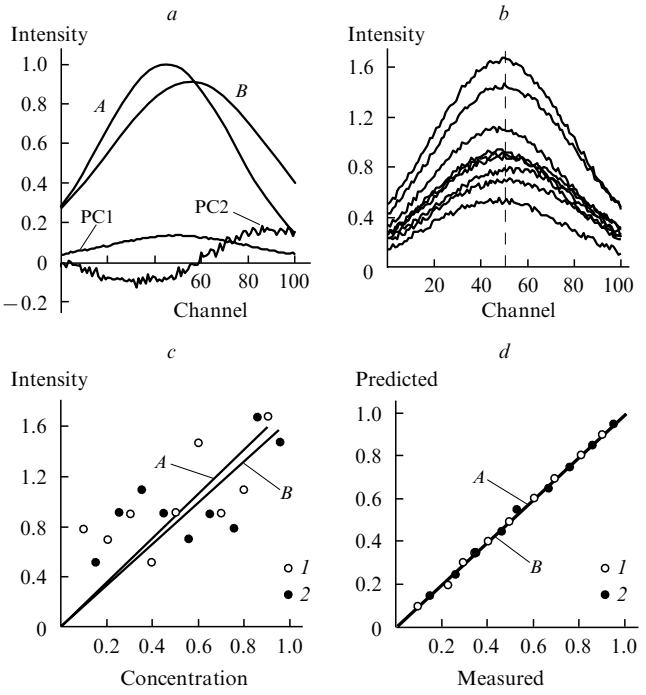


Figure 10. Examples of the construction of various calibrations. (*a*) spectra of pure (curves *A*, *B*) and principal components (curves PC1, PC2); (*b*) model spectral data; (*c*) one-dimensional calibration, $R^2_A = 0.50$, RMSEC = 0.26, $R^2_B = 0.46$, RMSEC = 0.219; (*d*) PCR calibration, $R^2_A = 0.999$, RMSEC = 0.011, $R^2_B = 0.998$, RMSEC = 0.012.

characterised by correlation coefficient R^2 between values **y** and $\hat{\mathbf{y}}$: the closer this to unity, the higher the calibration quality. The corresponding values are given in the caption to Fig. 10 *c*. The graphical dependences show that due to low instrument selectivity, the one-dimensional calibration is unsatisfactory. The calibration using multiple regression is considered below.

Consider the application of a multivariate model constructed by the PCR method.[86] The PCR method uses inverse calibration, so that $\mathbf{Y} = \mathbf{C}, \mathbf{X} = \mathbf{S}$. In terms of the PCA method, the matrix **X** can be expanded using formula (1); in our example, $A = 2$. Loading vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ thus found are shown in Fig. 10 *a* (curves PC1 and PC2). By comparing the plots presented in Fig. 10 *a*,*b*, one can see that the first principal component describes a smooth trend in the data, whereas the second component shows the noised deviations from this trend. The obtained score matrix **T** is used as the block of independent variables (predictors) in the regression to the response block **Y**, *i.e.*, $\mathbf{Y} = \mathbf{Tb}$. The results of PCR calibration are presented in Fig. 10 *d*, which shows the predicted concentrations $\hat{\mathbf{y}}$ as functions of the measured values **y** (points *1* for substance A and points *2* for substance B) and the regression lines, which coincide. The RMSEC and R^2 values demonstrate that PCR provides high 'mathematical' selectivity and gives estimates for the concentrations of A and B with much higher accuracy than the one-channel calibration. In the PCR method, the number of degrees of freedom appearing in Eqn (6) equals

$$F = I - A.$$

It has been noted above that each chemometric model requires a proper validation. In our example, the validation was carried out using a test set (test validation) comprising five samples (mixtures of A and B). The validation results for substance B are shown in Fig. 11 *a*. The data for nine samples participating in calibration and five test samples are given in the 'measured – predicted' coordinates. The RMSEC and RMSEP values and the correlation coefficients for the training (R^2_c) and test (R^2_t) sets are also given.

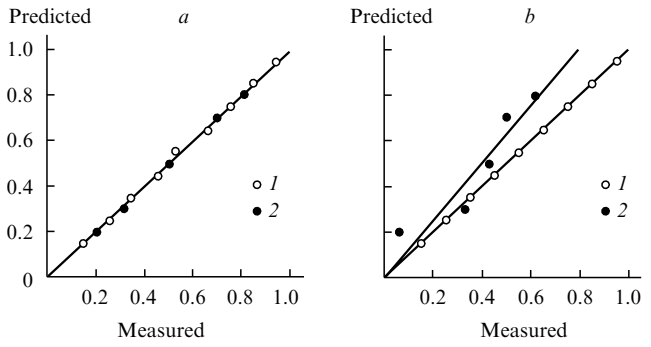


Figure 11. Validation of the calibrations in the model example using the principal component regression (*a*) and multiple regression (*b*). (*a*) $R_t^2 = 0.999$, RMSEC = 0.008, $R_c^2 = 0.998$, RMSEC = 0.12; (*b*) $R_t^2 = 0.86$, RMSEC = 0.14, $R_c^2 = 1.0$, RMSEC = 0; (*1*) training set, (*2*) test set.

The RMSEP values are calculated similarly to RMSEC [see Eqn (6)] but only for test samples. The number of degrees of freedom F is equal to the number of these samples. It can be seen that the principal component analysis passed the validation: the calibration and validation lines coincide.

Let us consider in this context the calibration by means of multiple regression. The training set consists of nine samples; therefore, not more than eight channels can be used for model construction ($I > J$); for example, the first, fourteenth, twenty-seventh, *etc*. The calibration and validation results for multiple regression are presented in Fig. 11 *b*. Since the number of samples is greater than the number of the channels only by one, the calibration line passes exactly through all points of the training set (points *1*); therefore, RMSEC = 0 and $R_c^2 = 1$. However, validation showed an unsatisfactory quality of this calibration: the accuracy is ten times worse than than the accuracy in the PCR method, and the validation straight line does not coincide with the calibration line. This is a typical example of model overfitting:[71] the accuracy of calibration is much higher than the accuracy of prediction.

The problem of balancing of data calibration is considered in many publications by Höskuldsson,[159] which introduced a new modelling concept, the so-called H-principle. According to this principle, the accuracy of modelling evaluated by the RMSEC parameter and the prediction accuracy evaluated by RMSEP are interrelated. A better RMSEC entails a poorer RMSEP; hence, they must be considered together. For this reason, the multiple linear regression, which always involves a redundant number of parameters, yields unstable models unfit for practical use.

Currently, PLS is the most widely used method for multivariate calibration in chemometrics. It resembles the PCR method; however, an important distinction is that PLS implies simultaneous decomposition of matrices **X** and **Y**

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}, \qquad (7)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^t + \mathbf{F}.$$

The projections are constructed in coordination, in such a way as to maximise the correlation between the X-score ($\mathbf{t}_a$) and Y-score ($\mathbf{u}_a$) vectors. Therefore, the PLS regression describes much better the complex relationships using a smaller number of principal component. The PLS method has been considered in detail in a book.[74] This approach has served as the basis for numerous calibration methods used in chemometrics, for example, SIMPLS,[160] PMN,[161] robust PLS,[162] ridge-PLS[163] and other.[164]

However, all these methods give predictions as point estimates, whereas in practice, an interval estimate including the prediction uncertainty is often required. The construction of confidence intervals by traditional statistical methods is impossible due to the problem complexity,[114] while the use of simulation methods[98] is hampered due to the long computation time.[101] Kantorovich[165] proposed to replace minimisation of the sum of squared deviations by a set of inequalities, which is solved by linear programming. In this case, the prediction gives an interval; therefore, this method was called 'simple interval calculation'.[36, 56] Several works in analytical chemistry have been performed using this method.[166]

2. *N*-Way regression

The mutivariate calibration methods are naturally extended to the case where blocks $\boldsymbol{X}$ and $\boldsymbol{Y}$ are represented by N-way matrices.[80] The regression can be constructed in different ways. Using PARAFAC and Tucker3, the predictor block $\boldsymbol{X}$ is represented as a product of 2D loading matrices, which are used to estimate the parameters. These methods can be regarded as extension of PCR to multiway data. An extension of PLS is the Tri–PLS decomposition of the 3D matrix $\boldsymbol{X}$, which can be represented as follows:[167]

$$^u\mathbf{X} \approx \mathbf{T}\,{}^u\mathbf{P}.$$

Here $^u\mathbf{X}$ is a 2D matrix (with the dimension $I \times KJ$) obtained by unfolding of the 3D matrix $\boldsymbol{X}$ (with the dimension $I \times K \times J$) (see Fig. 7); $\mathbf{T}$ is the 2D score matrix (with the dimension $I \times A$); $^u\mathbf{P}$ is the 2D weight matrix (with the dimension $A \times KJ$), which in turn represents the unfolding for the 3D matrix $\boldsymbol{P}$, represented as the tensor product of two 2D matrices

$$\mathbf{P} = {}^J\mathbf{P} \otimes {}^K\mathbf{P}.$$

The decomposition of block $\boldsymbol{Y}$ is carried out in a similar way

$$^u\mathbf{Y} \approx \mathbf{U}\,{}^u\mathbf{Q}.$$

Here, as in the usual PLS method, the scores $\mathbf{T}$ are chosen in such a way as to maximise the correlation between vectors $\mathbf{t}_a$ and $\mathbf{u}_a$. The regression problem $\mathbf{U} = \mathbf{TB}$ is solved by the conventional procedure.

The mathematical tool used in the multiway calibration is rather complex. However, the currently existing software ‡ allows chemists to overcome mathematical difficulties. Numerous examples of using multiway calibration in chemical analysis have been documented. For example, this method is used in spectrophotometry to determine pesticides,[168] in high performance liquid chromatography with diode matrix detection for peak resolution,[169] to determine trace concentrations of metals,[170] *etc*.

The use of gas chromatography–mass spectrometry to determine traces of clenbuterol in biological samples has been described.[171] In recent years, this method has been widely used to analyse traces of organic compounds. However, due to the complexity of biological products and low content of the analyte, the estimated detection limit depends appreciably on the method of mathematical processing of experimental data. In the study cited, seven standard samples with known clenbuterol concentrations were prepared. The mass-spectrometric detection was carried out both in the full-scanning mode (210 ions) and for eight particular ions. The obtained data are three-way, the first way being the samples, the second, the mass spectra and the third, chromatograms. The full-scanning mode gave a 3D predictor matrix $\boldsymbol{X}$ with the dimension $7 \times 210 \times 37$; the separate ion detection mode gave a $7 \times 8 \times 22$ matrix. The response block is a 1D vector $\mathbf{y}$, which includes seven concentrations.

The calibration was constructed using various three-way algorithms: PARAFAC, PARAFAC2, Tucker3 and Tri–PLS. The Tri–PLS method proved to be the method of choice as this gave the lowest detection limit. A comparison of the results obtained by this method and a standard univariate procedure

‡ See R Bro, C A Andersson *The N-Way Toolbox for MATLAB*. Version 2.02 (2003). Available at http://www.models.kvl.dk/source

showed a pronounced decrease in the detection limit, in particular, from 283 to 20.91 mg kg^{-1} in the full-scanning mode and from 73.95 to 26.32 mg kg^{-1} for scanning of particular ions. The detection limits were calculated in terms of the NAS concept.[124]

3. Non-linear calibration

In some cases, for example, in the titration problems considered above, it is impossible to build a linear calibration. In addition, the linear technique requires a large number of data, which are not always available. Two alternative approaches are possible. One implies multiple non-linear regression, while the other, multivariate non-linear calibration. Both approaches are considered below.

The non-linear regression analysis [172] can be successfully used to solve problems of quantitative analysis if the variables are few in number. In addition, a conceptual model relating the blocks **X** and **Y** is required. Apparently, the scope of such problems is not wide, including mainly kinetic (in particular, titration) problems.[95] This approach was employed, for example, in the analysis of the activity of antioxidants,[36] in solving the inverse kinetic problem [34,94] and for the above-mentioned titration.[173,174] A detailed analysis of the problems faced by a researcher who uses this approach has been reported.[57]

An alternative to the classical regression is the soft approach, which does not require the knowledge of a hard model but implies the presence of a large number of data.[78] To take into account non-linear effects, the INLR,[175] GIFI–PLS [176] and QPLS [177] methods representing the upgraded PLS method have been proposed.[81] In addition to non-linear PLS, the ANN method [178,179] simulating signal propagation in the cerebral brain cortex is used in chemometrics. This method is used successfully for function interpolation. About 10 years ago the neural network method attracted attention of chemists, which started to use it for classification,[180] discrimination [53] and calibration.[181,182] However, more recently, the interest somewhat attenuated and the use of ANN in chemometrics became more seldom. The reason is the above-noted model overfitting. When using neural networks, it is very difficult to evaluate correctly the degree of complexity of the model, which results in an unstable and unreliable prediction. Yet another interesting version of non-linear modelling that simulates biological processes is the GA method.[183,184] This method and its modification, the IA method, are useful in those cases where the chemical analysis problem fails to be formalised in terms of the usual objection functions, for example, for resolution of overlapping multicomponent chromatograms.[185] Examples of practical application of non-linear approaches in chemiluminescent analysis have been reported.[186]

V. Data preprocessing and signal processing

1. Data preprocessing

An important condition for correct modelling and, hence, successful chemical analysis is the preliminary data preprocessing, which includes various transformations of the initial ('raw') experimental values. The simplest transformations include centring and scaling.[187] Centring is subtraction of some matrix **M** from the initial matrix **X**

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{M}.$$

Usually, centring is carried out by columns: for each vector $\mathbf{x}_j$ the average values are calculated

$$m_j = \frac{x_{1j} + \ldots + x_{Ij}}{I}.$$

Then

$$\mathbf{M} = (m_1\mathbf{l}, \ldots, m_J\mathbf{l}),$$

where $\mathbf{l}$ is a vector of ones with the length I. In some cases, centring by rows is also carried out: the average values for the rows are calculated and these values are subtracted from the corresponding rows $\mathbf{x}_i^t$. In the case of multiway data, centring can be carried out separately for each way. Centring is required if the model is uniform, *i.e.*, has no absolute term, as in Eqns (1) and (7). After this operation, the chemical rank of the model decreases by unity and the accuracy of description may increase. Centring may be considered as projecting onto a zero principal component,[13] hence, it is always used in the PCA and PLS methods. However, centring should not be employed if there are gaps in the data.

The second simplest transformation of data is scaling. Unlike centring, this transformation does not change the structure of the data but only the weight of their parts on processing. Scaling can be carried out for each mode. Scaling by columns is multiplication of the initial matrix **X** by matrix **W** on the left

$$\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X},$$

where **W** is a diagonal matrix with the dimension $J \times J$. Usually, diagonal elements w_{jj} are equal to the reciprocal of the standard deviation

$$d_j = \sqrt{\sum_{i=1}^{I} (x_{ij} - m_j)^2 / I}$$

along column $\mathbf{x}_j$. Normalisation by rows is multiplication of matrix **X** by diagonal matrix **W** on the right

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}.$$

The dimension of **W** is $I \times I$, while its elements w_{ii} are the reciprocal values of the standard deviations of the rows $\mathbf{x}_i^t$. A combination of centring and scaling by columns

$$\tilde{x}_{ij} = \frac{x_{ij} - m_j}{d_j}$$

is called autoscaling.

Data scaling is often used in order to equalise the contribution of variables to the model (for example, in the liquid chromatography–mass spectrometry hyphenated method), to take into account the heteroscedastic errors, or for combined processing of different blocks of data. Scaling can also be considered as a method that allows one to stabilise the computation algorithms.[71] However, this type of transformation should be used with caution, because it can distort the results of qualitative analysis.[42]

Apart from linear transformations, non-linear transformations of experimental results are also used. For example, in analysis of the data of near-IR spectroscopy, the Kubelka–Munck transformation is often employed.[188] The purpose of this and other transformations, for example, the Box–Cox transformation,[33] is model linearisation. Simple operations with the data such as taking the logarithm [56] or extraction of the root [36] can often markedly improve the model.

Almost in all cases, the initial data contain errors both random and systematic. In order to reduce the influence of random noise, various methods of data smoothing are used, for example, moving average, the Savitzky–Golay method.[42,189] Reducing the effect of systematic errors, *i.e.*, removal of the systematic shift of the data, is more difficult. If this shift is invariable, it can be removed using centring. In the case of linear or square dependence on the variable (for example, wavelength), numerical differentiation may prove useful. In the case of more complex dependences, special methods are used; two of these are considered below.

The multiple signal correction method, called also multiplicative scattering correction (MSC),[70] was first developed [190] for near-IR spectroscopy and was based on ideas stated by Kubelka and Munck.[188] The MSC transformation procedure is simple. First, the 'basic spectrum' is determined as the average over all rows of matrix **X**.

$$\mathbf{m}^t = \frac{\mathbf{x}_1^t + \ldots + \mathbf{x}_I^t}{I}$$

Then a regression is constructed for each row $\mathbf{x}_i^t$

$$\mathbf{x}_i^t = a_i + b_i \mathbf{m}^t + \mathbf{e}_i^t$$

and the coefficients a_i and b_i are determined. The transformed data are obtained from the equation

$$\tilde{\mathbf{x}}_i^t = \frac{\mathbf{x}_i^t - a_i}{b_i}.$$

The MSC parameters a_i and b_i can be found only for some (floating) window rather than for all variables.[191]

The second method (more precisely, a group of methods), OSC [192] differs in the fact that the predictor matrix **X** is transformed using the response block **Y**. This method is applied for data preprocessing in solving problems of quantitative analysis. The idea of OSC is to remove, from block **X**, all the systematic dependences not related to the modelled response, *i.e.*, the part of matrix **X** that is orthogonal to matrix **Y**. This should result in an increase in the correlation coefficient R^2 and decrease the number of PLS components A needed for data modelling. There are many versions of this method, which was first proposed by Wold *et al.*[193] and developed by other researchers.[194, 195] The OSC procedure, like the PLS procedure, is accomplished successively by steps. In each step, a part of matrix **X** related to one OSC component is removed. The part of the matrix

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2,$$

orthogonal to **Y**, *i.e.*, such that

$$\mathbf{Z} = \mathbf{Y}^t\mathbf{X}_2 = 0,$$

is determined using an algorithm similar to PLS. The OSC method and MATLAB codes have been described in detail in a publication.[195]

An approach that improves the model quality by selection of variables is an alternative to the MSC and OSC methods. The utility of selection, *i.e.*, elimination of some columns $\mathbf{x}_j$ from the initial data **X**, was confirmed by both theoretical investigations and experimental results. This approach is used in both qualitative [196] and quantitative analysis.[197] The variable selection is performed using a number of methods, in particular,[198] Pareto optimisation [199] and 'jack knife'.[117] Of special importance is variable selection in those cases where the analytical signal continuously depends on the channel, for example, in the analysis of spectrometric data.[200] In this case, variables are selected by blocks, as in methods considered in Refs 195, 201. Apart from variable selection, the selection of samples, *i.e.*, rows $\mathbf{x}_i^t$ in matrix **X** (together with the corresponding values in the response matrix **Y**), is also used. The sample selection may also improve the model quality, but it is especially important for detection of outliers,[113] for calibration transfer from one instrument to another.[116, 202, 203] The new approach to sample classification and selection has been reported.[56]

2. Signal processing

The processing of analytical signals using various transformations and filters plays an important role in chemical analysis.[204, 205] The Fourier transform has actually revolutionised NMR, IR and X-ray spectroscopy during the last 20 years. The initial data are not recorded as primary spectra but as temporal series in which all the spectroscopic information is mixed, and recovery of the spectra requires a mathematical processing. A main reason for using Fourier spectrometry is increasing the signal-to-noise ratio; in addition, the experiment can be carried out about 100 times faster than with a conventional spectrometer. For example, this allowed ^{13}C NMR spectroscopy to become a routine analytical method, despite the low sensitivity of the ^{13}C nuclei to an external magnetic field. Using pulse NMR spectroscopy, one can accumulate and sum up signals for a large number of pulses. Simultaneously with Fourier transform spectroscopy, numerous methods appeared for improving the quality of data. These methods include the so-called Fourier deconvolution (signal separation), some operations with the initial data in the temporal domain and the subsequent use of the Fourier transform.

Yet another modern method for signal processing is wavelet analysis.[206] This allows encoding, compression and modelling of large data containing thousands of variables. This analysis is a natural development and continuation of the Fourier method. A drawback of the Fourier method is that its basis functions depend continuously on time; therefore, they cannot be used to present time-dependent data. Wavelet analysis employs basis functions with a limited range of variation of the argument, which satisfy special requirements of a scaled range. These functions shift along the signal axis and the spectra obtained upon verification provide a time-and-frequency representation with different resolutions depending on the range width. If wavelet analysis precedes the PCA or PLS method, these methods can be applied to analyse very large data without loss of information.[207] Wavelet analysis is also used for compression and smoothing of mono- and two-way IR spectra and NMR spectra.[208]

In some cases, it is necessary to perform a fast signal smoothing in real-time. The Kalman filter is a method developed for this purpose. It can be used, for example, to model the variation of the kinetics during a process. The general idea of the Kalman filter is correction of the model following the process development. As soon as new data become available, the model is supplemented and improved. With the advent of fast and powerful computers, the Kalman filter has become of virtually no use, although some works performed with this tool still appear (see, for example, Ref. 209).

VI. Conclusion

1. Process analytical technology

We have considered the key achievements of chemometrics during the last 15–20 years and their applications in analytical chemistry. Many topical issues and applications both related to and far from analytical chemistry remained beyond the scope of the review. One issue deserves special attention, as it reflects most vividly the trends and prospects of the development of chemometrics and analytical chemistry, in particular, we are speaking about methods for the process analytical technology.

In the 1930s, the American statistician Shewhart [210] proposed to use statistical methods for the control of industrial processes. His idea is very simple: if data on a normally operating industrial process over a long period are collected and statistically processed, then for a controlled operating parameter x, a range can be established, $[x_{\min}, x_{\max}]$, in which the process operates normally. The departure of the parameter x beyond these limits gives warning about some emergency that requires immediate intervention. This method was called statistical process control. It has been successfully used in practice (Shewhart charts). However, subsequently, as the process technology became more complicated, it turned out that each parameter x_i cannot be monitored separately independently of other parameters. This often resulted in faulty decisions such as false alarms, off-spec products, *etc.* The point is that the measured operating parameters $x_1, x_2, \ldots$, are usually interrelated (correlated) with each other and they should be considered together. The situation largely resembles analysis of multichannel chemical data (*e.g.*, spectra). Taking into account this analogy, MacGregor [211] developed a new approach to the solution of this problem, MSPC.[212] He proposed to use the PCA method for the analysis of multivariate data and to construct control limits in the score space by means of the Mahalanobis distance. The idea proved highly fruitful and found numerous supporters. Chemometric specialists developed methods for the multivariate control of batch (for example, biochemical) processes based on the three-way calibration;[24] hierarchical,[213] block [214] and path [82] methods have been developed for the analysis of complex processes. The authors of some studies proposed not only to monitor but also to optimise the processes.[215] These theoretical

results have been used in practice, first of all, in the food [216] and pharmaceutical [217] industries. Systems for the quality control in the production of polymers,[218] non-ferrous metals [219] and semiconductors [41] have been developed.

Thus, a new trend appeared in chemometrics [220] dealing with problems far from the problems of analytical chemistry. However, it was found that sensors and transducers used traditionally in various branches of industry do not provide information necessary for the control of complex (first of all, pharmaceutical) processes. There appeared urgent need for methods of real-time or even *in situ* monitoring of chemical reactions.[221] Traditional analytical methods, first of all, UV [94] and IR [222] spectrometry proved suitable for this purpose. The chemometric methods for the resolution of curves and estimation of kinetic constants from spectroscopic [223] and chromatographic [224] data were found to be useful for monitoring of chemical reactions and processes. A combination of MSPC with analytical monitoring and with process analytical chemistry [225] gave an impetus to the development of a new line of research in analytical chemistry, process analytical technology (PAT).[226] PAT is a system for designing, analysing and controlling the manufacture through timely measurements (*i.e.*, during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality

The major problem faced by modern industry is to ensure high quality of the final product. In view of the increasing global competition and fluctuating market demands, the necessity of effective control of industrial processes in real-time becomes obvious, and chemometrics plays a crucial role in solving this problem. The USA Food and Drug Administration [227] confirmed this role legislatively in September, 2004.

Within the framework of chemometrics, remarkable methods and algorithms have been created; however, they were approved very slowly and reluctantly by the regulating authorities of all developed countries. This is attributable to the fact that ideas of the multivariate approach are more difficult to perceive and visualise than the traditional univariate methods, which are not always able to reflect the full situation. For example, it is much easier to state that the product quality is described by the height of some peak at a particular wavelength than to explain that this quality is determined by whether or not the projection of the whole spectrum falls into a definite area in the PLS-score space found using the Mahalanobis distance, *etc*.

Throughout the whole history of the development of chemometrics, only one regulatory document has been adopted in this field.[228] After the PAT document is approved, chemometrics will necessarily become a legitimate tool for any company willing to follow the FDA guideline principles. In our opinion, approval of this regulatory document would mark a crucial turn in technology and a new industry paradigm, the mission of which is to to build quality into products. A fundamental distinction of this paradigm from the existing one is replacement of the standardisation and unification principles by flexible on-line control of the future product quality at every stage of the manufacture starting from raw material analysis. Consider a simple example: the products of catering facilities that work with the 'standardisation' paradigm have an obviously poorer quality than home-made food, which is produced with flexible on-line control. However, with the introduction of the PAT system, large-scale producers would be able to ensure the same high quality while maintaining large output.

2. Prospects of development

What is the role of analytical chemistry in the formation of this new technological paradigm? What can be chemists' answer to this question? In our opinion, the following trends would prevail in the development of analytics. First, the objects of analysis would become more complex and complicated. The industry requirements would not put peripheral questions such as how much of substance X is present in a sample but general questions of whether a product with the required quality will be obtained from the given raw material or whether the chemical reaction in the given column develops in a right way. Second, methods of analysis will be changed to gain the required data directly in the plant in real-time (*in line*) rather than in the laboratory (*at line*). Third, the amount of multiway and multivariate data will sharply increase. The role of hyphenated and composite methods of analysis will grow. Fourth, the desired chemical information will be deeply hidden in these data; moreover, it will be less formalised, which will call for the use of fine methods for its retrieval. Fifth, the organisation of the analytical experiment will change: instead of analysis of a single sample in one experiment, it will be necessary to use a system approach according to which numerous samples are analysed simultaneously by different methods in an automated mode under different conditions. This large-scale computer-assisted experiment (which already exists, for example, in the microarray approach) will become a routine analytical practice. Sixth, the analytical investigation will be focused on biological objects and biochemical processes and on industrial processes as a whole.

All these trends are already visible in analytical chemistry. The role of analytical chemist will change: he will inevitably become more an analyst than a chemist. The problems tackled by the researcher will be reduced to two key problems. The first is to set up an experiment that could give data suitable for retrieving the required information. The desired information may be a prognosis for the final state of the system rather than a quantitative (concentration) or qualitative (yes/no) result. The second problem is how to retrieve the desired information from the data and interpret it from the utility and quality standpoints. To solve these problems, a researcher must use the experience and tooling of chemometrics. Thus, chemometrics as an inherent part of analytical chemistry will largely determine the trends of its development.§

It should be emphasised that the pronounced extension of the scope of analytical methods should be based on close cooperation of chemometric specialists not only with chemists but also with other scientists, first of all, mathematicians, physicians and biologists.

The authors are grateful to O N Karpukhin (Institute of Chemical Physics, Moscow) and A Yu Bogomolov (EMBL, Hamburg) for valuable advice during preparation of the review and to K Esbensen (Aalborg University Esbjerg, Denmark) for his effort aimed at popularisation of chemometrics in Russia.

References

1. M Sharaf, D Illman, B Kowalski *Chemometrics* (New York: Wiley, 1986)
2. Yu A Zolotov *Analiticheskaya Khimiya: Problemy i Dostizheniya* (Analytical Chemistry: Problems and Achievements) (Moscow: Nauka, 1992)
3. Yu V Granovskii *Vestn. Mosk. Univ., Ser. 2, Khim.* **38** 211 (1997) [a]
4. Yu A Karpov, T M Polkhovskaya *Standartizatsiya i Metrologiya v Metallurgicheskom Proizvodstve* (Standardisation and Metrology in Metallurgy) (Moscow: Moscow Institute of Steel and Alloys, 1989)
5. P Geladi, K Esbensen *Chemom. Intell. Lab. Syst.* **7** 197 (1990)
6. D L Massart *Chemometrics: a Textbook* (New York: Elsevier, 1988)
7. S Wold *Chemom. Intell. Lab. Syst.* **30** 109 (1995)
8. M Blanco, I Villarroya *Trends Anal. Chem.* **21** 240 (2002)
9. B G Osborne, T Fearn *Near Infrared Spectroscopy in Food Analysis* (Harlow, Essex: Longman Scientific and Technical, 1986)

§ When applying the foregoing to the practical training of specialists in chemometrics, note that it would be expedient to develop university classes in this discipline. Obviously, raising an issue of starting the corresponding specialties in the master specialisation and of the formation of Candidate and Doctor Councils is also justified.

10. M Blanco, J Coello, H Iturriaga, S Maspoch, E Rovira *J. Pharm. Biomed. Anal.* **16** 255 (1997)
11. A Espinosa, D Lambert, M Valleur *Hydrocarbon Process.* **74** 86 (1995)
12. T Næs, C Irgens, H Martens *Appl. Stat.* **35** 195 (1986)
13. H Martens, T Næs *Trends Anal. Chem.* **3** 204 (1984)
14. W S Gosset ('Student') *Biometrika* **6** 1 (1908)
15. K Pearson *Philippine Mag.* **2** (6) 559 (1901)
16. R A Fisher *Statistical Methods for Research Workers* (Edinburgh: Oliver and Boyd, 1925)
17. R A Fisher *The Design of Experiments* (Edinburgh: Oliver and Boyd, 1935)
18. V Nalimov *Primenenie Matematicheskoi Statistiki pri Analize Veshchestva* (Application of Mathematical Statictics in the Analysis of Substances) (Moscow: Fizmatlit, 1960)
19. S Wold, K Esbensen, P Geladi *Chemom. Intell. Lab. Syst.* **2** 37 (1987)
20. G Golub, C van Loan *Matrix Computations* (Baltimore: John Hopkins University Press, 1996)
21. P Geladi, H Grahn *Multivariate Image Analysis* (Chichester: Wiley, 1996)
22. B Walczak, D L Massart *Trends Anal. Chem.* **16** 451 (1997)
23. A I Belousov, S A Verzakov, J von Frese *J. Chemom.* **16** 482 (2002)
24. P Nomikos, J F MacGregor *AIChE J.* **40** 1361 (1994)
25. P Geladi, K Esbensen *J. Chemom.* **5** 97 (1991)
26. M Schaeferling, S Schiller, H Paul, M Kruschina, P Pavlickova, M Meerkamp, C Giammasi, D Kambhampati *Electrophoresis* **23** 3097 (2002)
27. M M C Ferreira *J. Chemom.* **18** 385 (2004)
28. I E Frank, J H Friedman *Technometrics* **35** 109 (1993)
29. S Wold, A Berglund, N Kettaneh *J. Chemom.* **16** 377 (2002)
30. G Molenberghs *Biometrics* **61** 1 (2005)
31. A G Shmelev *Vopr. Psikhol.* (5) 34 (1982)
32. H Wold, in *Perspectives in Probability and Statistics* (Sheffield: University of Sheffield, Applied Probability Trust, 1975) p. 117
33. N R Draper, H Smith *Applied Regression Analysis* (New York: Wiley, 1981)
34. O Ye Rodionova, A L Pomerantsev *Kinet. Katal.* **45** 485 (2004) [b]
35. H-L Koh, W-P Yau, P-S Ong, A Hegde *Drug Discov. Today* **8** 889 (2003)
36. A L Pomerantsev, O Ye Rodionova *Chemom. Intell. Lab. Syst.* **79** 73 (2005)
37. L Gribov *Matematicheskie Metody i EVM v Analiticheskoi Khimii* (Mathematical Methods and Computers in Analytical Chemistry) (Moscow: Nauka, 1989)
38. K J Siebert *J. Am. Soc. Brew. Chem.* **59** 147 (2001)
39. K Varmuza, W Werther, F R Krueger, J Kissel, E R Schmid *Int. J. Mass Spectrom.* **189** 79 (1999)
40. G W Johnson, R Ehrlich *Environ. Forensics* **3** 59 (2002)
41. B M Wise, N B Gallagher, E B Martin *J. Chemom.* **15** 285 (2001)
42. R G Brereton *Chemometrics: Data Analysis for the Laboratory and Chemical Plant* (Chichester: Wiley, 2003)
43. N P Komar' *Osnovy Kachestvennogo Khimicheskogo Analiza* (Foundations of Qualitative Chemical Analysis) (Kharkov: Kharkov State University, 1955)
44. L A Gribov, V I Baranov, M E Elyashberg *Bezetalonnyi Molekulyarnyi Spektral'nyi Analiz. Teoreticheskie Osnovy* (Standardless Molecular Spectral Analysis. Theoretical Foundations) (Moscow: Editorial URSS, 2002)
45. M E Elyashberg *Usp. Khim.* **68** 579 (1999) [*Russ. Chem. Rev.* **68** 525 (1999)]
46. B M Mar'yanov, A G Zarubin, S V Shumar *Zh. Anal. Khim.* **58** 1126 (2003) [c]
47. V I Vershinin, B G Derendyaev, K S Lebedev *Metody Komp'yuternoi Identifikatsii Organicheskikh Soedinenii* (The Methods of Computer Identification of Organic Compounds) (Moscow: Nauka, 2002)
48. I G Zenkevich, B Kránicz *Chemom. Intell. Lab. Syst.* **67** 51 (2003)
49. I V Pletnev, V V Zernov *Anal. Chim. Acta* **455** 131 (2002)
50. N M Halberstam, I I Baskin, V A Palyulin, N S Zefirov *Usp. Khim.* **72** 706 (2003) [*Russ. Chem. Rev.* **72** 629 (2003)]
51. V I Dvorkin *Metrologiya i Obespechenie Kachestva Kolichestvennogo Khimicheskogo Analiza* (Metrology and Quality Support of Quantitative Chemical Analysis (Moscow: Khimiya, 2001)
52. Yu G Vlasov, A V Legin, A M Rudnitskaya *Usp. Khim.* **75** 141 (2006) [*Russ. Chem. Rev.* **75** 125 (2006)]
53. A V Kalach, Ya I Korenman, S I Niftaliev *Iskusstvennye Neironnye Seti — Vchera, Segodnya, Zavtra* (Artificial Neuron Networks — Yesterday, Today and Tomorrow) (Voronezh: State Technological Academy, 2002)
54. S P Kazakov, A A Ryabenko, V F Razumov *Opt. Spektrosk.* **86** 537 (1999) [d]
55. V F Razumov, M V Alfimov *Zh. Nauch. Prikl. Foto Kinematograf.* **46** 28 (2003)
56. O Ye Rodionova, K H Esbensen, A L Pomerantsev *J. Chemom.* **18** 402 (2004)
57. E V Bystritskaya, A L Pomerantsev, O Ye Rodionova *J. Chemom.* **14** 667 (2000)
58. A Bogomolov, M McBrien *Anal. Chim. Acta* **490** 41 (2003)
59. US P. 0126892-A1 (2004)
60. S Kucheryavski, V Polyakov, A Govorov, in *Progress in Chemometrics Research* (Ed. A L Pomerantsev) (New York: NovaScience Publishers, 2005) p. 3
61. N M Oskorbin, A V Maksimov, S I Zhilin *Izv. Alt. Univ.* (1) 35 (1998)
62. S V Romanenko A G Stromberg, E V Selivanova, E S Romanenko *Chemom. Intell. Lab. Syst.* **73** 7 (2004)
63. I E Vasil'eva, A M Kuznetsov, I L Vasil'ev, E V Shabanova *Zh. Anal. Khim.* **52** 1238 (1997) [c]
64. D L Massart, B G Vandeginste, L M C Buydens, S De Jong, P J Lewi, J Smeyers-Verbeke *Handbook of Chemometrics and Qualimetrics. Part A* (Amsterdam: Elsevier, 1997)
65. B G Vandeginste, D L Massart, L M C Buydens, S De Jong, P J Lewi, J Smeyers-Verbeke *Handbook of Chemometrics and Qualimetrics. Part B* (Amsterdam: Elsevier, 1998)
66. T Næs, T Isaksson, T Fearn, T Davies *Multivariate Calibration and Classification* (Christerer: Wiley, 2002)
67. R Kramer *Chemometric Techniques for Quantitative Analysis* (New York: Marcel Dekker, 1998)
68. K R Beebe, R J Pell, M B Seasholtz *Chemometrics: a Practical Guide* (New York: Willey, 1998)
69. E R Malinowski *Factor Analysis in Chemistry* (2nd Ed.) (New York: Wiley, 1991)
70. H Martens, T Næs *Multivariate Calibration* (New York: Wiley, 1989)
71. A Höskuldsson *Prediction Methods in Science and Technology* Vol. 1 (Copenhagen: Thor Publishing, 1996)
72. R A Kellner, J-M Mermet, M Otto *Analytical Chemistry. The Appoved Text to the FECS Curriculum Analytical Chemistry* (Weinheim: Wiley-VCH, 2001)
73. B M Mar'yanov *Izbrannye Glavy Khemometriki* (Selected Chapters of Chemometrics) (Tomsk: Tomsk State University, 2004)
74. K Esbensen *Analiz Mnogomernykh Dannykh* (Analysis of Multivariate Data) (Chernogolovka: Institute for Problems of Chemical Physics, Russian Academy of Sciences, 2005)
75. K Esbensen, O Rodionova, A Pomerantsev, O Startsev, S Kucheryavski *J. Chemom.* **17** 422 (2003)
76. O Ye Rodionova *Chemom. Intell. Lab. Syst.* **67** 194 (2003)
77. S Kucheryavski, C Marks, K Varmuza *Chemom. Intell. Lab. Syst.* **78** 138 (2005)
78. L Eriksson, E Johansson, N Kettaneh-Wold, S Wold *Multi- and Megavariate Data Analysis* (Umeå: Umetrics, 2001)
79. E Sanchez, B R Kowalski *J. Chemom.* **2** 247 (1988)
80. A Smilde, R Bro, P Geladi *Multi-way Analysis with Applications in the Chemical Sciences* (Chichester: Wiley, 2004)
81. S Wold, J Trygg, A Berglund, H Antti *Chemom. Intell. Lab. Syst.* **58** 131 (2001)
82. A Höskuldsson *J. Chemom.* **58** 287 (2001)
83. P Geladi, J Burger, T Lestanderet *Chemom. Intell. Lab. Syst.* **72** 209 (2004)
84. G H W Sanders, A Manz *Trends Anal. Chem.* **19** 364 (2000)
85. G E P Box, W G Hunter, J S Hunter *Statistics for Experimenters* (New York: Wiley, 1978)

86. E Z Demidenko *Lineinaya i Nelineinaya Regressii* (Linear and Non-liniar Regressions) (Moscow: Finansy i Statistika, 1981)
87. P Jy *Sampling for Analytical Purposes* (Chichester: Wiley, 1989)
88. W Kleingeld, J Ferreira, S Coward *J. Chemom.* **18** 121 (2004)
89. *Chemom. Intell. Lab. Syst.* **74** (Special Issue) 1 (2004)
90. B Walczak, D L Massart *Chemom. Intell. Lab. Syst.* **58** 15 (2001)
91. P R C Nelson, P A Taylor, J F MacGregor *Chemom. Intell. Lab. Syst.* **35** 45 (1996)
92. H Haario, V-M Taavitsainen *Chemom. Intell. Lab. Syst.* **44** 77 (1998)
93. E F Brin, A L Pomerantsev *Khim. Fiz.* **5** 1674 (1986) [e]
94. S P Gurden, J A Westerhuis, S Bijlsma, A K Smilde *J. Chemom.* **15** 101 (2001)
95. A L Pomerantsev, Doctoral Thesis in Physico-matematical Sciences, Institute of Chemical Physics, Russian Academy of Sciences, Moscow, 2003
96. D A Morales *J. Chemom.* **16** 247 (2002)
97. A de Juan, M Maeder, M Martínez, R Tauler *Chemom. Intell. Lab. Syst.* **54** 123 (2000)
98. B Efron *Ann. Stat.* **7** 1 (1979)
99. *EURACHEM/CITAC Guide, Quantifying Uncertainty in Analytical Measurement* (2nd Ed.) (Lisbon: EURACHEM, 2000)
100. K Faber, B R Kowalski *Chemom. Intell. Lab. Syst.* **34** 283 (1996)
101. A L Pomerantsev *Chemom. Intell. Lab. Syst.* **49** 41 (1999)
102. A Pulido, I Ruisánchez, R Boqué, F X Rius *Trends Anal. Chem.* **22** 647 (2003)
103. V I Vershinin *Accredit. Qual. Assur.* **9** 415 (2004)
104. N M Faber *Chemom. Intell. Lab. Syst.* **64** 169 (2002)
105. N M Faber, R Bro *Chemom. Intell. Lab. Syst.* **61** 133 (2002)
106. A Lorber *Anal. Chem.* **58** 1167 (1986)
107. J Ferré, N M Faber *Chemom. Intell. Lab. Syst.* **69** 123 (2003)
108. R Boqué, N M Faber, F Xavier Rius *Anal. Chim. Acta* **423** 41 (2000)
109. R Boqué; J Ferré, N M Faber, F Xavier Rius *Anal. Chim. Acta* **451** 313 (2002)
110. I Berget, T Næs *J. Chemom.* **18** 103 (2004)
111. D Jouan-Rimbaud, D L Massart, C A Saby, C Puel *Anal. Chim. Acta* **350** 149 (1997)
112. M Meloun, J Militký, M Hill, R G Brereton *Analyst* **127** 433 (2002)
113. J A F Pierna, F Wahl, O E de Noord, D L Massart *Chemom. Intell. Lab. Syst.* **63** 27 (2002)
114. K Faber *Chemom. Intell. Lab. Syst.* **52** 123 (2000)
115. N M Faber, X-H Song, P K Hopke *Trends Anal. Chem.* **22** 330 (2003)
116. E Bouveresse, D L Massart *Vib. Spectrosc.* **11** 3 (1996)
117. F Westad, H Martens *J. Near Infrared Spectrosc.* **8** 117 (2000)
118. M Hubert, S Verboven *J. Chemom.* **17** 438 (2003)
119. H R Keller, D L Massart *Chemom. Intell. Lab. Syst.* **12** 209 (1992)
120. E R Malinowski *J. Chemom.* **6** 29 (1992)
121. P J Gemperline *Anal. Chem.* **58** 2656 (1986)
122. S Wold *Pattern Recogn.* **8** 127 (1976)
123. J-H Jiang, Y Liang, Y Ozaki *Chemom. Intell. Lab. Syst.* **71** 1 (2004)
124. F C Sanchez, B van de Borgaert, S C Rutan, D L Massart *Chemom. Intell. Lab. Syst.* **34** 139 (1996)
125. H Shen, B Grande, O M Kvalheim, I Eide *Anal. Chim. Acta* **446** 313 (2001)
126. W Windig, J Guilment *Anal. Chem.* **63** 1425 (1991)
127. A Bogomolov, M Hachey, in *Progress in Chemometrics Research* (Ed. A L Pomerantsev) (New York: Nova Science Publishers, 2005) p. 119
128. J Diewok, A de Juan, M Maeder, R Tauler, B Lendl *Anal. Chem.* **76** 641 (2003)
129. A Yu Bogomolov, T N Rostovshchikova, V V Smirnov *Zh. Fiz. Khim.* **69** 1197 (1995) [f]
130. H A Seipel, J H Kalivas *J. Chemom.* **18** 306 (2004)
131. S R Crouch, A Scheeline, E S Kirkor *Anal. Chem.* **72** 53 (2000)
132. R I Shrager *Chemom. Intell. Lab. Syst.* **1** 59 (1986)
133. R De Maesschalck, D Jouan-Rimbaud, D L Massart *Chemom. Intell. Lab. Syst.* **50** 1 (2000)
134. J M Andrade, M P Gomez-Carracedo, W Krzanowski, M Kubista *Chemom. Intell. Lab. Syst.* **72** 123 (2004)
135. O Ye Rodionova, L P Houmøller, A L Pomerantsev, P Geladi, J Burger, V L Dorofeyev, A P Arzamastsev *Anal. Chim. Acta* **549** 151 (2005)
136. L X Sun, K Danzer *J. Chemom.* **10** 325 (1996)
137. A J Myles, S D Brown *J. Chemom.* **17** 531 (2003)
138. D González-Arjona, G López-Pérez, A G González *Talanta* **49** 189 (1999)
139. H Mark *Anal. Chem.* **59** 790 (1987)
140. P J Gemperline, N R Boyer *Anal. Chem.* **67** 160 (1995)
141. H L Mark, D Tunnell *Anal. Chem.* **57** 1449 (1985)
142. U Indahl, N S Sing, B Kirkhuus, T Næs *Chemom. Intell. Lab. Syst.* **49** 19 (1999)
143. G Downey, J Boussion, D Beauchene *J. Near Infrared Spectrosc.* **2** 85 (1994)
144. G R Flaten, B Grung, O M Kvalheim *Chemom. Intell. Lab. Syst.* **72** 101 (2004)
145. T Næs, U Indahl *J. Chemom.* **12** 205 (1998)
146. J McElhinney, G Downey, T Fearn *J. Near Infrared Spectrosc.* **7** 145 (1999)
147. S Zomer, R Brereton, J F Carter, C Eckers *Analyst* **129** 175 (2004)
148. V V Zernov, K V Balakin, A A Ivaschenko, N P Savchuk, I V Pletnev *J. Chem. Inf. Comput. Sci.* **43** 2048 (2003)
149. M Sarker, W Rayens *J. Chemom.* **17** 166 (2003)
150. A Herrero, S Zamponi, R Marassi, P Conti, M C Ortiz, L A Sarabia *Chemom. Intell. Lab. Syst.* **61** 63 (2002)
151. R Manne, B-V Grande *Chemom. Intell. Lab. Syst.* **50** 35 (2000)
152. S Bijlsma, A K Smilde *J. Chemom.* **14** 541 (2000)
153. R Bro *Chemom. Intell. Lab. Syst.* **38** 149 (1997)
154. H Kiers *J. Chemom.* **14** 151 (2000)
155. N M Faber, R Bro, P K Hopke *Chemom. Intell. Lab. Syst.* **65** 119 (2003)
156. C A Andersson, R Bro *Chemom. Intell. Lab. Syst.* **52** 1 (2000)
157. F J del Rio, J Riu, F X Rius *J. Chemom.* **15** 773 (2001)
158. R G Brereton *Analyst* **125** 2125 (2000)
159. A Höskuldsson *J. Chemom.* **2** 211 (1988)
160. S de Jong *Chemom. Intell. Lab. Syst.* **18** 251 (1993)
161. B Li, A J Morris, E B Martin *Chemom. Intell. Lab. Syst.* **72** 21 (2004)
162. M Hubert, K Vanden Branden *J. Chemom.* **17** 537 (2003)
163. E Vigneau, M Devaux, M Qannari, P Robert *J. Chemom.* **11** 239 (1997)
164. P Geladi *Chemom. Intell. Lab. Syst.* **60** 211 (2002)
165. L V Kantorovich *Sib. Mat. Zh.* **3** 701 (1962)
166. V M Belov, V A Sukhanov, F G Unger *Teoreticheskie i Prikladnye Aspekty Metoda Tsentra Neopredelennosti* (Theoretical and Practical Aspects of the Uncertainty Centre Method) (Novosibirsk: Nauka, 1995)
167. R Bro *J. Chemom.* **10** 47 (1996)
168. Y Ni, C Huang, S Kokot *Chemom. Intell. Lab. Syst.* **71** 177 (2004)
169. Z P Chen, J Morris, E Martin, R-Q Yu, Y-Z Liang, F Gong *Chemom. Intell. Lab. Syst.* **72** 9 (2004)
170. F M Fernández, M B Tudino, O E Troccoli *Anal. Chim. Acta* **433** 119 (2001)
171. I Garcia, L Sarabia, M C Ortiz, J M Aldama *Anal. Chim. Acta* **515** 55 (2004)
172. Y Bard *Nonlinear Parameter Estimation* (New York: Academic Press, 1974)
173. D M Barry, L Meites *Anal. Chim. Acta* **68** 435 (1974)
174. B Mar'yanov, in *Khimiki TGU na Poroge Tret'ego Tysyacheletiya* (The Chemists of Tomsk State University at the Threshold of the Third Millenium) (Tomsk: Tomsk State University, 1998) p. 48
175. A Berglund, S Wold *J. Chemom.* **11** 141 (1997)
176. A Berglund, N L U Kettaneh, S Wold, N Bendwell, D R Cameron *J. Chemom.* **15** 321 (2001)
177. S Wold *Chemom. Intell. Lab. Syst.* **14** 71 (1992)
178. J Zupan, J Gasteiger *Anal. Chim. Acta* **248** 1 (1991)

179. J Zupan, J Gasteiger *Neural Network for Chemists. An Introduction* (Weinheim: VCH, 1993)
180. W Wu, B Walczak, D L Massart, E Heuerding, F E Erni, I R Last, K A Prebble *Chemom. Intell. Lab. Syst.* **33** 35 (1996)
181. J R M Smits, W J Melssen, L M C Buydens, G Kateman *Chemom. Intell. Lab. Syst.* **22** 165 (1994)
182. W J Melssen, J R M Smits, L M C Buydens, G Kateman *Chemom. Intell. Lab. Syst.* **23** 267 (1994)
183. D B Hibbert *Chemom. Intell. Lab. Syst.* **19** 277 (1993)
184. R Leardi *J. Chemom.* **15** 559 (2001)
185. X Shao, Z Chen, X Lin *Chemom. Intell. Lab. Syst.* **50** 91 (2000)
186. L A Tortajada-Genaro, P Campins-Falcó, J Verdú-Andrés, F Bosch-Reig *Anal. Chim. Acta* **450** 155 (2001)
187. R Bro, A K Smilde *J. Chemom.* **17** 16 (2003)
188. P Kubelka, F Munck *Z. Tech. Phys.* **12** 593 (1931)
189. A Savitzky, M J E Golay *Anal. Chem.* **36** 1627 (1964)
190. P Geladi, D MacDougall, H Martens *Appl. Spectrosc.* **3** 491 (1985)
191. T Isaksson, B Kowalski *Appl. Spectrosc.* **47** 702 (1993)
192. J Trygg, S Wold *J. Chemom.* **17** 53 (2003)
193. S Wold, H Antti, F Lindgren, J Öhman *Chemom. Intell. Lab. Syst.* **44** 175 (1998)
194. T Fearn *Chemom. Intell. Lab. Syst.* **50** 47 (2000)
195. A Höskuldsson *Chemom. Intell. Lab. Syst.* **55** 23 (2001)
196. Q Guo, W Wu, D L Massart, C Boucon, S de Jong *Chemom. Intell. Lab. Syst.* **61** 123 (2002)
197. M Forina, S Lanteri, M C Cerrato Oliveros, C Pizarro Millan *Anal. Bioanal. Chem.* **380** 397 (2004)
198. R Leardi, R Boggia, M Terrile *J. Chemom.* **6** 267 (1992)
199. J H Kalivas *Anal. Chim. Acta* **505** 9 (2004)
200. N Benoudjit, E Cools, M Meurens, M Verleysen *Chemom. Intell. Lab. Syst.* **70** 47 (2004)
201. U Indahl, T Næs *J. Chemom.* **18** 53 (2004)
202. R N Feudale, N A Woody, H Tan, A J Myles, S D Brown, J Ferré *Chemom. Intell. Lab. Syst.* **64** 181 (2002)
203. E L Sulima, V A Zubkov, L A Rusinov, in *Progress in Chemometrics Research* (Ed. A L Pomerantsev) (New York: Nova Science Publishers, 2005) p. 196
204. J-H Jiang, Y Ozaki, M Kleimann, H W Siesler *Chemom. Intell. Lab. Syst.* **70** 83 (2004)
205. P W Hansen *J. Chemom.* **15** 123 (2001)
206. C K Chui *Introduction to Wavelets* (New York: Academic Press, 1992)
207. J Trygg, S Wold *Chemom. Intell. Lab. Syst.* **42** 209 (1998)
208. S-P Reinikainen, in *Progress in Chemometrics Research* (Ed. A L Pomerantsev) (New York: Nova Science Publishers, 2005) p. 21
209. Y Pan, C K Yoo, J H Lee, I-B Lee *J. Chemom.* **18** 69 (2004)
210. W A Shewhart *Economic Control of Quality of Manufactured Product* (New York: Van Nostrand, 1931)
211. J MacGregor, Th Kourti *Contr. Engin. Pract.* **3** 403 (1995)
212. A L Pomerantsev, O Ye Rodionova *Metod. Menedzh. Kachest.* **6** 15 (2002)
213. Th Kourti, J MacGregor *Chemom. Intell. Lab. Syst.* **28** 3 (1995)
214. J A Westerhuis, Th Kourti, J F Macgregor *J. Chemom.* **12** 301 (1998)
215. A L Pomerantsev, O Ye Rodionova, in *Progress in Chemometrics Research* (Ed. A L Pomerantsev) (New York: Nova Science Publishers, 2005) p. 209
216. R Bro *Chemom. Intell. Lab. Syst.* **46** 133 (1999)
217. J Gabrielsson, N-O Lindberg, T Lundstedt *J. Chemom.* **16** 141 (2002)
218. C K Yoo, J-M Lee, P A Vanrolleghem, I-B Lee *Chemom. Intell. Lab. Syst.* **71** 151 (2004)
219. M Baroni, P Benedetti, S Fraternale, F Scialpi, P Vix, S Clementi *J. Chemom.* **17** 9 (2003)
220. H Martens, M Martens *Multivariate Analysis of Quality: an Introduction* (Chichester: Wiley, 2001)
221. R M Dyson, M Hazenkamp, K Kaufmann, M Maeder, M Studer, A Zilian *J. Chemom.* **14** 737 (2000)
222. K Pöllänen, A Häkkinen, S-P Reinikainen, M Louhi-Kultanen, L Nyström *Chemom. Intell. Lab. Syst.* **76** 25 (2005)
223. T J Thurston, R G Brereton, D J Foord, R E A Escott *Talanta* **63** 757 (2004)
224. E Bezemer, S C Rutan *Chemom. Intell. Lab. Syst.* **59** 19 (2001)
225. J Workman Jr, K E Creasy, S Doherty, L Bond, M Koch, A Ullman, D J Veltkamp *Anal. Chem.* **73** 2705 (2001)
226. S P Gurden, E B Martin, A J Morris *Chemom. Intell. Lab. Syst.* **44** 319 (1998)
227. *Guidance for Industry PAT — a Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance* U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Veterinary Medicine (CVM) Office of Regulatory Affairs (ORA), September 2004, Pharmaceutical CGMPs
228. *ASTM Standard E1655* Standard Practices for Infrared Multivariate Quantitative Analysis, 1997

[a] — *Moscow Univ. Bull. (Engl. Transl.)*
[b] — *Kinet. Catal. (Engl. Transl.)*
[c] — *J. Anal. Chem. (Engl. Transl.)*
[d] — *Opt. Spectrosc. (Engl. Transl.)*
[e] — *Chem. Phys. Rep. (Engl. Transl.)*
[f] — *Russ. J. Phys. Chem. (Engl. Transl.)*