FEATURE ARTICLE

CrossMark

Chemometrics in analytical chemistry—part II: modeling, validation, and applications

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]

Abstract
The contribution of chemometrics to important stages throughout the entire analytical process such as experimental design, sampling, and explorative data analysis, including data pretreatment and fusion, was described in the first part of the tutorial "Chemometrics in analytical chemistry." This is the second part of a tutorial article on chemometrics which is devoted to the supervised modeling of multivariate chemical data, i.e., to the building of calibration and discrimination models, their quantitative validation, and their successful applications in different scientific fields. This tutorial provides an overview of the popularity of chemometrics in analytical chemistry.

Keywords Calibration · Discrimination · Validation · Prediction · Omics · Hyperspectral imaging

Introduction

Chemometrics is a highly interdisciplinary field (see Fig. 1) whose relevance among the chemical disciplines, in general, and, analytical chemistry, in particular, has considerably grown over the years. Despite this, it is still largely unknown to many analytical chemists and often misused or not completely understood by practitioners who rely on those very few techniques which are implemented in the most widespread commercial software. However, chemometrics represents a wealth of possibilities for the (analytical) chemists, as it is a discipline which accompanies the analytical workflow at all stages of the pipeline. In this context, the present paper represents the second part of a feature article aiming at highlighting the fundamental role

that chemometrics has within analytical chemistry and the different tools which could and should be used to address key issues in the field. Starting from these premises, in the first part of this tutorial article [1], topics like sampling, experimental design, data preprocessing projection methods for data exploration and factor analysis, and data fusion strategies were covered and contextualized in the framework of specific goals within analytical chemistry. On the other hand, the present paper deals with aspects related to predictive modeling (both for quantitative and qualitative responses) and validation, and presents some successful examples of the application of chemometric strategies to different "hot topics," together with a few considerations about how the discipline will evolve in the next years.

✉ 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 Nijmegen, 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

Fig. 1 Chemometrics as an interdisciplinary field
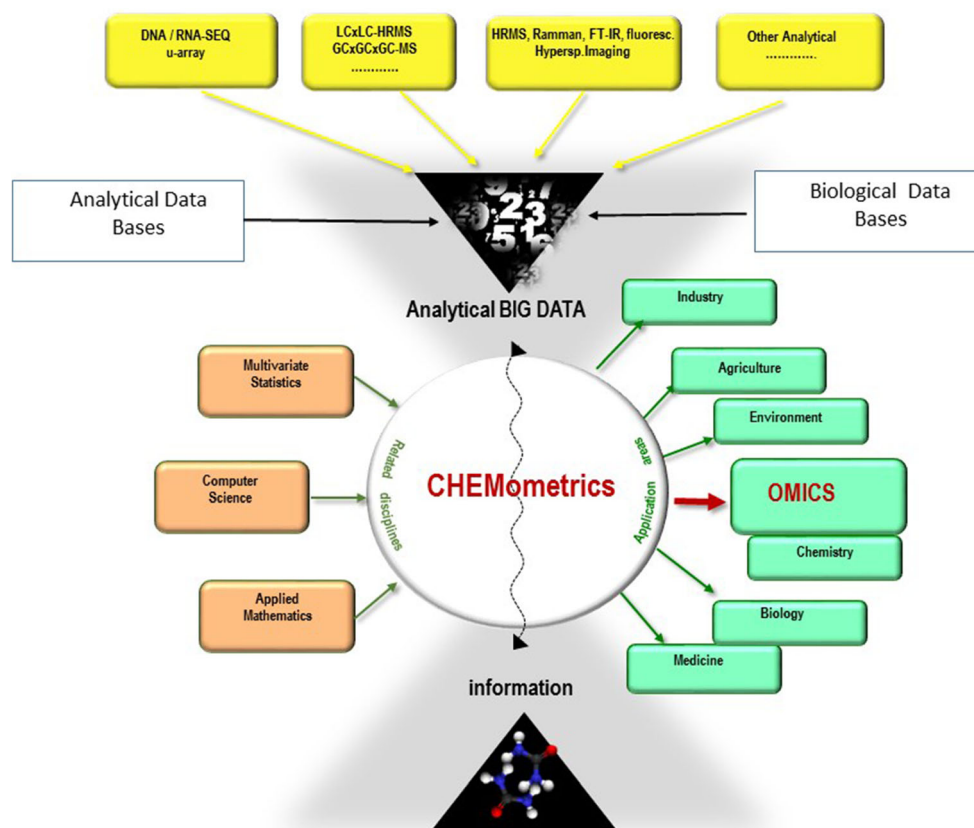
Predictive chemometrics modeling

Many applications of chemometrics in analytical chemistry are associated with the so-called predictive modeling [2], where a mathematical model represents the relationship between a chemical and/or physical property of a sample with generally easier and cheaper to acquire instrumental signals. Typical examples are modeling of protein content in wheat samples based on their NIR signals, modeling of octane number of fuel based on the NIR signals, modeling of antioxidant properties of tea extract based on their chromatographic fingerprints, modeling of biological activity of studied compounds based on their descriptive parameters, medical diagnostics based on genomic, proteomic and/or metabolomics fingerprints, etc.

Chemometric models may predict both, quantitative (continuous-valued, e.g., protein content) and qualitative (discrete, such as healthy/ill or compliant/non-compliant, class membership) properties. Calibration, classification, and discrimination type of problems (see Fig. 2) are common in chemometrics. The same considerations hold in other predictive modeling situations.

In the other words, the main goal of multivariate modeling is the prediction of parameter(s), the measurement of which would be time- and cost-intensive, based on fast and cheap measurements such as NIR, or on calculated parameters

(energy of interactions, topological indices, etc.). The responses to be predicted are called dependent variables and the variables used to perform this prediction are called independent variables.

Multivariate calibration

Assuming that the set of centered dependent variables is denoted as $\mathbf{Y}$ ($m \times k$) and a set of centered independent variables
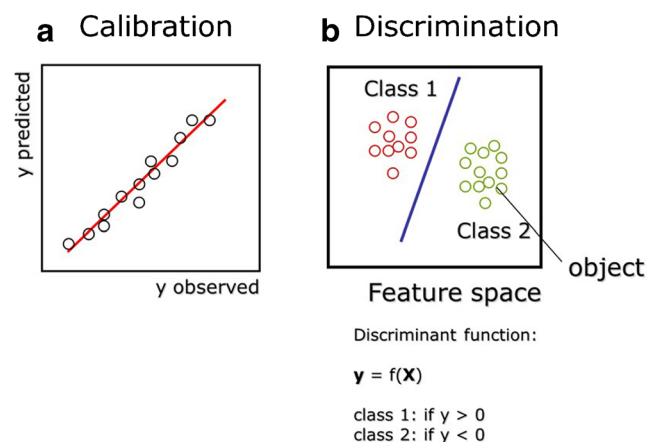

Fig. 2 Schematic representation of calibration and discrimination principles

is denoted as $\mathbf{X}$ ($m \times n$), the constructed calibration model can be represented as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

where the matrix $\mathbf{B}$ ($n \times k$) collects the k vectors of regression coefficients, and $\mathbf{E}$ ($m \times k$) denotes the residuals.

If only one parameter is to be predicted, the above equation becomes:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ and $\mathbf{e}$ are vectors of dimensionality $m \times 1$ and $\mathbf{b}$ is a vector of dimensionality $n \times 1$. In the following, it is assumed that only one parameter is predicted for the sake of parsimony, but this approach may be extended to the prediction of multiple dependent variables.

How the model is constructed depends mainly on data dimensionality and its correlation structure. If matrix $\mathbf{X}$ contains fewer variables than samples ($n < m$) and these variables are uncorrelated, then the vector of regression coefficients may be calculated as:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

If matrix $\mathbf{X}$ contains more variables than samples ($n > m$), and/or if these predictor variables are highly correlated, then classical multivariate methods like multiple linear regression (MLR) cannot be applied. This is because matrix ($\mathbf{X}'\mathbf{X}$) cannot be inverted due to rank deficiency. In this case, Eq. (3) may be replaced by a more general equation:

$$\mathbf{b} = \mathbf{X}^{+}\,\mathbf{y} \tag{4}$$

where $\mathbf{X}^{+}$ represents a "generalized inverse" of $\mathbf{X}$.

Several common multivariate methods in common use, such as Ridge regression (RR), principal components regression (PCR), and partial least squares regression (PLSR), differ in the estimation of this inverse $\mathbf{X}^{+}$. RR is more popular among statisticians than in chemometrics [3], probably due the widespread use of PCR and PLS loading and score plots for model visualization and interpretation in chemometrics.

Once the model is build (i.e., the vector of regression coefficient is calculated), it may be used to predict the y value for a new sample(s):

$$y_{\text{pred}} = \mathbf{x}_{\text{new}}\,\mathbf{b} \tag{5}$$

Multivariate classifiers

Many chemical problems are associated with classification, i.e., the assignment of studied samples to one or more classes based on a set of measurements recorded to characterize them.

One-class classification (OCC) methods (also known as class-modeling methods) are directed toward modeling of individual classes new [4] independently of others. They are focused on similarities among the samples from the same class rather than the differences between classes.

In chemometrics, the most popular OCC method is soft independent modeling of class analogy (SIMCA) new ref. [5]. It forms a closed acceptance area that delineates the target class in the multivariate space using disjoint PCA (performed just on the training set in the target class rather than all objects studied). SIMCA calculates two types of distance, the D statistic (within the disjoint PC space) and the Q statistic (from an object to the disjoint PC space), also known by various different names according to author. There are various ways to interpret these distances according to author and software package new ref. [6].

In contrast, two or multi-class classifiers (often known as supervised discriminant methods) are used to form boundaries between classes rather than around classes. The oldest and most known is linear discriminant analysis (LDA) proposed by Fisher [7]. It looks for linear surfaces (hyperplanes) which best separate the samples from different categories in the multidimensional space, the identification of which relies on the relative position of the barycenters of the groups (between-class scatter B) and on the within-class variance/covariance W. However, as it happens in regression, the presence of a higher number of variables than samples ($n > \text{m}$) or collinearity among the predictors result in the inapplicability of the method, so that alternatives should be found to deal with the multi- or megavariate data, commonly encountered in modern chemical problems. A possible way of overcoming this limitation, analogously to what already discussed for calibration, is to project $\mathbf{X}$ onto a suitable low-dimensional subspace of orthogonal variables, where discriminant analysis can be performed. This can be done, for instance, by a preliminary PCA decomposition of the $\mathbf{X}$ matrix (PCA-DA) or a PLS between $\mathbf{X}$ and suitably defined $\mathbf{Y}$ (PLS-DA) [7], new [8, 9].

To perform discrimination among k classes via PLS-DA, information about class membership of each sample is in the form of a binary coded classification vector or matrix. Usually, a separate set of calibrations are performed against a classification vector for each class in the model (PLS1), although a single model can alternatively be performed against a classification matrix with as many columns as classes (PLS2). Usage of this binary coding makes it possible to transform the discrimination problem into a regression problem. Assignment of a sample to the studied classes is then accomplished using a variety of decision criteria according the predicted values in the classification vector or matrix. Using PLS methods, in addition to predicting class membership of individual samples, it is also possible to investigate which variables are most appropriate as discriminators or markers.

Classification can be performed using different methods, but each problem requires a targeted approach and there is no universally agreed way. Discrimination can perfectly attribute a new sample if it is a member of one of a number of predefined classes. However, in case it does not unambiguously belong to any of such predefined classes, many two or multi-class methods may often have problems with outliers (or ambiguous samples). OCCs can distinguish the target class from any other objects and classes; therefore, they are more usually employed for authentication new [10] in chemometrics, not requiring samples to belong to predefined classes. The alternative (Bayesian) methods although common in machine learning are not well established within chemometrics.

More details about chemometrics modeling

The quality of the model is determined by the quality of the data used for its building. One of the most important issues which influences model performance is the "representativity" of the model. Samples included into the model dataset should represent all sources of data variability within the scope of the experiment or the observed system.

While constructing the PCR, PLSR, PCA-DA, or PLS-DA model, it is necessary to properly estimate model complexity (the number of latent variables used). Too many latent variables lead to overfitting (an increase of the variance of the predicted values), too few of them cause underfitting (an increase of the bias of the predicted values). Model with a proper number of latent variables ensures an appropriate compromise (balance) between the variance and bias. Model complexity is usually determined, based on the cross-validation procedure (vide infra).

Model performance

Each model has to be carefully evaluated. Model fit is calculated for the samples from the model set. Of course, the most important is model predictive power, i.e., its ability to correctly predict a dependent variable(s) for new samples (not used for the model building). Estimation of the model predictive power can be performed based on the independent and representative test set, or using the cross-validation procedure (application of the cross-model validation is a recommended approach, [11]).

Good models have low complexity (i.e., few latent variables) and low, yet similar, cross validation and test errors. If predictive power of a model is unsatisfactory, we should consider the following reasons:

- data do not contain necessary information for modeling the dependent variable;
- data contains outliers and
- the studied relationship is highly non-linear.

The first case can be diagnosed based on, e.g., the UVE-PLS method [12]. In the arsenal of chemometric tools, there are also the robust versions of PCR and PLSR, which allow dealing with the outliers [13]. We are also equipped with efficient approaches to deal with non-linearities [14].

Kernels and dissimilarity matrices

Area of applicability of multivariate methods, such as PCR or PLSR, can be highly extended to very complex non-linear systems, when instead of the model described by Eq. (1), the following model is constructed:

$$\mathbf{Y} = \mathbf{K}\,\mathbf{A} \tag{7}$$

where $\mathbf{K}$ ($m \times m$) denotes the kernel or dissimilarity matrix, and $\mathbf{A}$ ($m \times k$) represents the k vectors of regression coefficients.

The simplest kernel is defined as $\mathbf{K} = \mathbf{X}\mathbf{X}'$. It is the linear kernel, but there are many other interesting non-linear kernels, the Gaussian kernel being the most popular one. New possibilities of modeling complex non-linear systems are also offered by different types of dissimilarity matrices such as the Euclidean distance matrix [14].

Validation

Classical linear regression modeling relies on strong assumptions about the fulfillment of the model and optimal least-squares estimation of the parameters under a set of specific constraints. Chemometrics modeling on the other side uses less assumptions. It extracts and represents the information in the collected experimental data through models which, depending on the specific application, provide approximations of the system under study or predictions to be drawn. Here, it should be stressed that given the problem under investigation and, in particular, the available data, there is always more than one single model a researcher can, in principle, calculate. However, when dealing with soft models, i.e., models which are based mostly only on the experimentally measured data, various additional factors (e.g., the number of available samples and their representativeness via an appropriate experimental design (see part I of this article, [1]), the peculiar characteristics of the method, the specific algorithm for computing the solution) concur in defining their performances, so that not all the possible models have the same quality.

In this framework, validation is a fundamental step of the chemometrics pipeline which is aimed at evaluating whether reliable conclusions may be drawn from a model [15, 16]. In a fundamental review [15], Richard Harshman suggests that the validation process should include the appropriateness of the model, the computational adequacy of the fitting procedure, the statistical reliability of the solution, and the generalizability and explanatory validity of any resulting interpretations.

As the definition suggests, checking for appropriateness of the model means to verify how appropriate the model is for the specific questions/problems one has to deal with. In the context of soft modeling, this aspect implies examining, e.g., whether a latent structure should be expected and, if so, whether it should be bilinear or multilinear, or whether orthogonality of the loadings should be imposed to achieve a hierarchical relation among the components. Investigation of whether the data should be preprocessed or some sort of data transformation (e.g., logarithmic or square root) should be used also falls within this aspect.

On the other hand, computational correctness has mainly to do with the fitting procedure used to calculate the model parameters. In particular, one should investigate if an iterative algorithm has converged to the desired global optimum or, instead, a local minimum is reached, whether the solution is independent or not on the choice of the starting point, and how reproducible it is.

Addressing the issue of statistical reliability means to investigate how appropriate the distributional assumptions are, if any, how stable and parsimonious the model is with respect to resampling, and whether the correct number of components has been chosen. In particular, estimating what is the stability of the solution across subsamples allows to identify whether the results are reliable enough to allow interpretation and, possibly, if the conclusions can be generalized to new samples.

Lastly, the aspect of explanatory validity deals with assessing how well the characteristics of the system under study are captured by the model. Addressing this issue means, e.g., examining to what extent are the results interpretable and how are the latent factors calculated by the model related to the external information available on the individuals or the variables, verifying whether there could be nonlinearities or, in general, whether the residuals contain further (unmodeled) systematic variation, indicating that not all the structure of the data has been captured/explained by the model. Questions about whether highly influential observations or extreme points/outliers, which could bias the solution, are present or the discussion of whether other results confirm or conflict with the conclusions of the model also fall within this aspect.

Having identified the main issues embraced by the validation process, then appropriate diagnostics should be defined accordingly in order to evaluate the quality of the model(s). These diagnostics, which may be expressed in a graphical form (allowing a straightforward evaluation), quantitatively or in the formalism of statistical testing, can rely on the investigation of model parameters, or of the residuals. In the latter case, to avoid overoptimism, it is essential not to use for validation the residuals of the fitted models (i.e., those evaluated on the same data used for calculating the model parameters, i.e., the so-called calibration or training set), as in almost all the cases they cannot be considered as representative (neither in terms of structure nor of magnitude) of the residuals that would be obtained by applying the model on new data. Strategies for obtaining more realistic residuals include the use of an external test set and cross-validation.

Test set validation consists in applying the model to a new set of data (the test set) and calculating the corresponding residuals (which of course implies that the necessary information for this computation—e.g., the true values of the responses in the case of predictive modeling—is available for these new samples). Accordingly, this validation strategy is the one which most closely resembles how they will be actually used, i.e., to make predictions on future observations. Ideally, the test set should be built to be as representative as possible of the population of future measurements (and, hence, to be as independent as possible from the calibration set), e.g., by including samples collected a sufficient amount of time after the training ones or coming from different locations, suppliers, and so on. This is the only way in which a reliable estimate of the residuals to be expected for real data can be obtained.

Cross-validation is an internal resampling method which simulates test set validation by repeatedly splitting the available data in two subsets, one for model building and one for model evaluation. At each iteration, a part of the data is left out and a model is calculated without these values; the model is then used to predict the left out data and the corresponding residuals are computed. The procedure is repeated until each observation has been left out at least once or when the desired number of iterations is reached. This strategy is particularly suitable for small data sets but since the calibration and validation sets are not truly independent on one another may result in a biased estimate of the residuals and, hence, of the overall model quality. Indeed, cross-validation is often used for model selection or model comparison but should not be used for the final assessment of the reliability of the model, where test set validation should be preferred.

When the environmental conditions change—e.g., the temperature of the spectrometer—the measured data are impacted and the model may not stay valid. This specific problem falls into the so-called robustness problem and must be treated with specific diagnostic and correction methods which are out of the scope of this paper.

Successful applications of chemometrics in analytical chemistry

Instrumental methods of analysis provide rich information about chemical systems which can be arranged in data structures of different complexity [17] and can be analyzed by chemometric methods which take profit of their intrinsic structure (Fig. 3).

Examples of successful applications of chemometrics in Analytical Chemistry appeared already in the 1980s and 1990s of the past century in the fields of food/agriculture analysis [18], oil octane determinations [19], and process analytical chemistry in industry [20]. These successful applications were dominated by the development and use of pattern recognition, cluster analysis and multivariate calibration methods [21]. Chemometric methods were also expanding rapidly in other fields like in source apportionment environmental chemistry studies [22] or in QSAR studies in medicinal chemistry [23]. This tendency was then expanded at the beginning of this century by the incorporation of new applied fields, especially with the development of high throughput spectroscopic and chromatographic analytical methods providing megavariate data [24]. Other significant examples of this tendency nowadays are in the explosion of omics, spectroscopic imaging, and all new big analytical data applications, which are described in more detail below as example.

Omics

The emergence of the "omics" paradigm led to dramatic shifts in analytical chemistry (see Fig. 4 and [25]). Widely used principles for chemical detection and quantification of chemical mixtures, such as nuclear magnetic resonance (NMR) and liquid and gas chromatography coupled to mass spectrometry, are now used for comprehensive quantification of proteins and
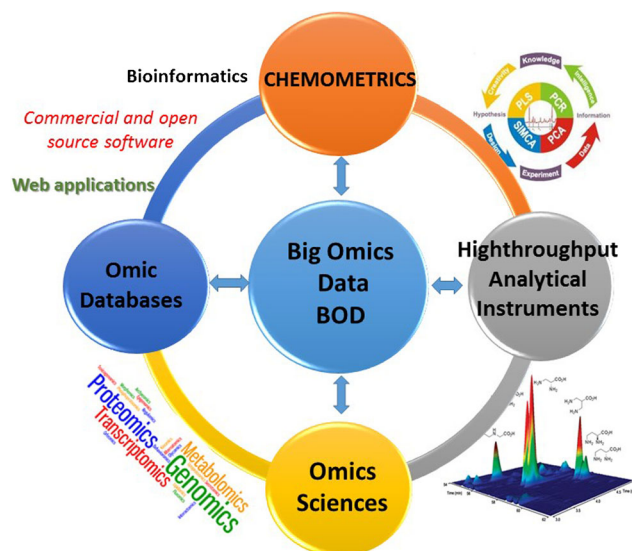
Fig. 4 Big omics data analysis and their interdisciplinary relationships (adapted and printed with permission from [25])

metabolites—genes and their transcripts require more specific technology. This, however, requires considerable chemometric adaptations to provide informative models.

In omics data, the vast majority of measured compounds is unrelated to the studied biochemical process. Such unrelated compounds will contribute considerably to chemometric models. "Sparse" methods [26] limit the compounds that contribute to those that describe the studied biochemical process well, providing information-dense and process-focused models. Sparse methods, however, specifically aim for the smallest set of informative biomarkers, such that many relevant molecular species, related to included compounds, may be ignored. Developing a method to find a comprehensive set of compounds involved in a studied biochemical process is one of the imminent challenges in chemometrics.

Fig. 3 Analytical data types
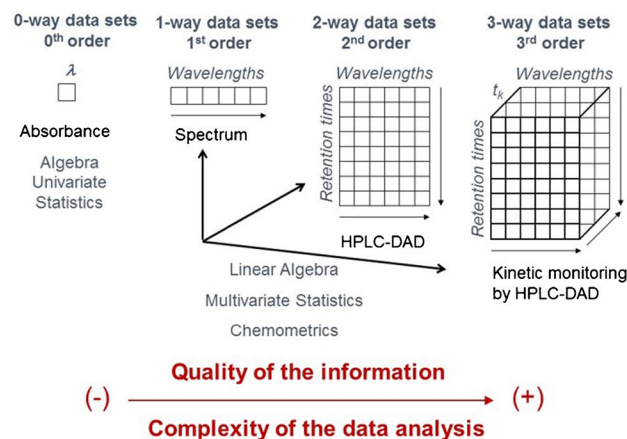
Conventional applications of chemometrics may aim at classifying sample groups or predicting a difficult-to-measure sample characteristic. Many omics studies however aim for a comprehensive overview on the metabolites and/or proteins within a biological system that respond to combinations of experimental factors like time, diet, sex, and species. Principal component analysis and PLSR cannot extract such information interpretably. Combining information on the experimental design by analysis of variance (ANOVA) with multivariate PCA, for example the ASCA method [27], has proven invaluable to understand contributions of different experimental factors to the metabolic and protein complement of an organism.

Chromatography, mass spectroscopy (MS), and NMR-based detection may introduce considerable artifacts, e.g., chromatographic misalignment, that needs to be removed prior to modeling omics data. Concentrations may vary many orders of magnitude between metabolites or proteins, as does the sensitivity of analytical detection and the bioactivity of each compound; the values that an omics analysis returns may therefore be incomparable between metabolites or proteins. Thirdly, the sample origin itself may reduce information content of the resulting data: the amount of water excreted from the body greatly affects urinary metabolite concentrations, such that the resulting data needs to be normalized for this. Several data pre-processing methods are available to mitigate such issues [28], but a comprehensive study on preprocessing metabolomics data shows that data preprocessing may considerably affect how well the model distinguishes samples and indicates relevant biomarkers. In [29], different data analysis strategies for LC-MS metabolomics studies and alternatives for an optimal selection of the different alternatives are given.

Hyperspectral imaging

Hyperspectral imaging is an active area of research that has grown quickly during the last years. Hyperspectral images are measurements that contain spatial and spectral information and they provide chemical information and detailed knowledge of the distribution of the sample constituents in the surface (or volume) scanned (Fig. 5). Hyperspectral images result from spectroscopic readings of hundreds of contiguous spectral channels at each spatial position (pixel) of the target sample under study. Hyperspectral imaging techniques can be based on different spectroscopies like Raman, infrared, and fluorescence. They are useful methods in different areas, such as polymer research, materials science, biomedical diagnostic, pharmaceutical, industry, analytical chemistry, process control, and environmental analysis [30, 31].

Combination of hyperspectral imaging with chemometrics is especially useful for the quantification of the compounds in a product and for heterogeneity control. The application of chemometric tools is needed at different stages of image data analysis such as in compression (i.e., wavelets), pretreatment (i.e., correcting baseline, background), and in exploration. Several methods have been proposed to extract the maximum amount of information from the available spectral imaging data. Multivariate image analysis (MIA) and PCA have been applied in this context. Quantitative analysis of image constituents has been known under the denomination multivariate image regression (MIR) and it has often been performed by multivariate calibration methods, like PLSR. Another possible approach for image analysis is multivariate curve resolution (MCR), which is based on a bilinear model of the image, i.e., concentration profiles (folded back into distribution maps) and pure spectra of the image constituents. The MCR-ALS (alternating least squares) method has been adapted particularly well to hyperspectral image resolution due the ease of introduction of external spectral and spatial information about the image as a constraint and due to its ability to work with single and multiset image arrangements [32, 33]. At present, different alternatives for hyperspectral image data fusion have been proposed [34].

Mass spectrometry imaging (MSI) is an extremely useful tool for the study of complex mixtures in real biological samples such as cells or tissues [35, 36]. Its usefulness is due to its high chemical specificity to simultaneously analyze multiple compounds in a very wide mass range, from small (i.e., metabolites) to large molecules (i.e., proteins). In addition to the qualitative information about the presence or absence of a particular molecule, MSI gives the spatial distribution of these molecules over the analyzed sample surface. Thus, MSI couples the spatial information provided by the spectral imaging techniques with the chemical specificity based on the mass accuracy of the high-resolution mass spectrometry techniques (and possible MS/MS analysis) that allows unambiguous identification of the detected molecules. Application of chemometric methods to MSI data faces a bottleneck concerning the vast size of the experimental data sets. The standard approach for MSI data compression consists in binning mass spectra for each pixel to reduce the number of m/z values. New approaches based on the selection of the regions of interest (ROI) have been proposed in this context [37].

Applications of chemometric methods rely on the use of appropriate and easily accessible software. In Table 1, a list of the most popular software used by chemometricians at present is listed together with their internet links. Some of these tools are commercial products, usually with low-cost licenses for the academic purposes, and others are freely available as open source products. Selection of them can be done according to the working environment and purposes. For instance, in academics and research, the MATLAB (The Mathworks Inc) is a commonly used underlying computer and visualization environment, also due to the considerable number of advanced toolboxes specifically designed to solve a very large number of scientific and technological problems. They provide the state
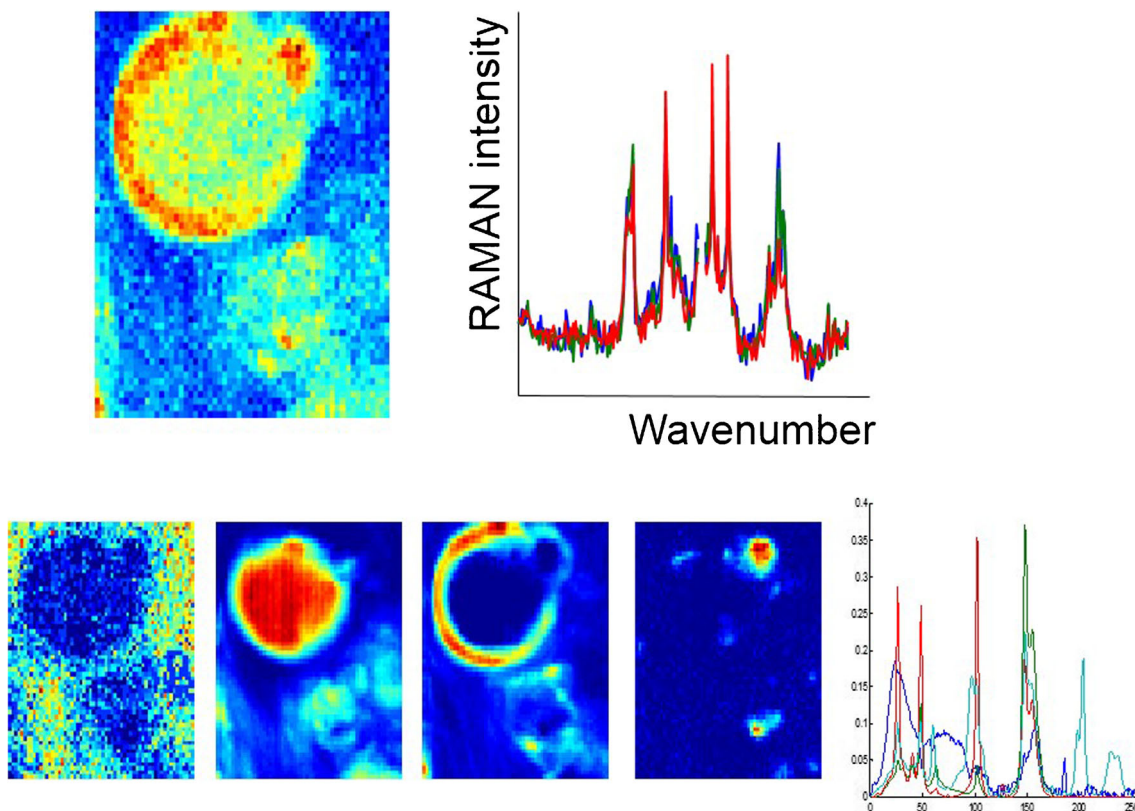
Fig. 5 Hyperspectral imaging: spatial and spectral information of the sample constituents

Table 1   Chemometrics software

Software name	Owner/Author	Link	Purpose
MARLAB and Toolboxes	The Mathworks	https://www.mathworks.com/	General
PLS Toolbox (MATLAB)	Eigenvector Inc	http://www.eigenvector.com/	General
Tensorlab	N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, University of Leuven	http://www.tensorlab.net	N-way data analysis
Factor Analysis and Chemometrics Toolbox	Applied Chemometrics, Inc.	http://www.chemometrics.com	Applied chemometrics
Multivariate calibration	A. Olivieri, IQUIR	http://www.iquir-conicet.gov.ar/eng/div5.php?area=12	Multivariate calibration
Multivariate Curve Resolution	R. Rauler, A. de Juan and J. Jaumot IDAEA-CSIC and University of Barcelona	https://mcrals.wordpress.com/	Multivariate curve resolution
Classification toolbox	D. Ballabio and R. Todeschini University of Milan	http://michem.disat.unimib.it/chm/download/classificationinfo.htm	Classification and QSAR
N-way toolbox	Rasmus Bro, University of Copenhagen	http://www.models.life.ku.dk/nwaytoolbox	N-way data analysis
Unscrambler	CAMO	http://www.camo.no	General
Pirouette	Infometrix	http://www.infometrix.com	General
SIMCA	Umetrics, Sartoius	http://www.umetrics.com	Pattern Recognition
Prosensus Pro MV	MACC/ProSensus	http://www.prosensus.ca	Process control/monitoring
Multivariate Statistical Analysis in Chemometrics in R	Peter Filzmoser and Kurt Varmuza University of Wien	https://cran.r-project.org/web/packages/chemometrics/index.html https://www.rdocumentation.org/packages/chemometrics/versions/1.4.2	General
R-based chemometrics software	Chemometrics Group of the Italian Chemical Society	http://gruppochemiometria.it/index.php/software	General

of the art solutions to many of these problems. These toolboxes are either purchased directly to the vendor or provided freely by a large number of developers worldwide. In the same MATLAB environment, there are also other commercial products like the PLS Toolbox, which is one of the tools more profusely used by chemometricians in their work. This toolbox is constantly evolving and improving, incorporating many of the tools developed by chemometricians worldwide. Other more specific toolboxes in this context also implemented for the MATAB environment are given in the table. Among them, the TENSORLAB toolbox gives a very large number of tools for the investigation of complex multi-way data sets using tensor products and most of the currently available variants of multilinear model-based methods. Other toolboxes are available for multivariate calibration (MVC packages for 1st, 2nd, and 3ed order calibration), multivariate curve resolution (MCR), or classification and discrimination type of problems. Commercial software developed as end products for most current applications in the industry are also available in the market as the Unscrambler (CAMO), Pirouette (Infometrix), or SIMCA (Sartorious) products. Recently, due to the increasing number of users of open source R language applications, especially in the biosciences fields, R language platforms are gaining popularity in the chemometrics field as well and there are already chemometrics software packages under this open source environment.

Where are we going?

Put 1000 chemometricians in a room and there will be 1001 opinions as to its future directions. There is no Nostradamus of chemometrics, so we can only guess. However, at least some of chemometrics is about predicting future trends.

One thing that is certain is that the subject will become very diverse and chemometrics plays to very diverse audiences. For example, some will be advanced statisticians with a strong grounding in methods for estimation and distributions and hypothesis testing. There will be computer scientists who are good at algorithm development. And at the other end, synthetic chemists who just want a package to optimize their reactions. There is no unifying knowledge base.

However, classically, most of modern chemometrics was first applied within analytical chemistry, and many of the major conferences, journals, and texts are still primarily based within analytical chemistry, which will remain an important but not exclusive cornerstone. In the context of learning, chemometrics though still has trouble incorporating itself into the core knowledge base of the analytical scientist's syllabus. This is partly because the core corpus of knowledge is crowded and any new material must displace older topics. Basic univariate statistics has always been part of the core understanding of analytical chemists, including concepts such

as precision, accuracy, univariate calibration, uncertainty, etc. Skoog and West [38] contain only univariate analysis in their opening chapters, as do Christian and colleagues [39]. Spreadsheets are considered important for basic analytical chemistry, but multivariate analysis is not.

Chemometrics is likely only to become part of the basic education of analytical scientists if other, more classical, methods are removed. As instrumental techniques replace classical tests, the need to have some familiarity with computational and statistical approaches for analyzing instrumental data will hopefully become essential learning for the analyst, although this may take many years or even decades. There is also a problem that some excellent laboratory-based analysts are not all mathematically oriented, and course organizers, wanting to attract student to fill their places may not want to load their courses with too much maths.

Until that time, chemometrics will be viewed as an advanced topic for analytical chemists, primarily for specialism at graduate level. A few universities with very active chemometricians will fight to incorporate this into the basic syllabus, but chemometrics will primarily be encountered by analytical chemists at graduate level, or in professional development courses. A few enthusiasts or members of dedicated groups will encounter chemometrics personally without the need for formalized courses. However, as instruments become ever more sophisticated, the quantity of data expands, and the need for chemometrics over classical analytical tests increases in real life. Only over several decades will applied multivariate statistical approaches replace traditional laboratory testing, when the latter becomes redundant. And at the same time, in other scientific disciplines and communities like in bioinformatics and data sciences, new developments and proposals emerge, with significant challenges in big data storage and processing. Effectively analyzing all currently available data through directed statistical analysis is very difficult [24, 25, 40].

Within research, though, as megavariate datasets become more common, there will be an increased need for chemometrics in research, sometimes though unfortunately standing on weak foundations as basic education is lacking. Two key future growth points are identified.

Metabolomics is a very rapid point for expansion. This both involves destructive analysis primarily from spectroscopy and chromatography, and also in situ approaches such as hyperspectral imaging. The growing need will be to educate users about the quality of data, both to improve instrumental resolution and signal quality, and to obtain sufficient and well representative samples. Unfortunately, the majority of metabolomic data is of insufficient quality for sophisticated chemometrics analysis, often leading to expensive but inconclusive experimentation. An urgent need is for experimenters to consult chemometrics experts before rather than after data is collected. Although there are some very elegant descriptions of the application of chemometrics to well designed and

Table 2 Chemometrics textbooks and reference works

Title	Authors	Editorial	Details/Comment
Comprehensive Chemometrics Chemical and Biochemical Data Analysis	Brown S., Tauler R. and Walczak B. eds	© Elsevier 2009–2018 https://www.elsevier.com/books/comprehensive-chemometrics/brown/978-0-444-52701-1	General, Comprehensive (4 volumes 2896 pages). New edition is in preparation
Data Handling in Science and technology	Walczak B. ed.	Elsevier 1984–2017 https://www.sciencedirect.com/bookseries/data-handling-in-science-and-technology/volumes	30 volumes 1984–2017, covering all aspects of chemical data analysis
40 years of Chemometrics—From Bruce Kowalski to the Future	B. Lavine, S. Brown and K.S.Booksh eds	ACS Symposium Series Nr 1199, 2015	Historical overview and recent contributions
Chemometrics in Excel	Pomerantsev A.L.	Wiley 2014	Chemometrics tools in Excel
Chemometrics in R	Wehrens R.	Springer, 2011	R tools
Introduction to Multivariate Statistical Analysis in Chemometrics	K. Varmuza and P. Filzmoser	CRC Press 2009	General, Introductory
Chemometrics: Statistics and Computer Application in Analytical Chemistry	Otto M.	Wiley-VCH 2nd edition 2007	General, Introductory
Applied Chemometrics for Scientists	Brereton R.G.	John Wiley & Sons, 2007	General with applications
Practical Data Analysis in Chemistry	Maeder, M.; Neuhold, Y.-M.	Elsevier. 2007	Hard/Soft modeling
Chemometrics in Spectroscopy	Mark, H.; Workman, J.	Academic Press-Elsevier, 2007	Spectroscopy
Practical Guide to Chemometrics	Gemperline, P. J., ed.	CRC Press, 2006, 2nd ed.	
Multi-way analysis with applications in the chemical sciences	Smilde, A. K.; Bro, R.; Geladi, P.	Wiley, 2004	Multi-way/multilinear modeling
Chemometrics: Data Analysis for the Laboratory and Chemical Plant	Brereton R.G.	John Wiley & Sons, 2003 DOI:https://doi.org/10.1002/0470863242	General, Comprehensive
Multivariate Analysis of Quality: An Introduction	Harald Martens, Magni Martens	John Wiley, 2001) ISBN 0–471–97,428-5 $320.00	Introductory
Hand book of Chemometrics and Qualimetrics: Part A & Part B	Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J.	1998, Elsevier	General, Comprehensive
Chemometric Techniques for Quantitative Analysis	Kramer, R.	CRC Press. (1998)	General, Introductory
Chemometrics: A Practical Guide	Beebe, K. R.; Pell, R. J.; Seasholtz, M. B.	Wiley. (1998)	General, Introductory
Chemometrics in Environmental Analysis	J.W. Einax, H.W. Zwanziger, S. Geiss	VCH (1997) ISBN 3–527–28,772-8 $230.00	Evironmental applications
Factor Analysis in Chemistry	Malinowski, E. R.; Howery, D. G.	New York: Wiley. ISBN 0471058815. 1980,1989, 1991, 2002	Factor analysis
Multivariate Calibration.	Martens, H.; Naes, T.	New York: Wiley, 1989, ISBN 0471909793	Calibration
Chemometrics: a textbook..	Massart, D. L.; Vandeginste, B. G. M.; Deming, S. M.; Michotte, Y.; Kaufman, L.	Amsterdam: Elsevier. 1988 ISBN 0444426604	General, Introductory
Chemometrics	Sharaf, M. A.; Illman, D. L.; Kowalski, B. R.	1986. New York: Wiley. ISBN 0471831069	General, Introductory

controlled experiments, allowing the use and development of new algorithms, such datasets are very much the exception rather than the rule, and a major job is to educate laboratory-based scientists who will have missed out of chemometrics during their basic training.

Heritage Science is another important growth point where chemometrics will play an important role in the future. Especially important is the ability to study works of art non-invasively using methods such as infrared and Raman spectroscopy, hyperspectral imaging, and also X-ray fluorescence, bringing the laboratory to the museum. Underpaintings, restorations, and even forgeries can be uncovered by looking at the layers. Chemometrics has a major role to play resolving and interpreting these complex spectroscopic fingerprints.

While the number of users will expand dramatically, the number of investigators developing fundamental new methods is unlikely to increase. As more established workers retire, or pass away, or move onto other jobs and interests, new colleagues step into their footsteps. The majority of chemometric data is not of sufficient quality to benefit from most of the frontline methods.

In computing and maths, it is well known that most novel approaches remain in dusty books and rarely read papers. But a small number do break through and can be revolutionary. There will always be conferences and fora for the more advanced theoreticians. A few will be lucky enough to work with colleagues who do see the benefit of designing experiments to take advantage of the latest in innovative data analysis. Some emerging areas are listed below.

Combining multivariate methods that take into account interactions between variables as well as experimental factors, such as multilevel approaches, ANOVA-PCA and ASCA is likely to be an important growth area where chemometricians can develop niche methods.

The interaction between the large machine learning community and the chemometrics community is potentially a significant future avenue. Approaches such as support vector machines [41], self-organizing maps [42], and related methods [43] may play a greater role in extracting data from complex analytical measurements. However, an important challenge is the transition of these methods, that have proven highly successful for vast amounts of information-sparse data that has been poorly characterized, to the analysis of analytical data from technology on which much is known about the measurement principle, artifacts, and quantitative methods to remove these. Data, that may also have been collected on experimentation with technical, analytical, and biological replication that is invaluable to assess and characterize a priori the information content in the data. Among all these techniques, support vector machines are gaining more popularity because, by the use of suitable transformations like the ones described in the section Kernels and dissimilarity matrices, they are able to handle non-linear problems conserving many of the advantages of linear-based models [44–46].

Computationally intense approaches are also likely to be of interest to experts. Traditionally, chemometrics was developed in the 1970s and 1980s using very economical methods for validation such as cross-validation. More computationally demanding approaches such as the bootstrap or double cross-validation are now feasible in realistic timescales even for moderately demanding datasets, and have become an important area of development.

These advances, among others, will continue to occupy the minds of pioneering researchers. Different approaches have come in waves, examples being Kalman filters, and wavelets, sometimes filling conferences and journals for a few years, and then step back. Others will develop into fundamental tools of the chemometrician, encountering ever more complicated but in some cases messy data, with, at the same time, ever more access to available computing power.

The future will be exciting, but diverse.

Conclusions

In the previous first part of this feature article (1), we have revised different aspects of chemometrics including aspects like experimental design, sampling, data pre-processing, projection methods, data fusion, and also gave a fast historical overview of the evolution of the field. In this second part, we have covered other important aspects of this subject like modeling, calibration, discrimination, validation, prediction, and revised some classical and recent important applications of chemometrics in the omics and hyperspectral imaging fields. This second feature article ended with a discussion about the perspectives and future evolution of the field. Overall, the two papers have intended to give a fast and easy to read summary of the chemometrics field from the authors' perspective. It is clear however, that this only provides a preliminary overview of the field and that more advanced reading for specific topics is found in more comprehensive works. In Table 2, a list of major reference works and textbooks in chemometrics is given for those interested in a deeper insight of the field. Among them, the Data Handling in Science and technology series and the Comprehensive Chemometrics Chemical and Biochemical Data Analysis major reference work are continuously evolving to cover and summarize the new advances in the chemometrics and related fields. After 40 years of development, chemometrics has become a mature scientific field and the tools developed by chemometricians are helping to solve very challenging analytical problems in different applied fields.

Compliance with ethical standards

References

1. Brereton RG, Jansen J, Lopes J, Marini F, Pomerantsev A, Rodionova O, et al. Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. Anal Bioanal Chem. 2017;409:5891–9.

2. Kalivas JH, Calibration Methodologies in Comprehensive Chemometrics, Brown S, Tauler R, Walczak B (Eds.). Amsterdam:Elsevier; 2009, Vol.3, chapter 3.01.

3. Belsley DA, Kuh E, Welsch RE. Identifying influential data and sources of collinearity. New York: John Wiley & Sons; 1980.

4. Brereton RG. One Class Classifiers. J Chemometr. 2011;25:225–46.

5. Wold S, Sjostrom M. SIMCA: a method for analyzing chemical data in terms of similarity and analogy, in Kowalski, BR (Ed) Chemometrics Theory and Application, American Chemical Society Symposium Series 52, Wash., D.C.:American Chemical Society; 1977, 243–282.

6. Pomerantsev A, OYe R. Concept and role of extreme objects in PCA/SIMCA. J Chemometr. 2014;28:429–38.

7. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugenics. 1936;1936:179M.

8. Barker M, Rayens W. Partial least squares for discrimination. J Chemom. 2003;17:166–73.

9. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. J Chemom. 2014;28:221–35.

10. Rodionova YO, Titova AV, Pomerantsev AL. Discriminant analysis is an inappropriate method of authentication TRAC trends. Anal Chem. 2016;78(4):17–22.

11. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation Chemomet. Intell Lab Syst. 2006;84:69–74.

12. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste B. Sterna C. Anal Chem. 1996;68:3851–8.

13. Serneels S, Filzmoser P, Croux C, Van Espen PJ. Chemometr Intell Lab Syst. 2005;76:197–204.

14. Zerzucha P, Walczak B. Concept of (dis)similarity in data analysis TRAC trends. Anal Chem. 2012;38:116–28.

15. Harshman R. How can I know if it's real? A catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling. In: Law HG, Snyder Jr CW, Hattie J, Mc Donald RP, editors. Research Methods for Multimode Data Analysis. New York: Praeger; 1984. p. 566–91.

16. Westad F, Marini F. Validation of chemometric models—a tutorial. Anal Chim Acta. 2015;893:14–24.

17. Booksh KS, Kowalski BR. Theory of analytical chemistry. Anal Chem. 1994;66(15):782A–91A.

18. Forina M, Lanteri S, Armarino C. Chemometrics in food chemistry, in Chemometrics and species identification. Berlin: Springer; 1987. p. 91–143.

19. Kelly JJ, Barlow CH, Jinguji TM, Callis JB. Ana Chem 1989;61(4);313–320.

20. Wise BM, Gallagher NB. J Process Contr 1996;6(6);329–348.

21. Sharaf MA, Illman DL, Kowalski BR. Chemometrics, chemical analysis, vol. 82. New York: John Wiley and Sons; 1986.

22. Hopke PK. Receptor Modling in Environmental Chemistry, New York: John Wiley Sons; 1981; Hopke PK. Modeling for air quality management, Amsterdam:Elsevier; 1991.

23. Eriksson L, Johansson E. Multivariate design and modeling in QSAR. Chemometr Intell Lab. 1996;34:1–19.

24. Eriksson L, Byrne T, Johansson E, Trygg J, Wikström C. Multi- and megavariate data analysis basic principles and applications, Umeå. 3rd ed. Sweden: Umetrics academy; 2013.

25. Parastar H, Tauler R. Big (bio)chemical data mining using Chemometric methods: a need for chemists. Angew Chem Int. 2018; https://doi.org/10.1002/anie.201801134.

26. Cao K, Lê Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. Bioinformatics. 2011;12:253.

27. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers RJAN, van der Greef J, Timmerman ME. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics. 2005;21:3043–8.

28. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.

29. Gorrochategui E, Jaumot J, Lacorte S, Tauler R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. TrAC - Trends in Anal Chem. 2016;82:425–42.

30. Grahn HF, Geladi P, editors. Techniques and applications of hyperspectral image analysis. Chichester, UK: John Wiley & Sons Ltd; 2005.

31. Geladi P, Grahn H. Multivariate image analysis in chemistry and related areas: chemometric image analysis. Chichester UK: Wiley; 1996.

32. Olmos V, Benítez L, Marro M, Loza-Alvarez P, Piña B, Tauler R, et al. Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images. TrAC-Trends in Anal Chem. 2017;94:130–40.

33. Felten J, Hall H, Jaumot J, Tauler R, de Juan A, Gorzsás A. Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS). Nat Protoc. 2015;10:217–40.

34. Piqueras S, Bedia C, Beleites C, Krafft C, Popp J, Maeder M, et al. Handling different spatial resolution in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets. Anal Chem. 2018;90(11):6757–65.

35. Setou M. (Ed.) Imaging mass spectrometry. Protocols for Mass Microscopy, Berlin:Springer; 2010.

36. Rubakhin SS, Sweedler JV (Eds), mass spectrometry imaging. Principles and protocols. New York: Humana Press; 2010.

37. Bedia C, Tauler R, Jaumot J. Compression strategies for the chemometric analysis of mass spectrometry imaging data. J Chemom. 2016;30:575–88.

38. Skoog DA, West DM, Holler FJ, Crouch SR. Fundamentals of analytical chemistry. Ninth ed. Belmont, CA: Brooks/Cole; 2014.

39. Christian GD, Dasgupta PN, Schug KA. Analytical chemistry. seventh ed. New York: Wiley; 2013.

40. Zomaya AY, Sakr S. Handbook of big data technologies. Berlin: Springer; 2017.

41. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Know Disc. 1998;2:121–67.

42. Kohonen T. Self-Organizing maps. Third ed. Berlin: Springer; 2001.

43. Schmidhuber J. Deep learning in neural networks: an overview http://arxiv.org/abs/1404.7828, 2014.

44. Lutsa J, Ojedaa F, Van de Plasa R, De Moora B, Van Huffel S, Suykens JAK. A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal Chim Acta. 2010;665:129–45.

45. Nia W, Nørgaard L, Mørup M. Non-linear calibration models for near infrared spectroscopy. Anal Chim Acta. 2014;813:1–14.

46. Thissen U, Pepers M, Ustun B, Melssen WJ, Buydens LMC. Comparing support vector machines to PLS for spectral regression applications. Chemometr Intell Lab Syst. 2004;73:169–79.

Springer

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*.
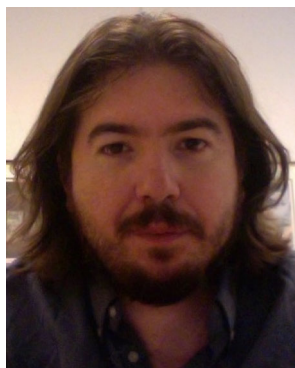
Federico Marini is Researcher and Professor of Chemometrics at the University of Rome La Sapienza. In 2006, he was awarded the Young Researcher Prize from the Italian Chemical Society and in 2012 he won the Chemometrics and Intelligent Laboratory Systems Award. His research activity is focused on all 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, multi-set and multi-way modeling 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.

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 also collaborates intensively with industry, governmental institutions, and developers of analytical technology to apply these methods in production processes, societal safety, and individual health.

Alexey L. Pomerantsev is Principal Researcher at 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).

João A. Lopes is a Chemical Engineer and currently Professor of Pharmaceutics at the Faculty of Pharmacy of University of Lisbon. He obtained a PhD in Chemical Engineering in 2001. His scientific activity has 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 published 92 peer review 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.

Oxana Ye. Rodionova is Leading Researcher at 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 the 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.

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 *Chemometrics and Intelligent Laboratory Systems*. He was 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 Chemometrics, especially in the development of multivariate curve resolution methods. http://www.cid.csic.es/homes/rtaqam/.

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