Anal Bioanal Chem (2017) 409:5891–5899
DOI 10.1007/s00216-017-0517-1

FEATURE ARTICLE

Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools

Richard G. Brereton[1] · Jeroen Jansen[2] · João Lopes[3] · Federico Marini[4] · Alexey Pomerantsev[5] · Oxana Rodionova[5] · Jean Michel Roger[6] · Beata Walczak[7] · Romà Tauler[8]

Received: 6 March 2017 / Revised: 23 June 2017 / Accepted: 10 July 2017 / Published online: 3 August 2017

Abstract Chemometrics has achieved major recognition and progress in the analytical chemistry field. In the first part of this tutorial, major achievements and contributions of chemometrics to some of the more important stages of the analytical process, like experimental design, sampling, and data analysis (including data pretreatment and fusion), are summarised. The tutorial is intended to give a general updated overview of the chemometrics field to further contribute to its dissemination and promotion in analytical chemistry.

Keywords Chemometrics · Experimental design · Sampling · Data preprocessing · Projection methods · Data fusion

All participants belong to the chemometrics study group of the Division of Analytical Chemistry of EuCheMS.

✉ Romà Tauler
Roma.Tauler@idaea.csic.es

[1] School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, UK

[2] Institute for Molecules and Materials, Radboud University, Postvak 61, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

[3] Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Av. Prof. Gama Pinto, 1649-003 Lisbon, Portugal

[4] Department of Chemistry, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Rome, Italy

[5] Institute of Chemical Physics RAS, 4, Kosygin Str, 119991 Moscow, Russia

[6] Irstea, UMR ITAP, 361 Rue Jean-François Breton, 34000 Montpellier, France

[7] Institute of Chemistry, University of Silesia , 40-006, Katowice, Poland

[8] IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

Introduction (historical perspective)

Chemometrics in analytical chemistry originated from several sources.

An important historic catalyst was quantitative analytical chemistry. Although early acknowledged reports stretch back to the sixteenth century, with attempts to assay the amount of gold and silver, it was not until the early twentieth century that quantitative analysis became a widespread discipline. All analytical chemists are aware of concepts of precision, accuracy, errors (or uncertainty), which gradually coevolved together with quantitative analytical chemistry.

By 1949 Mandel describes the use of least squares regression, experimental designs and analysis of variance, ANOVA [1] in analytical chemistry, nearly 25 years before the word "chemometrics" was invented. Over many years, statistical methods have always been a cornerstone of modern analytical chemistry.

The second catalyst, multivariate methods, was first reported in a modern context of physico-analytical chemistry as a method for determining the number of components in spectra of mixtures in the early 1960s [2, 3]. However, these early pioneering papers, appearing throughout this decade, were primarily written from the point of view of theoretical chemistry. Determining the number of components in a spectrum of mixtures was viewed as a problem of similar interest to that of obtaining a quantum mechanical model for the lines in a spectrum. These methods however did not immediately reach laboratory-based analysts with more limited access to computing power.

Another, separate, influence over multivariate methods came via applied statistics following on from pioneers including Pearson and then Fisher [4] whose work in the 1920s and 1930s defined much of our modern thinking and terminology in multivariate analysis. The early work of these applied statisticians was primarily in agricultural statistics, although applications widened to psychology and finance over that period. Ideas such as principal component analysis, factor analysis and discrimination were developed.

Only in the 1970s did the two strands of multivariate thinking, from physical chemistry and from applied statistics, start to converge. Terminology in both areas was quite different and most of our modern terminology arises from statisticians. Departmentalisation of academic research and publishing at the time meant that mainstream statisticians rarely encountered quantitative chemists, and with separate libraries and usually buildings, were rarely aware of each other's work.

A third and important catalyst in the 1960s was the growth of available computer power. Most computers accessible to scientists up to the late 1970s were large off-line mainframes, programmed in languages such as Fortran. Originally only mathematicians had good access to computers. Many of the multivariate methods were of very limited applicability without reasonable computer power. However, as the decades progressed, it was increasing possible for applied scientists to get access to computers. Crystallography was a major early application of computers to the analysis of instrumental chemical data.

In the late 1960s, there started to emerge an interest in machine learning. A strong influence had been the NASA moon mission which led organic chemists to develop the area of artificial intelligence for structure elucidation [5]. This led to a variety of spin-off projects within chemistry, of which the pioneering work of Kowalski, Jurs and Isenhour introduced computerised pattern recognition to analytical chemists [6].

During the 1970s, the three themes of statistics in analytical chemistry, multivariate statistics and computing started to converge. The first paper to use the name chemometrics (in Swedish) was published by Wold but was on the topic of cubic splines [7]. However, it took about a decade for this term to become widespread. Many of the other pioneers used the term "chemical pattern recognition" instead.

By the 1980s, several events provided visibility for the subject. Of particular importance was a NATO sponsored meeting held in Cosenza, Italy, where many of the early experts presented their work [8]. After this, the name "chemometrics" took off. There were conferences, journals, software packages and books that became widespread. In Europe, several research groups, mostly in analytical chemistry, embraced the field of chemometrics, with Massart and colleagues producing one of the earliest comprehensive texts [9].

The early applications over this period were primarily in quantitative analytical chemistry such as NIR calibration, HPLC resolution and UV/Vis deconvolution. There was also a growing application in the area of Multivariate Statistical Process Control.

In the twenty-first century, another revolution occurred—the rapid growth of cheap computing power, allowing powerful algorithms that in the past took hours on desktops, or required access to mainframes, to become routine tools for the laboratory chemist. Hand in hand with this was the growth of rapid, automated, instruments so large datasets (often called megavariate data) could be generated, using approaches such as hyphenated and multidimensional chromatography or NMR.

This data explosion meant that chemometrics was no longer primarily focused on improving the quantitative performance of analytical instruments. Pattern recognition [8–10] became a widespread tool. Applications included biomedical data, especially metabolomics [11] but also food chemistry [12] as well as more recently developing areas including forensics and cultural heritage studies among others. Chemometrics has developed as a widespread tool for the applied analytical chemist as well as a more theoretical method to assist the improvement and development of instrumental methods.

Chemometrics has moved far from the original vision of the pioneers of the 1960s who would have envisaged it very much as restricted to a tool from physical chemistry improving quantitative analysis, into a widespread technique integrated into many applications of applied laboratory-based analytical science. Figure 1 displays the workflow of chemometrics, from data, information to knowledge and the usual steps involved.

Experimental design

Early uses of statistical experimental design appeared in the scientific literature of the eighteenth and nineteenth centuries, introducing concepts such as randomization and blocking, but it was not until the 1920s that statisticians started to introduce a formalised approach. R.A. Fisher and colleagues at Rothamsted Research Centre in the UK introduced many of the concepts that we regard as the building blocks of modern statistical thinking and his 1935 book "The Design of

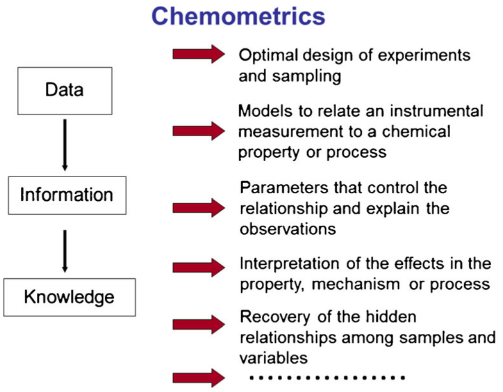

Chemometrics
Data
Information
Knowledge
Optimal design of experiments and sampling
Models to relate an instrumental measurement to a chemical property or process
Parameters that control the relationship and explain the observations
Interpretation of the effects in the property, mechanism or process
Recovery of the hidden relationships among samples and variables
....

Fig. 1 Chemometrics workflow and steps involved

Experiments" is the first comprehensive treatise of the area [13].

In the 1940s, statistical principles were gradually being adapted by the chemical manufacturing industry and a new generation of statisticians promoted this thinking, with emphasis on efficient process optimisation. G.E.P. Box was a notable leader, originally working within Imperial Chemical industries, ICI, until he moved to the USA. His book together with Hunter and Hunter [14] is widely regarded as a benchmark for the modern era. As the 1970s passed, readily available computing power started to become a tool of the scientist, and so an understanding of statistical experimental design was no longer the exclusive province of the professional statistician. Analytical chemists became interested, for example to tackle problems of HPLC optimisation or improving extraction procedure, where many factors influenced the desired outcome. S.N. Deming was a powerful advocate [15] within the analytical literature of the 1970s and 1980s. Using user-friendly packages and spreadsheets, common designs, originally requiring complex mathematical calculations, became widely available and understood.

There are many motivations for statistical experimental design. One of the foremost among analytical chemists is that different factors interact, and so changing one factor at a time can lead to false optima. For example, if a process is dependent on both pH and temperature, to first keep the temperature constant and then change the pH, then use this optimum pH and change the temperature, may not lead to the true optimum. This is because the temperature-dependent behaviour differs according to pH and these factors cannot be considered independently. The traditional approach is called One Factor At-a Time (OFAT). In fact, OFAT can be used successfully in science, for example, if we want to study Newton's second law, a good idea would be to keep the mass of an object constant and then change the force, or to keep the force constant and change the mass, in both cases to measure acceleration. However, for more complex cases where the exact parametric relationship is unknown (and would be too time-consuming to study), or where we believe factors interact, we need a formalised set of statistical rules.

Statistical experimental designs (or Design of Experiments, DoE) are also needed to minimise experimental time. For example, if we wanted to study the effect of pH and temperature on a process, we could study all possible combinations of 10 temperatures and 10 pH values in 100 experiments, but this may be impracticable. In cases such as optimising an extraction, where there may be 10 or more factors, we would not have the resources to perform 10^{10} or 10 billion experiments. Hence, we need formalised approaches to safely get to our answer with a short amount of effort.

Formalised designs follow several steps. Firstly, identify the factors that are likely to influence an experimental outcome, and also the response(s) you may be interested in. Then, code the factors, which involves transforming physical variables such as temperatures onto a common mathematical scale, typically +1 for high and −1 for low, over the experimental range. Third, perform a set of experiments, usually using a well-established statistical protocol, such as a factorial design, or a central composite design or a mixture design. In some cases, we already know which factors are significant, but in other cases, we first have to 'screen' the factors, to reduce them to perhaps 3 or 4 that are really interesting: under such circumstances the initial step might involve a Plackett Burman or partial factorial design [13–15], prior to detailed modelling of the influence of a small subset of factors. The fourth stage involves modelling the data, which could for example provide an optimum or a quantitative model or both, as desired.

Analytical chemists tend to be less interested in important concepts such as stratification, blocking or randomisation, which are hugely important in biology, psychology, social sciences and so on, primarily because a laboratory experiments can be controlled quite effectively. For example, a reaction using specified conditions should give very similar results anywhere in the world as almost everything is known about the reaction conditions. For a biological experiment, this will not usually be so, as genetics, phenotypes, environment etc. cannot be perfectly controlled so we have to be careful about evening out these unknown variations. Hence, a book aimed at a chemist may have a different flavour to that aimed in medicine. Nevertheless, it is useful that analytical chemists have some consideration of these concepts, for example, a GCMS instrument may vary with time and we do not wish to confound these factors with the ones we are interested in.

Most literature over the past century has been concerned with univariate responses. If more than one is to be studied, traditionally each response is approached independently to reach a consensus. However, there has been considerable interest in multivariate responses over the past decade. This has led to the application of a variety of methods to tackle cases where both the factors (or x block) and responses (or y block) are multivariate, including multilevel methods, ANOVA simultaneous component analysis (ASCA) [16], ANOVA principal component analysis [17] (ANOVA-PCA) and ANOVA-target projection [18]. Such approaches are gaining acceptance in metabolomics where there may be complex responses from designed experiments.

Sampling

One thing chemometricians know for a long time is that proper sampling is fundamental for achieving a good description of the system being analysed. Why are then sampling issues so often neglected in chemometric-based published material? It is true that it is actually very difficult to find in the literature papers using chemometrics at different levels that properly describe issues directly related with sampling. There are many reasons behind this evidence but none can be explained by the lack of scientific background or theoretical models as there is extensive literature in this field [19, 20]. Geladi and Esbensen [21] warn against the indiscriminate use of chemometric methods focussing only on algorithms, and that do not consider the importance of population assumptions, of sampling plans, of the theory of sampling and that perform an indiscriminate use of resampling strategies (e.g. cross-validation) to compensate for inadequate sampling strategies. Sampling may be required for multiple purposes: (1) describing an object/material/lot globally or in detail; (2) monitoring an object or a system and (3) controlling a system or process. As stated in Petersen et al. (2005), ensuring system representativeness prior to analytical/data analysis is critical and known for more than 60 years. Errors derived from non-appropriate sampling procedures can exceed the analytical errors by 10–1000 times. It is also true that the growth of chemometrics since the 80s had a particular influence on the sampling strategies, where tackling correlation within objects allows a more effective sampling process.

Very often, data used for chemometric analysis are produced according to good practices such as resourcing to experimental design (DoE). This usually suffices to justify the quality of the data and the effective variance encompassed in the population of samples that derive from that DoE. Very seldom however, there is reference to the sampling procedures which are totally distinct from the DoE definition and analytical methodologies used to analyse the produced samples. Moreover, DoE dedicated to define optimal sample plans are proposed in [21, 22]. How samples will actually describe the heterogeneous material from which they were derived is still neglected and very often explain the incongruences in the reported results. Indeed very often there are references to homogeneous (e.g. a gas inside a container or an aqueous solution of some analyte) and heterogeneous (e.g. soil sample or soybean beans) materials being the latter more difficult to sample. Real homogeneous materials are rare or even should not be defined as such and all materials or systems should be treated as heterogeneous. Heterogeneous materials may be characterised by some random distribution of the property (when samples can be intentionally produced) or more often characterised by a correlated distribution of random properties [19, 23].

The theory of sampling (TOS) was in many ways a consequence of the works by Pierre Gy [19] and it is now perfectly established and fitted with all aspects that a scientific theory requires. The theory of sampling, that will be not described in detail here, refers to a scientific-based process of obtaining representative samples from a heterogeneous material (or lot) [22], (which intrinsically is a mass reduction step). TOS provides the fundamental background for an appropriate sampling procedure. It gives the tools to guarantee representative samples. TOS defines specimens as samples that do not represent the material from which they were collected or extracted. In TOS, lot is the sampling target or the material subject to sampling (other TOS definitions can be found in [22]). TOS states that only a full inspection of a sampling process can ensure representative samples. It distinguishes truly representative samples (those derived from a validated sampling process or *samples*) from incorrect samples (undocumented samples also designated by *specimens*). In general, the TOS intends to minimise the sampling error (or grade deviation) defined as the difference between the mass fraction of the analyte in the sample and the corresponding mass fraction in the lot divided by the latter. According to the TOS, the Fundamental Sampling Principle (FSP) encompasses the criteria for ensuring correctness of the sampling procedure, namely stating that the sampling process should be accurate (zero bias). Sampling errors or correct errors (that cannot in general be prevented at all) are segregated in fundamental sampling error and grouping and segregation error. Incorrect errors exist when the fundamentals of the TOS are not followed and therefore can be minimised if the good practices defined in the TOS are followed (a comprehensive explanation of these errors is available in [22, 23]). The most frequent operations encompass heterogeneity characterisation, mixing, use composite sampling and reduction in particle size. In the context of the chemometric or statistic literature, this sampling process should be referred as *physical sampling*.

Sampling or resampling is referred in the chemometric literature with different meanings. Notably more often, this terminology is found in the literature referring to a methodology that resources on a training set to optimise models' structure, access models' prediction performance, estimating uncertainty among others. This is *statistical sampling* where one draws individuals from a population assuming that they are similar in all aspects (except for the amount of some property of interest). S*tatistical sampling* is totally different from the *physical sampling* where objects are materially drawn from a pool of objects originated from the system under study and therefore constitute a subpart of some lot material. In particular, statistical sampling or resampling is often used and designated as cross-validation. Cross-validation is intended to optimise models resourcing on a training set towards optimal performance for future datasets. In [21], it is shown that the extensive and generalised use of cross-validation methods based on segmented procedures does not find a theoretical justification for the purpose. Indeed, cross-validation is known to be suboptimal for simulating a model performance in the presence of a test set. However, this is being completely ignored by the community as the methodology keeps being used indiscriminately regarding application or chemometric method. This approach simply ignores several fundamental issues, of which ignoring the sampling variance is probably the most critical.

The TOS is critical within the principles of proper statistical validation [21, 22] as it provides the tools to incorporate in models all variance-related factors, fundamentally addressing the critical so-called incorrect sampling errors that cannot in general be corrected with statistical/chemometric methods. A thorough discussion on this issue is given by Geladi & Esbensen (2010) [21], which demonstrates undoubtable insufficiency of cross-validation methods and highlights the critical importance of following TOS principles for proper validation (e.g. using external datasets).

It is known from TOS that it is not possible to keep constant the sampling bias when dealing with heterogeneous materials. Additionally, it is not generally possible to compensate this variability with statistical methods, which is precisely what it is found in the literature where statistical or chemometric methods are being used precisely for this function [24]. The paper from Dardenne et al. (2000) [25] illustrates very well using different case studies that the performance of chemometric models (they used from multiple linear regression to feedforward artificial neural networks) is much more dependent on the quality of the data than of the method itself.

Despite the existence of sufficient and accessible information in the literature for many years regarding good practices for performing sampling and alerts concerning the result of neglecting appropriate sampling much more effort must be done. Even if training events on the TOS are available periodically, essentially in the form of post-graduate courses, these concepts should be better integrated in undergraduate courses. This will considerably impact transversal areas like chemistry, health sciences, engineering, environmental sciences, among many others.

Data preprocessing

Experimental data should be often pretreated. In many cases, the necessary transformations are determined by the type of instrument used for data acquisition. For example, warping methods are used for peak alignment. They *are* employed for NMR and chromatographic data. Data preprocessing implies in many circumstances signal processing and signal extraction procedures (like filters and wavelets) and they should be also considered an integral part of chemometrics. Multiplicative scatter correction (MSC) is used to eliminate the light scattering from the near infrared spectra. Standard normal variate (SNV), baseline correction, and de-trending are applied for various types of spectral data. More general transformations, such as smoothing and differentiation, are also used for various multivariate data. The above-mentioned transformations are simultaneously applied to all variables of an object or sample, i.e. the entire spectrum or chromatogram is transformed according to the selected method. These transformations are referred to as row-wise methods.

Column-wise methods represent another type of preprocessing. They are applied to all objects or samples of every variable. Centering and weighting/scaling are the most popular methods. Column-wise centering is applied in more than 90% of applications, because it shifts the zero of the coordinate system to the center of the multivariate data cloud.

Column-wise scaling is mainly applied when variables are collected from various sources and/or expressed in various units, e.g. when dealing with process control variables or multicomponent ICP-MS trace element analysis. Here, column-wise scaling by standard deviation of the corresponding column brings all variables to a common base and enables simultaneous analysis of variables of a different kind.

Thus, the main objectives of data preprocessing are (1) elimination of artefacts caused by a specific instrument and sample geometry/condition; (2) data de-noising; (3) pretreatment of a raw data set in a way that makes it suitable for further data processing. The subject of data pretreatment has been covered in detail in different works such as [26–29].

It is important to emphasise the following issue. When gathering data for further multivariate analysis, it is strongly recommended to make sure that all results are obtained under the same experimental set-ups, unless the parameters of the experiment are considered to be variables of interest. For example, it is recommended to conduct spectra acquisition in a consistent range and resolution for all samples; it is recommended to keep the value of pH on the same level for all mixtures, etc. Otherwise, the collected data should be additionally re-sampled with extra noise introduced in the data prior to conducting simultaneous analysis.

Projection methods

Data analysis tools can be classified as univariate, multivariate or megavariate, depending on the number and amount of variables considered in the analysis of a single sample at a time. Multivariate and megavariate chemical data, as well as methods used to handle and analyse them, are the realm of chemometric methods. They can also be classified as first-, second-, third- or higher-order methods [30] using notation from tensor algebra, or, similarly, taking into account the number of directions, ways or modes of measurement, as zero-, one-, two- three- or multiway or multimode methods, using notation from other related data analysis fields, such as psychometrics. Typical cases are the analysis of two-way, or second-order, data obtained for a sample using chromatographic methods coupled with spectroscopic multichannel detection, or the analysis of spectral data collected for a set of samples. More complex data structures are obtained when more than two directions are collected during a particular analysis producing data cubes, such as data from excitation-emission fluorescence spectroscopic analysis of a set of samples.

Many chemometric tools can be introduced as data projection linear methods [30], which compress raw data, uncover hidden correlations, and separate useful information from noise. Projection methods provide a very intuitive and visual approach for data analysis. Factor analysis [31] and principal component analysis (PCA) [32, 33] are some of them and they represent the cornerstone of the majority of tools used in chemometric exploratory analysis. PCA transforms a set of data with correlated variables into a set of uncorrelated principal components (PCs), which are obtained as linear combination of initial variables. PCs are calculated in such a way that the maximum portion of variance is explained by the first principal component and, progressively, smaller shares of variance are explained by each subsequent component. The PCA model is presented by an orthogonal decomposition of a data matrix, $\mathbf{X} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E}$, with a specified number of principal components. The model consists of a structural part, $\mathbf{TP}^{\mathbf{T}}$, and an error part, **E**. The structural part (information) is intended to be used for interpretation or prediction, whereas the error part (noise) should be balanced for the model to be reliable. One important aspect is the presentation of data complexity using a minimal number of components that explain the experimental data within the limits imposed by experimental error.

PCA models are developed in a way that the structure of data matrix X can be understood better than by just looking at the raw data. **T** matrix, called the scores matrix, presents objects in a new, reduced space. The scores plot, also called "the map of samples", shows the inter-location and inter-connections of samples under investigation and may be used for unsupervised data clustering. **P** matrix, which is called the loadings matrix, reflects the importance of each variable in the projection. The loadings plot, also called "the map of variables", shows the influence and inter-connections of the variables in the data set. The axes defined by PCA may be rotated to enhance the interpretability of the data variance sources, for instance in terms of their possible physical nature, and several procedures can be found in the literature [31, 32].

In general, presentation of the original matrix **X** as a product of two matrices (structure part, $\mathbf{TP}^{\mathbf{T}}$) is called bilinear decomposition. Other types of projection methods have been proposed for similar bilinear decompositions of multivariate data sets. Two of them have become popular and widely used in the field of chemometrics. One of them is independent component analysis (ICA) [34] and the other one is multivariate curve resolution (MCR) [35]. In ICA, the components are extracted using the criterion of their independency. ICA and its variants can provide more interpretable projections in comparison with PCA, in case additional constraints like non-negativity are implemented.

MCR is a family of methods used to solve a ubiquitous problem of mixture analysis. During the last 40 years, the MCR methods [35] have slowly evolved as powerful tools used to investigate data of (partially) known chemical origin. The goal of MCR is to decompose mixed raw data into a bilinear decomposition of physically understandable pure component profiles, for example, a product of a matrix of pure concentrations and a matrix of pure spectra. The possibility to resolve a particular multicomponent system without ambiguities depends on many circumstances, including application of constraints related with the physical nature of the factors or components, with their selectivity, and with the previous knowledge of the system under study.

It is worth emphasising that the bilinear decompositions expressed in the different methods, either PCA, ICA or MCR, are similar, but their objectives and the ways how these decompositions are performed are different.

Analytical data sets of higher-order (multiway or multimode) structures can be investigated using unfolding methods, which present multiway data in a simplified matrix mode, or by mean of a direct multilinear decompositions [36], using such methods as Tucker 3 and PARAFAC [37]. For example, the PARAFAC method has been successfully applied in the mixture analysis of fluorescent compounds or in environmental source apportionment studies over time. Multiway data analysis is still a growing chemometric subfield which finds more and more applications due to the increasing complexity of data sets, and to the advent of the data fusion and big data analysis new paradigm.

Data fusion

Data fusion methods are nowadays the subject of active research in computational statistics and chemometrics [38]. These methods allow for the simultaneous analysis of datasets coming from different analytical platforms, omic levels, organisms or sample types. From a chemometrics point of view, data fusion strategies can be applied at low-level, mid-level and high-level [39]. High-level fusion (i.e. integration) implies optimal preprocessing and modelling procedures for each data block separately. The outputs of the different models are then jointly evaluated to provide a global overview. In the case of high-level data fusion methods, each dataset is analysed individually using traditional chemometric methods for feature detection [40], before feature integration for a global interpretation. In contrast, low-level and mid-level fusion strategies aim to combine original raw data blocks to obtain later an improved joint interpretation. Low-level fusion generates big size fused data with a large number of variables; whereas mid-level fusion is based on a previous dimensionality compression of data blocks, where only a reduced number of new variables (either the most relevant or the latent variables) from each data block are fused and jointly interpreted. The methods found in the literature for mid-level data fusion try to identify the common and specific variance coming from each one of the analysed blocks after a feature selection to reduce the size of the dataset, like GSVD (generalised singular value decomposition) [41], O2PLS (two-way orthogonal projections to latent structures) [42] as well as OnPLS (multiblock orthogonal projections to latent structures) [43], DISCO-SCA (distinctive and common components with simultaneous component analysis) [44], JIVE (joint and individual variation explained) [45], CMTF (coupled matrix and tensor factorization) [40] and CCSWA (common components and specific weights analysis) [46] methods. Some of these mid-level data fusion methods can also be used for low-level data fusion depending on the raw data characteristics (appropriate block scaling often is required). Finally, other methods are also available for direct low-level data fusion like the multivariate curve resolution alternating least squares (MCR-ALS) [35, 47] method. MCR-ALS allows the joint analysis of multiple datasets from different samples (experiments) or techniques, or from different samples and techniques simultaneously.

Concluding remarks

The increasing number of hyphenated and multidimensional analytical instruments and of all type of chemical measurement devices, providing huge amounts of analytical (big) data and information about complex natural samples require the use of more advanced chemometric data analysis tools. Chemometrics has emerged in the last years as a very successful data analysis approach in the Chemistry field, reaching multiple milestones. Chemometrics has been spread over a large number of applications, especially in analytical sciences, where it has penetrated with force, revolutionising most of the analytical process steps and contributing at the same time, to the solution of more involved and difficult analytical problems, related in many circumstances to new challenges and societal needs.

In this first part of this feature article on chemometrics, several fundamental aspects of this discipline have been covered including sampling, experimental design, data preprocessing and data fusion strategies, and projection methods for data exploration and factor analysis. Many of these aspects are closely related with analytical chemistry and its goals. In the second part of this tutorial, other aspects will be covered, like regression methods, method validation, some successful applications of chemometric methods and the future perspectives of this discipline.

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

References

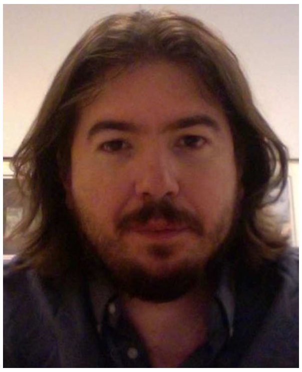
1. Mandel J. Statistical methods in analytical chemistry. J Chem Educ. 1949;26:534–9.
2. Weber G. Enumeration of components in complex systems by fluorescence spectrophotometry. Nature. 1961;190:27–9.
3. Wallace RM. Analysis of absorption spectra by multicomponent systems. J Phys Chem. 1960;64:899–901.
4. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.
5. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J. Applications of artificial intelligence for organic chemistry: the DENDRAL project. New York: McGraw-Hill; 1980.
6. Kowalski BR, Jurs PC, Isenhour TL, Reilly CN. Computerized learning machines applied to chemical problems-multicategory pattern classification by least squares. Anal Chem. 1969;41:695–700.
7. Wold S. Spline functions, a new tool in data-analysis. Kem Tidskr. 1972;3:34–7.
8. B.R. Kowalski (editor), Chemometrics, mathematics, and statistics in chemistry. NATO ASI Series C, Mathematical and Physical Sciences. Vol. 138 D., 1984, Reidel Publishing Company: Dordrecht.
9. D.L. Massart, B.G.M.Vandeginste, S.N.Deming, Y. Michotte and L.Kaufman. Chemometrics: a textbook., Elsevier, Data Handling in Science and Technology, Volume 2, Amsterdam 1988.
10. Brereton RG. Chemometrics for pattern recognition. Chichester: Wiley; 2009.
11. van der Greef J, Smilde AK. Symbiosis of chemometrics and metabolomics: past, present, and future. J Chemometrics. 2005;19:376–86.
12. Marini F, editor. Chemometrics in food chemistry. Amsterdam: Elsevier; 2013.
13. Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
14. Box GEP, Hunter WG, Hunter JS. Statistics for experimenters. New York: Wiley; 1978.
15. Deming SN, Morgan SL. Experimental design: a chemometric approach. Amsterdam: Elsevier; 1987.
16. J.J.Jansen, H.C.J.Hoefslood, R.J.Lalmers, J. van der Greef, M.E. Tiemmerman, A.K. Smilde, Anova simultaneous component analysis, (ASCA): a new tool for analysing designed metabolomics data. Bioinformatics. 2005, 3043–3048.
17. Harrington PB, Viera NE, Espinoza J, Nien JK, Romero R, Lergeyt AL. Analysis of variance-principal component analysis: a soft tool for proteome discovery. Anal Chim Acta. 2005;544:118–27.
18. F.Marini, D. de Beer, E. Joubert and B. Walczak, Analysis of variance designed chromatographic data sets: the analysis of variance-target projection approach, J Chromatogr. 2015, 94–102.
19. Gy PM. Sampling for analytical purposes. The Netherlands: John Wiley and Sons; 1998.
20. Einax JW, Zwanziger HW, Geis S. Sampling and sampling design. In: Chemometrics in environmental analysis. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA; 1997. p. 95–137.
21. Esbensen KH, Geladi P. Principles of proper validation: use and abuse of re-sampling for validation. J Chemom. 2010;24:168–87.
22. Petersen L, Minkkinen P, Esbensen KH. Representative sampling for reliable data analysis: theory of sampling. Chemom Intell Lab Syst. 2005;77:261–77.
23. Petersen L, Esbensen KH. Sampling in practice: a tos toolbox of unit operations. In: Pomerantsev A, editor. Progress in chemometrics research. US: Nova Science Publishers; 2005.
24. G. Kateman Chemometrics—sampling strategies, pp. 43–62. In: Chemometrics and species identification, Topics in current chemistry, Vol.141, Springer Verlag, FRG, 1987.

25. Dardenne P, Sinnaeve G, Baeten V. Multivariate calibration and chemometrics for near infrared spectroscopy: which method? J Near Infrared Spectrosc. 2000;8:229–37.
26. Engel J, Gerretzen J, Szymanska E, Jansen J, Downey G, Blanchet L, et al. Breaking with trends in pre-processing? Trends Anal Chem. 2013;50:96–106.
27. Data preprocessing chapters in comprehensive chemometrics, Vol2, Section Ed. J.Trygg, General Ed. S.D. Brown, R.Tauler, B.Walczak, Elsevier, Amsterdam, The Netherlands, 2009.
28. Beebe KR, Pell RJ, Seasholtz MB. Chemometrics. A practical guide. New York: Wiley; 1998.
29. Booksh KS, Kowalski BR. Theory of analytical chemistry. Anal Chem. 1994;66:782A–91A.
30. Linear soft modelling chapters in Comprehensive chemometrics, Vol2, Section Ed. A. de Juan, General Ed. S.D. Brown, R. Tauler, B. Walczak, Elsevier, Amsterdam, The Netherlands, 2009.
31. Malinowski ER. Factor analysis in chemistry. New York: John Wiley & Sons; 2002.
32. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer Verlag; 2002.
33. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst. 1987;2:37–52.
34. Lee TW. Independent component analysis—theory and applications. Dordrecht: Kluewer Academic Publishers; 1998.
35. Tauler R. Multivariate curve resolution applied to second order data. Chemom Intell Lab Syst. 1995;30:133–46.
36. Smilde A, Bro R, Geladi P. Multiway analysis: applications in the chemical sciences. New York: John Wiley & Sons; 2004.
37. Bro R. PARAFAC tutorial and applications. Chemom Intell Lab Syst. 1997;38:149–71.
38. Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proc IEEE. 2015;103:1449–77.
39. Blanchet L, Smolinska A. In: Jung K, editor. Statistical analysis in proteomics. New York, NY: Springer New York; 2016. p. 209–23.
40. Acar E, Rasmussen MA, Savorani F, Næs T, Bro R. Understanding data fusion within the framework of coupled matrix and tensor factorizations. Chemom Intell Lab Syst. 2013;129:53–63.
41. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. PNAS. 2003;100:3351–6.
42. Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. Plant J. 2007;52:1181–91.
43. Löfstedt T, Trygg J. OnPLS - a novel multiblock method for the modelling of predictive and orthogonal variation. J Chemom. 2011;25:441–55.
44. Schouteden M, Van Deun K, Pattyn S, Van Mechelen I. SCA with rotation to distinguish common and distinctive information in linked data. Behav Res Methods. 2013;45:822–33.
45. Kuligowski J, Perez-Guaita D, Sanchez-Illana A, Leon-Gonzalez Z, de la Guardia M, Vento M, et al. Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE). Analyst. 2015;140:4521–9.
46. Qannari EM, Courcoux P, Vigneau E. Common components and specific weights analysis performed on preference data. Food Qual Prefer. 2001;12:365–8.
47. E. Ortiz-Villanueva, F. Benavente; B. Piña; V. Sanz-Nebot; R. Tauler; J. Jaumot. Data fusion strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data. Submitted.

Richard Brereton did his undergrad, postgrad and postdoc studies in the University of Cambridge, after which he moved to the staff of the University of Bristol, and is a Fellow of the Royal Society of Chemistry, Royal Statistical Society and Royal Society of Medicine. He has published around 400 articles including 7 books, and is Editor-in-Chief of the journal *Heritage Science*.

Jeroen Jansen is Assistant Professor and acting group leader of the Department of Analytical Chemistry & Chemometrics, Radboud University in Nijmegen, The Netherlands. He develops novel analysis methods for chemical data from new types of experiments in medical, environmental and industrial science. He designs these methods to integrate domain knowledge, information on the measurements and informative matrix algebraic operations to provide dedicated answers to these new questions. He collaborates intensively with industry, governmental institutions and developers of analytical technology to apply these methods in production processes, societal safety and individual health.

João Lopes is a chemical engineer and currently Professor of Pharmaceutics at the Faculty of Pharmacy of University of Lisbon. He received his PhD in Chemical Engineering in 2001. His scientific activity has been focused on vibrational spectroscopy, chemometrics and process analytical technology applied to (bio)pharmaceutical, agro-food and environmental processes. He coordinates the research unit in chemometrics and process analytical technology of the Portuguese Associated Lab LAQV/REQUIMTE. He has published 92 peer-reviewed scientific papers and coordinated more than 30 R&D projects involving collaborations with the private sector, mainly pharmaceutical companies. He is currently a member of the steering committee of the EUFEPS PAT/QbD network and Vice-President of the Portuguese Society for Pharmaceutical Sciences.



Federico Marini is a researcher and Professor of Chemometrics at the University of Rome La Sapienza. In 2006, he was awarded the Young Researcher Prize from the Italian Chemical Society and in 2012 he won the Chemometrics and Intelligent Laboratory Systems Award. His research activity is focused on all aspects of chemometrics, ranging from the application of existing methods to real world problems in different fields to the design and development of novel algorithms with particular focus on nature-inspired methods, multiset and multiway modelling and classification. He is currently the coordinator of the Chemometric group of the Italian Chemical Society and a member of the Chemometric study group of EUCheMS.

Alexey Pomerantsev is a principal researcher at the Semenov Institute of Chemical Physics of Russian Academy of Sciences in Moscow, Russia, and a founding member and Chair of the Russian Chemometrics Society. He is developing statistical fundamentals for chemometric methods (non-linear regression and curve resolution, projection methods) and implements them in Chemometrics Add-In software, which is presented in the book '*Chemometrics in Excel*' (Wiley, 2014).

Oxana Rodionova is a leading researcher at the Semenov Institute of Chemical Physics of Russian Academy of Sciences in Moscow, Russia, and founding member and Secretary of the Russian Chemometrics Society. She has been working for several years on the development of chemometric tools (SIC method, DD-SIMCA) for the analysis of the authenticity of food and drugs, for process control in pharmaceutical and nuclear industries.

Jean-Michel Roger is specialized in chemometrics applied to Near Infrared Spectroscopy (NIRS). His research produced some methods to solve problems of calibration robustness. Some specific methods and applications were derived to address the problems of calibration transfer, drift compensation online, inter seasonal adjustment or compensation of the moisture effect for NIR based characterization of soil. He is involved in International Societies which promote chemometrics and NIRS.

Beata Walczak is head of the Analytical Chemistry Department at the University of Silesia, Katowice, Poland. She is involved in the development of chemometric methods for data preprocessing (de-noising, warping, normalisation, etc.) and data analysis (RBF-PLS, Dissimilarity-PLS, ANOVA-TP).

Romà Tauler is Research Professor at the Institute of Environmental Assessment and Water Research (IDÆA), CSIC, in Barcelona, Spain. He is Chief Editor of the *Journal of Chemometrics* and International Laboratory Systems President of the Catalan Chemistry Society, 2008–2013. He has published more than 350 papers in ISI journals (WoS h-index 51). His main research interests are in chemometrics, especially in the development of multivariate curve resolution methods (http://www.cid.csic.es/homes/rtaqam).