Accepted Manuscript

Modeling second-order data for classification issues: data characteristics, algorithms, processing procedures and applications

Silvana Azcarate, Adriano de Araújo Gomes, Arsenio Muñoz de la Peña, Héctor C. Goicoechea

PII: S0165-9936(18)30232-2

DOI: [10.1016/j.trac.2018.07.022](https://doi.org/10.1016/j.trac.2018.07.022)

Reference: TRAC 15207

To appear in: *Trends in Analytical Chemistry*

Received Date: 24 May 2018

Revised Date: 27 July 2018

Accepted Date: 29 July 2018

Please cite this article as: S. Azcarate, A. de Araújo Gomes, A. Muñoz de la Peña, H.C. Goicoechea, Modeling second-order data for classification issues: data characteristics, algorithms, processing procedures and applications, *Trends in Analytical Chemistry* (2018), doi: 10.1016/j.trac.2018.07.022.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Modeling second-order data for classification issues: data characteristics, algorithms, processing procedures and applications

Silvana Azcarate,^{a,b} Adriano de Araújo Gomes,^{c,*} Arsenio Muñoz de la Peña^{d, e} and

Héctor C. Goicoechea^{b,f,*}

^a *Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa, and Instituto de Ciencias de la Tierra y Ambientales de La Pampa (INCITAP), Av. Uruguay 151 (6300) Santa Rosa, La Pampa, Argentina*

^b *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290 CABA (C1425FQB), Argentina.*

^c *Departamento de Química Inorgânica – DQI, Instituto de Química – IQ, Universidade Federal do Rio Grande do Sul. Av. Bento Gonçalves 9500 Agronomia 91501970 - Porto Alegre, RS - Brasil.*

^d *Department of Analytical Chemistry, University of Extremadura, 06006, Badajoz, Spain*

^e *Research Institute on Water, Climate Change and Sustainability (IACYS), University of Extremadura, 06006, Badajoz, Spain*

^f *Laboratorio de Desarrollo Analítico y Quimiometría, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Ciudad Universitaria, 3000 Santa Fe, Argentina*

* Corresponding authors: Adriano A. Gomes, E-mail: araujo.gomes@ufrgs.br, Tel.: +55 51 33086862 and Héctor C. Goicoechea, E-mail: hgoico@fbc.unl.edu.ar, Tel.: +54 342 4575206 x190

Abstract

This review outlines some of the challenging aspects of applying second-order data modeling in classification issues analysing selected examples of current applications. The main analytical platforms used for classification are briefly discussed from the application point of view, and the utilization of the generated data is illustrated. After a critical discussion of the advantages concerning the general features of the available algorithms and their underlying models, one example is presented and discussed in detail with the purpose of illustrating the high potentiality of second-order data modeling in the classification field. In addition, advanced data pre-processing tools, prior to multivariate analysis, are explained, and relevant tools are displayed. Finally, novelty prospects in multi-way classification area are presented.

Keywords: Second-order data modeling; Classification; Chromatography; Excitation-emission fluorescence matrix; Three-way

1. Introduction

Classification has been a challenging analytical problem along the time addressed to solve endless issues in many research fields. In the last decades, it has witnessed a progressive increase of data production and storage. In this context, there is an ever-increasing trend in advanced classification procedures to use more and more sophisticated analytical platforms, which generate large and complex data, which in turn require more and more sophisticated data analysis tools for converting data into information, i.e. analytical problems (and, so, the measurements undergone to address them) generate large amounts of data which must be handled afterwards by chemometrics [1].

Diverse available analytical platforms are differing by their sensitivity, selectivity, and high-throughput capacity, but all of them have often the capability to generate several thousand variables per sample. Besides, depending on the technique, the data may have two, three, and more modes (dimensional structure). Therefore, for a single sample it can be recorded data with a two-dimensional (2D) array, which are commonly called second-order data denoting the number of modes for the data array [2-3].

In multivariate classification setting, second-order data handling is producing a significant impact on the development of analytical strategies, especially for determining characteristic patterns of analytes of interest in complex matrices, such as those found in environmental, biological and food samples, among others [4]. For this reason, second-order data modeling applications on the classification context are in growing development.

Although there is a wide variety of analytical techniques that could be used in order to obtain higher-order data, only some of them have been used with the purpose of

classification. In fact, modern analytical platforms of separation followed by different detection techniques, e.g. tandem gas chromatography-mass spectrometry (GC-MS) [5-7], liquid chromatography-mass spectrometry or diode-arrays spectrophotometer (LC-MS [8,9], LC diode-array spectroscopy (LC-DAD) [10,11]), and capillary electrophoresis (CE) with DAD [12], have been scarcely applied. Conversely, fluorescence spectroscopy is the direct detection technique that has been mostly exploited in order to solve classification issues. In fact, this analytical platform is emerging as a competitive technique in the field of characterization due to its moderate selectivity and high sensitivity, which has been applied to the detection of a very wide range of analytes [13]; in addition, it is a relatively fast and inexpensive analytical technique [14,15].

Regarding the use of second-order data for classification studies, the scientific community has mainly focused their concern on the development of analytical methodologies for food authenticity and traceability [4,13]. Some of them have been applied on various foodstuff such as wines [12,16-23], oils [5,24-30], vinegar [31,32], mayonnaise [33], orange juice [34], fish [35], meat [36], tomatoes [8-9], paprika [37], and honey [38,39] in order to determine characteristic patterns of compounds or parameters related to a geographical origin, the adulteration of samples or some specific conditions (e.g. processes, storage, harvest or variety). Other relevant works have been orientated to more specific research, such as the forensic analysis of textile fibers [40,41], the characterization of organic constituents of soils [42], discrimination of geographical origin and detection of adulteration of burdock and kudzu roots [43,44], and for the identification of biodiesel and petroleum oils with environmental purposes [7,45,46]. Several recent reviews have been published about this ever increasing research field where relevant applications can be found [47-53].

Independently of the considered sample or faced classification problem, all applications of second-order classification methods reported in the literature have been focused and applied as a need to handle a data structure of high complexity, when the data set was acquired from a modern analytical platform. Apparently, few reports have motivated their research on the need to increase the data order as a strategy for enhancing the quality of information for a sample set and, subsequently to solve a classification problem. In this context, it is very important to stress that the second-order classification advantage when compared to first-order classification has not been stated yet as was for the calibration world [3].

Referring to data analysis, a wide assortment of chemometric algorithms has been thoroughly implemented and compared [4]. In addition, it has been demonstrated that second-order data modeling can be far more effective when using multi-way methods compared to unfolding procedures (re-arranging the multi-way data into a two-way matrix structure) [4,24,54,55]. In this context, the concept of second-order advantage in classification has been introduced by Escandar et al., although it calls for a more detailed study and discussion [45].

The aim of this review is to provide awareness of the increasing application of second-order data generation and modeling, and their advantages in the classification context. First, a study on the data generation, structure and properties is presented. Then, the most important algorithms for modeling second-order data in classification are carefully analyzed. Remarkable differences in the screening capability of the employed algorithms are shown and discussed. Finally, one case study work applying: a) two-way and b) three-way structured data, is used to show the potentiality of the use of matrix data with focus on classification ability results [33].

2. Second-order data

2.1. Structure and low rank trilinearity property

Samples often belong to complex and heterogeneous systems including mixtures of different molecular families and therefore complicated analytical platforms are required, which should be able to detect the wide range of molecules that might be present at rather different concentration levels [56].

In this context, second-order data can be recorded in two ways: a) direct detection techniques using a single instrument, such as a spectrofluorimeter registering excitation-emission luminescence matrices (EEMs); or b) a separation technique followed by detection dispositive, as GC-MS or LC-DAD, etc. [2,3,57]. Independently of the applied technique, a data matrix (second-order data) is attained for each sample. As can be appreciated in Fig. 1, the data set can be arranged into: (A) a single three-dimensional array or cube, stacking each individual matrix on top of each other [$\mathbf{X}(I \times J \times K)$], or (B) ‘unfolding’ or ‘vectorizing’ the original matrices, which indicates the process of reshaping a matrix in a vector, and is performed by concatenating two-way arrays (matrices). The later process can be performed in six different ways: if I (samples) matrices have dimension J (first mode: e.g. emission wavelength or time) $\times$ K (second mode: e.g. excitation wavelength), the following two way augmented matrices can be derived: $\mathbf{X}(IJ \times K)$, $\mathbf{X}(JI \times K)$, $\mathbf{X}(I \times KJ)$, $\mathbf{X}(I \times JK)$, $\mathbf{X}(K \times IJ)$ and $\mathbf{X}(K \times JI)$. Nevertheless, when dealing with classification issues, it is discernible that the first mode of the data array (the sample dimension I), should be conserved. Consequently, only two kinds of augmented matrices can be built: $\mathbf{Xc}(I \times JK)$ and $\mathbf{Xr}(I \times KJ)$ as displayed in Fig. 1B. As can be seen, the three-way array is represented by $\mathbf{X}$, using both boldface and italics, while the matrix is represented by $\mathbf{X}$, using both boldface and normal letter.

Figure 1. Several sample matrices (I) recorded for classification purposes, with dimension ($J \times K$) can be staked in a three-way structure $X(I \times J \times K)$ (A), or appended, but maintaining I and unfolding the first and second instrumental modes in two different ways: augmented matrices $X_c(I \times JK)$ and $X_r(I \times KJ)$ can be built (B).

When second-order data are processed for a set of samples in the first way mentioned above (Fig. 1.A), it is important if the three dimensional array built with these data fulfills or not with the so-called low rank trilinearity condition, i.e. that the three-way data array built with a set of second-order matrices can be modeled through Eq. (1):

$$X_{ijk} = \sum_{n=1}^N a_{in} b_{jn} c_{kn} + E_{ijk} \quad (1)$$

where N is the total number of chemical constituents generating the measured signal, a_{in} is the relative concentration or score of component n in the i -th sample, and b_{jn} and c_{kn} are the intensities in the instrumental channels (or data dimensions) j and k , respectively. The values of E_{ijk} are the elements of the three-dimensional array E ,

representing the residual error, and having the same dimensions as $\mathbf{X}$. The column vector $\mathbf{a}_n$ (containing the a_n values of the I samples) is collected in the scores matrix $\mathbf{A}$, while vectors $\mathbf{b}_n$ and $\mathbf{c}_n$ are collected in the loading matrices $\mathbf{B}$ and $\mathbf{C}$ (usually both vectors $\mathbf{a}_n$ and $\mathbf{b}_n$ are normalized to unit length). A three-dimensional array built with second-order data matrices will be low rank trilinear whether the signal for a given sample is bilinear and the component profiles are constant across the different samples.

The first condition implies that a single component data matrix can be decomposed into the product of two vectors, each containing the component profile in one of the two data dimensions. An excitation-emission fluorescence matrix represents a well known illustration of bilinear data. It can be decomposed into the excitation and the emission spectra for a given fluorescent component. By contrast, total synchronous fluorescence matrix (TSFS) is not low rank bilinear, because each offset of a single compound has a specific spectrum in the second dimension, making it impossible to describe the TSFS data in terms of one profile in each dimension.

The second requirement implies that the shape of the profiles in all dimensions for a given component must be the same, with intensity variations being due only to different concentrations in different samples.

In those cases in which multi-way data are involved for classification issues, the choice of the appropriate multi-way approach will be decisive in the validity of the solution found. In this context, from a practical point of view, it is of paramount concern whether a three-dimensional data structure $\mathbf{X}$ ($I \times K \times J$) fulfills the requirements of low rank trilinearity. The latter can be attained by inspection of the eigenvalues of matrices $\mathbf{Xc}$ ($I \times JK$) and $\mathbf{Xr}$ ($I \times KJ$) obtained from frontal planes of $\mathbf{X}$ ($I \times K \times J$), as was commented above and shown in Fig. 1.

If the multi-way structure is in fact low rank trilinear, it is expected to obtain equal significant amount of eigenvalues in $\mathbf{Xc}$ and $\mathbf{Xr}$ after principal component analysis. This scenario suggests that the same sources of variance are present in both instrumental modes, and that it maintains the same profile (shape and position) when changing from sample to sample, varying only in intensity. In Fig. 2, the eigenvalues for $\mathbf{Xr}$ and $\mathbf{Xc}$ obtained for a real case involving the analysis of white wine samples by using EEMs [16] (Fig. 2A) and capillary electrophoresis with DAD [12] (Fig. 2B) are shown.

Figure 2: Eigenvalues for $\mathbf{Xc}$ (square-line) and $\mathbf{Xr}$ (circle-line) matrices corresponding to (A) EEMs and (B) CE-DAD data.

As it can be seen in Fig. 2A, the same amount of significant eigenvalues were necessary to explain most of the variability, corroborating the fact that a three-dimensional array built with EEMs data will result low rank trilinear. On the other hand, misalignment and deformation of the peaks in the electropherograms make the three-dimensional array built with CE-DAD data non-trilinear, because a different number of

significant eigenvalues are necessary for matrices $\mathbf{Xr}$ and $\mathbf{Xc}$ (Fig. 2B) in order to explain the same amount of variability.

Interestingly, several issues in the classification world can emerge when the characteristic functions are seriously overlapped and the decomposition is ambiguous (nontrilinear data) [58]. Nevertheless, the real problems investigated and used to write this review present some interesting solutions, showing that in some cases a compromise can be attained.

2.2. Luminescence matrix

Fluorescence is characteristically multidimensional. The fluorescence emission process contains a wealth of independent information related to the fluorophore and its surroundings. A total fluorescence matrix from a sample is obtained by systematically varying the excitation and emission wavelengths [16,19,31,32,41,45,59]. Additionally, the introduction of new commercially available instruments for fluorescence analysis has allowed inquiring into the use of other alternatives such as synchronous and front-face fluorescence spectroscopy. These new approaches have increased the varieties of samples that can be analysed by fluorescence [15,60].

The sampling geometry can have a substantial effect on the obtained fluorescence signal. In fluorescence spectroscopy (FS), the excitation light travels into the sample from one side, and the detector is positioned at right angle to the sample centre. Particularly, front-face fluorescence spectroscopy (FFFS) can be applied when the substances present large absorbance (exceeding the 0.1 value) and significant scattering of light. In this situation, the traditional right angle FS technique cannot be applied due to emission and excitation spectra are both decreased and excitation spectra are distorted [61, 62]. With the object of minimizing these effects, it is recommended

not to place the sample with its surface oriented at an angle of 45° to the incident beam, but rather between 30° or 60° to the light source and the detector [61]. In this context, several researchers have clashed to analysis of turbid or opaque samples, and to avoid these problems have used FFFS, since the signal becomes more independent of the penetration of the light through the sample [20, 26, 33-35, 42].

Figure 3 shows a typical EEM landscape employed for classification of Argentinean white wines according to the grape variety [16]. The data set comprised forty-one EEMs recorded in the excitation range of 245-341 nm each 2 nm and emission range of 375-500 nm each 5 nm.

Figure 3. Typical EEM landscape employed for classification of Argentinean white wines according to the grape variety [16].

On the other hand, in synchronous fluorescence spectroscopy (SFS) the sensitivity associated with fluorescence is maintained while offering a narrowing of spectral bands and an enhancement in selectivity by spectral simplification in multicomponent analysis as main advantage [63]. Spectra are obtained by

simultaneously scanning of both excitation and emission monochromators, keeping a fixed wavelength interval, named offset between them. Inasmuch as SFS technique can be developed on both modes, FS or FFFS, it can be applied on an extensive variety of samples, independently of its environment [36,46]. Therefore, this versatile application is in a recent growth. Interestingly, a total synchronous fluorescence matrix can be obtained recording spectra at different offsets [64], which has been shown good results in classification problems. As was commented above, a three-way array built with these matrices will not be low rank trilinear.

2.3. Separation techniques (liquid and gas chromatography/capillary electrophoresis) coupled to UV-Vis, fluorescence or mass detection

The application of separation techniques with classification purposes has been increased in the last years. Nowadays, the most commercial analytical techniques are based on the separation-coupling concept with chromatography or electrophoresis for the sensitive and selective determination of the non-targeted and/or exhaustive analytes in a sample.

In classical chromatography/electrophoresis, where complete resolution of constituent peaks is required, second-order data generation is not fully necessary, because chromatograms/electropherograms present resolved analyte peak areas. This approach is no longer useful in complex samples where baseline resolution cannot be achieved, and multiple constituents overlap in the elution/migration time mode. Here is when second-order signals generation is needed. However, even when complete resolution is obtained for a given analytical system, sometimes this is at the expense of using complex chromatographic conditions (pH and/or solvent gradients) and/or very

long experimental run times. In these cases, the use of simpler and faster conditions may lead to unresolved analyte peaks in a reasonably short time

Consequently, hyphenation from separation techniques such as GC, LC or CE beside spectroscopic techniques such as DAD or, fast scanning fluorescence detection (FSFD), can provide many possible analytical strategies for classification studies [12,30,37]. Furthermore, due to advanced technologies in mass detectors, in recent years, the coupling of LC, GC or CE with MS has been employed in various classification studies [5–9,65].

Another fact that should be considered when recording and modeling these data, is that some pretreatments such as baseline, centering and misalignment removal are applied in the same way, regardless of whether data are unfolded or retained as three-way (see below).

Contrasting with EEM data, separation measurements always involve some lack of reproducibility in the profiles of the various constituents in the elution time mode. If for some reason they can be decreased to a point where the data can be reasonably considered of low rank trilinear, this is important information to be known because the brand of modeling to be implemented is depending on this property. Usually the problem increases in the following order: gas chromatography < liquid chromatography << capillary electrophoresis.

Figure 4 shows a typical time-wavelength matrix obtained by CE/DAD for a wine sample, with dimensions of 421 elution times (from 1.17 to 7.39 minutes) $\times$ 71 wavelengths (from 227 to 367 nm) [12]. Using these data matrices, it was possible to discriminate samples of Argentinean white wine by a proper modeling, which allowed identifying the loading profiles in spectral mode related to different wine samples.

However, it is relevant to consider that, due to instrumental imperfections, a specific analyte may not elute at the same time in all samples; moreover, when for example MS detection is used, the mass spectrum obtained at this time each fragment can appear at different position (m/z) [66]. For that, a non-trilinear structure is obtained (see above). Taking into account that the subsequent data modeling is highly dependent on their structure, depending on the algorithm to be applied, implementing an adequate data pre-processing is often mandatory before the modeling.

Figure 4. Typical time-wavelength matrix landscape obtained by CE/DAD for a wine sample, with dimensions of 421 elution times (from 1.17 to 7.39 minutes) $\times$ 71 wavelengths (from 227 to 367 nm) [12].

3. Data pre-processing

3.1. Scattering correction in luminescence matrices

Several phenomena related to the nature of the sample as well as the concentration and the molecular environment of the natural fluorophores could affect the fluorescence intensity. The effect of quenching, the concentration, and the molecular

environment of fluorophores plus scattering phenomena have been described in detail in the literature [61,62,67].

One important aspect to be taken into account previous to analyze EEM data is the handling of Rayleigh and Raman scattering [68]. Although, these effects do not own any chemical information about the fluorophores in the solution, they can have a substantial effect on the obtained fluorescence signal and disturb the mathematical modeling of the chemical compounds of interest. Naturally, the fluorophores signal follows a low rank trilinear structure that is scrambled by this phenomenon, generating specific zones where changes in the scattering peaks shape and position by excitation wavelength are produced. For this reason, the scattering peaks do not present a trilinear behavior [68, 69].

Generally, three-way arrays built with EEM data present ideal trilinear structure, which makes such data properly compatible with the parallel factor analysis (PARAFAC) model (Eq. 1). However, when the ideal trilinear structure of EEM data is broken by scattering signal, the PARAFAC model cannot be described by a few factors. Alternatively, it is advisable to remove the scattering before modeling. For that, several established ways to deal scattering effects in relationship to PARAFAC modeling have been proposed in the literature, but only some of them have been applied on the classification field [68, 69].

Accordingly, the simplest implemented strategies, when the affected areas do not contain relevant information, are reducing the areas of analysis, previously to the measurements, selecting the wavelength range from which the scattering effects are not produced [40,41]. However, important information about some sample fluorophores can be lost [41].

Notwithstanding, when the fluorophore signal is interfered for Rayleigh and/or Raman scattering, other perspectives have been proposed in order to conserve the original information, in addition to the removing of the affected areas (Fig. 5-A/B). One of them is inserting missing values inside the scattering areas [17,24,31] (Fig. 5-C); another is replacing the subtracted areas with interpolated values [32, 33]. The interpolation purpose is to replace the data in the affected areas with new data consistent with the data in the remaining parts of the EEM. A critical comparison of how missing or interpolated data can affect the PARAFAC modeling of EEMs has been worthily developed concluding that only interpolated values to remove scattering areas helps PARAFAC to converge faster preventing additionally the convergence to local minima [69]. In fact, both alternatives could be implemented after suitable scattering signal analysis.

Figure 5. (A) EEM registered with Rayleigh scattering (first and second harmonic). (B) Zoom of the zone marked in Fig. 4(A) in which an area was selected with the object of eliminating the scattering effect. (C) Inserting missing values inside the scattering areas. (D) Replacing the subtracted scattering area with interpolated values.

3.2. Corrections applied in separation techniques

Data acquired using hyphenated separation techniques (see above) possess artificial variation related to the noise and other non-sample related variations. Chemometric data analysis methods are sensitive to this type of variations that might hide the true information and lead to misinterpretations. For this reason, it is highly advisable to carry out a data pre-processing prior to application of any classification algorithm. The task of any pre-processing method is to make samples as similar as possible in order to transform the instrumental output into analysable and comparable information for the analytes [70, 71].

In order to attend these aims, a group of methods such as baseline correction, peak alignment, warping, smoothing (noise reduction), scaling, normalization, etc., can be applied as data pre-processing. The most widely implemented strategies, focusing classification analysis, are commented in the following.

3.2.1 Baseline correction

This pre-processing is implemented to reduce the negative influence of a background in a chromatogram, the baseline subtraction can remove systematic artifacts produced by the column or matrix material. Accordingly, to find better estimate of the baseline, numerous algorithms have been discussed in the literature [72-74]. Eilers et al. [72] have introduced an algorithm for baseline elimination based on asymmetric least squares spline regression (AsLSSR) approach, in which the following cost function is minimized:

$$Q = \sum_i v_i (y_i - f_i)^2 + \lambda \sum_i (\Delta^2 f_i)^2 \quad (2)$$

where y is the experimental signal, f is a smooth approximation of baseline trend (y), Δ is the derivative of f , i denotes successive values of the signal, the positive parameter λ

is a regularization parameter that weighs the second term, and ν is weight. The positive deviations from the estimated baseline (peaks) have low ν values, whereas the negative deviations (baseline) obtain high ν values. In the multidimensional extension of baseline correction method, Eilers et al. proposed the splines-based approach for smoothing instrumental signal [the penalty term in Equation (2)]. Details of the implementation of the mentioned method can be found in the literature [72]. As an illustrative example, Fig. 6 shows how the baseline is computed with the described algorithm and subtracted [75].

Figure 6. LC-DAD original chromatogram (blue line), its computed baseline (green line) and the corrected chromatogram obtained by subtracting the background matrix (red line), all of them plotted at 200 nm [75].

3.2.2. Peak alignment and warping

Variations presented in chromatographic/electrophoretic data in retention/migration times should be corrected in some specific cases for developing appropriate multivariate modeling. In order to ensure that the same analyte is correctly

assigned across samples in time and/or to achieve the data low rank trilinearity, these unwanted peak shifts can be corrected using algorithms for retention time alignment.

Numerous methodologies are able to warp the chromatograms by shifting, compressing and/or stretching them along the time axis, varying peak positions and shapes. Several algorithms of this type have been reported in the literature; deserving special attention the following ones: parametric time warping (PTW) [76], peak alignment by means of a genetic algorithm (PAGA) [77], interval correlation optimized shifting (ICOSHIFT) [78], correlation optimized warping (COW) [79], and dynamic time warping (DTW) [79].

The algorithm most used as data preprocessing in order to correct these variations has been COW. This algorithm is based on a piecewise or segmented data preprocessing technique where alignment of a sample chromatogram toward a reference chromatogram is performed by allowing small changes in segment lengths on the sample chromatogram [80]. As an illustrative example, Fig. 7 shows a chromatogram before (A) and after shift correction with COW (B) [81].

Figure 7. Chromatogram before (A) and after COW alignment (B) [81].

3.2.4. Normalization

Usually, normalization is necessary for efficient application of multivariate data analysis methods when different samples are compared. This pre-processing technique corrects the systematic variation occurring among spectra, due to differences in the analyte concentration in the samples, added to variation in the instrument detector sensitivity [82]. Thus, from qualitative data it is possible to extract patterns (relative metabolite intensities) that are found in different classes of samples. Then, for the correction of the overall concentration variations among samples, a scaling factor based normalization method can be performed dividing each analyte in a sample by a factor,

such as unit norm, total area, and total sum of intensities, calculated for a sample [65,71].

4. Algorithms and data modeling

As was commented in Fig. 1, the set of matrices attained each one for each single sample can be arranged in a cube ($\mathbf{X}$) or in an augmented matrix ($\mathbf{Xc}$ or $\mathbf{Xr}$), i.e. unfolding of original matrices. The kind of arrangement and the algorithm to be applied will depend on the kind of data and their inherent properties. For instance, it is not the same building the data set structure with EEMs or GC-MS data, which, as was commented above, bear different properties.

4.1. Unfolding of second-order data followed by traditional two-way algorithms application

It should be important to remark that the unfolding strategy originates loss of the so-called second-order advantage in the calibration world, although it can be say that the advantage of unfolding is that it allows for employing all the many classification methods available for two-way data.

On the other hand, unfolding the array can present some disadvantages: a very large data matrix is often obtained upon unfolding (the matrix will have I objects or samples by JK variables). Therefore, the number of irrelevant variables becomes larger than the really meaningful ones, originating numerical problems in the computation of the models. Additionally, interpretation of the results in terms of the more relevant variables in the definition of the classification model can result difficult, and sometimes, impossible.

Nevertheless, once the unfolding step is performed, the matrices can be processed by the standard two-way classification methods. Following, the most important of them will be briefly described: a) soft independent modeling of class analogies (SIMCA), b) unfolding partial least squares-discriminant analysis (U-PLS-DA) and c) linear discriminant analysis (LDA).

4.1.1. SIMCA

This method is based in that the similarity between samples belonging to a class can be caught by a principal component model of proper dimensionality. Thus, the verification of whether a sample is accepted or rejected by the class model reduces to outlier detection according to some kind of distance to the latent variable model [4]. In order to build the model of a particular category (class A), a principal component analysis (PCA) model is computed based on the data collected on the training samples coming from that class, $\mathbf{X}_A$, according to Equation (3):

$$\mathbf{X}_A = \mathbf{T}_A \mathbf{P}_A^T + \mathbf{E}_A \quad (3)$$

where $\mathbf{T}_A$ and $\mathbf{P}_A$ are the scores and loading matrices for class A, respectively, and $\mathbf{E}_A$ contains the residuals, i.e. the portion of the total variability in $\mathbf{X}_A$ not accounted for by the PC model. Once the PCA is carried out, it is possible to define the distance of the k th sample to the model of the class, $d_{k,A}$ as:

$$d_{k,A} = \sqrt{(\text{OD}_{k,A})^2 + (\text{SD}_{k,A})^2} \quad (4)$$

where $\text{OD}_{k,A}$ and $\text{SD}_{k,A}$ represent the orthogonal and the score distances of the individual to the model of class A, respectively. The first is a measure of the distance of a sample to the PC subspace and is connected to the extent of the residuals. The second specifies

how far the object is from the training samples of the class in the PC space, and is connected to the value of the score vector.

Two statistics are computed: T and Q , which represent the squared Mahalanobis distance of a sample to the center of the score space and the sum of the squared residuals, respectively. They are used to explain for the score and orthogonal distance terms in Equation (4). The distance of the k th sample to the model of the class is expressed as:

$$d_{k,A} = \sqrt{(T^2_{redk,A})^2 + (Q^2_{redk,A})^2} \quad (5)$$

where the subscript 'red' indicates that the variables are normalized by dividing each term for the 95th percentile of the corresponding distribution under the null hypothesis. By definition, since each of the two terms in Equation (5) is divided by the corresponding critical limit, a threshold of $\sqrt{2}$ is normally chosen to assess whether a sample is accepted by the class or not. Indeed, if $d_{k,A} \leq 2$, then the sample k is accepted by the model of class A; otherwise it is rejected. It is important to remark that Equation (5) shows one of the possible implementation of SIMCA classification rule.

4.1.2. U-PLS-DA

The theory of PLS is well known [83]. U-PLS-DA models multidimensional data which data structure is lost by unfolding as was shown in Fig. 1B. It involves a training step in which the relation between the variables and sample codes is estimated from a set of training samples. Then, a prediction step in which the training parameters are used to estimate the codes for unknown samples.

Training provides a loading $\mathbf{Q}$ and a weight loading matrix $\mathbf{W}$ (both of size $JK \times M$, where M is the number of latent factors) and regression coefficients $\mathbf{v}$ (size $M \times 1$). The model is given by Eq. (6):

$$p_{il} = \sum_{m=1}^M t_{im} q_{lm} + e_{il} \quad (6)$$

where p_{il} is the signal measured for sample i and l runs from 1 to (JK) , M is the number of latent variables, t_{im} is an element of the score matrix $\mathbf{T}$, q_{lm} is an element of the loading matrix $\mathbf{Q}$, and e_{il} is a residue not fitted by the model. Prediction of the coded class for an unknown sample (c_u) can be made according to Eq. (7):

$$c_u = \mathbf{t}_u^T \mathbf{v} \quad (7)$$

where $\mathbf{t}_u$ is the test sample score, obtained by projecting the vectorized data for the test sample $[\text{vec}(\mathbf{X}_u)]$, where $\text{vec}(\cdot)$ implies the vectorization operator] onto the space of the M latent factors:

$$\mathbf{t}_u = (\mathbf{W}^T \mathbf{Q})^{-1} \mathbf{W}^T \text{vec}(\mathbf{X}_u) \quad (8)$$

As an example, let us show the work presented by Escandar et al. [45] in which crude oils obtained from different sources of similar geographic origin, as well as light and heavy lubricating oils, were analyzed. Room-temperature phosphorescence matrix signals were recorded and processed by applying discriminant both unfolded and discriminant multidimensional partial least-squares, among other algorithms [45]. U-PLS-DA was applied on a training set with 20 EEMs of oil denominated types 0, 1, 2, and 3. Figure 8 shows predicted vs. nominal coded values for training and test samples. The confidence interval for each category was estimated as the product of the calculated standard deviations of the results for the training samples (0.18, 0.09, 0.17, and 0.19 for María Inés, Medanito, and light lubricating and heavy lubricating oils, respectively) and the Student t -value with $n-1$ degrees of freedom ($t_{(4,0.05)} = 2.78$). As can be clearly seen, the prediction ability of U-PLS-DA algorithm is poor, with a significant overlapping between the crude oil populations. As will be shown below, a better performance of the algorithm maintaining the cube structure is attained in this particular case.

Figure 8. Plot of the U-PLS-DA predicted vs nominal coded values for test samples: María Inés (blue crosses) and Medanita (green crosses) crude oils, and light (red crosses) and heavy (black crosses) lubricants. The geometrical symbols to the right of each group indicate the values of the corresponding training samples and were shifted for clarity. The gray regions indicate the confidence interval for each oil class (see text) [45].

4.1.3. LDA

This is a well-known algorithm, which calculates a surface separating the sample groups by establishing a linear discriminant function that maximizes the ratio of the between class and the within-class variances [84]. Categories are supposed to follow a multivariate normal distribution and be linearly separated. With the $\mathbf{A}(I \times JK)$ (or $I \times KJ$) variable matrix and the dummy matrix $\mathbf{Y}(I \times g)$ of binary digits representing the group assignments (g is the number of categories), the best representation is obtained if the ratio of the between-class variance $\mathbf{Bc}$ matrix and the within-class variance $\mathbf{Wc}$ matrix is maximized. Suitable expressions for the matrices $\mathbf{Bc}$ and $\mathbf{Wc}$ are:

$$\mathbf{Bc} = (g - 1)^{-1} \mathbf{A}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{A} \quad (9)$$

$$\mathbf{Wc} = (I - g)^{-1} [\mathbf{A}^T \mathbf{A} - (g - 1) \mathbf{Bc}] \quad (10)$$

It can be shown that the canonical variate (CV) scores contain the successively maximized between-groups variance/within groups variance. They are obtained by

principal component analysis of the matrix ($\mathbf{W}\mathbf{c}^{-1}\mathbf{B}\mathbf{c}$) and projection of the data matrix $\mathbf{A}$ onto the first loadings. The samples are then plotted on a two or three-dimensional space defined by the first CV scores for each sample. It is important to note that the $\mathbf{A}$ dimension is $I \times JK$; thus, selection of variables, or as will be shown below, data compression and use of scores should be carried out before applying LDA because it should be satisfied that $I > JK$.

4.2. Three-way algorithms for compression followed by two-way algorithms application

The first step is applying an algorithm for data compression. The most popular algorithms are PARAFAC [16], multivariate curve resolution alternating least squares (MCR-ALS) [85], Tucker3 [86], independent component analysis (ICA) [87], among others [88]. However, owing that they are not classifiers, the ultimate goal is reached using a first order classifier (see above) on the output data of the multi-way model.

The first step is to choose the method that can be applied to model the input data that can be arranged in: a) a three dimensional array, or b) an augmented matrix, as was explained in Fig. 1. This step should take into account the data characteristics, for example it should be considered if the three-way array $\mathbf{X}$ ($I \times J \times K$) fulfils the low rank trilinearity property [89]. After that, the second step is to estimate the rank of the data, i.e. to choose the appropriate number of factors from the multi-way model.

Afterward, the matrix of scores can be used to assess similarities and dissimilarities between samples and to build supervised classification models. In this case it is necessary to take into account, that the input data are scores of the multi-way model.

Finally, it should be considered that the validation phase (both internal, e.g. by cross-validation, or external, e.g. by a test set) is a two-steps process as well, first

projecting test (or left out) samples on the compression model to get the scores and then to feed this to the classification model for prediction. In fact, compressing the whole data set at once (especially in the case of internal validation) is a most frequent encountered mistake.

4.2.1. Principal component analysis on multi-way data set

Principal component analysis (PCA) is one of the most used algorithm for previous data analysis in order to evaluate their discriminative power [90]. From PCA, similarities and differences between samples on the bi- or three-dimensional space can be visualized. In addition, the existence of outliers can be also evidenced [91].

The simplest approach for three-way PCA consists on the reduction to conventional PCA, i.e. unfolding of the data array by combining two modes to a single one. Such a procedure seems reasonable in some specific situations like multivariate image analysis, but in general, combined modes do not meet the aim of data reduction. A more advanced way of unfolding which yields separate component matrices for each mode is the Tucker 1 method [90].

PCA multi-way models can be obtained as indicated in Fig. 9A. According to the pictorial description, the three-way array, $\mathbf{X}$ ($I \times J \times K$), is unfolded in a larger matrix $\mathbf{X}$ ($I \times JK$), which is decomposed by traditional PCA. The loading vector, sized ($JK \times 1$), for each principal component is refolded generating a matrix of loadings ($J \times K$). Note that each score interacts with JK combination of loadings. Dissimilar to Tucker3-based models (see below), this multi-way PCA approach does not allow cross interaction between the factors [92].

Figure 9. Pictorial representation of PCA multi-way (A), Tucker3 (B) and PARAFAC models (C).

4.2.2. Tucker 3

In the multi-way universe, the Tucker3 method is presented as a generalization of PCA [93]. A pictorial representation of Tucker3 is displayed in Fig. 9B. Tucker3 algorithm can be expressed by Eq. (11):

$$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk} \quad (11)$$

According to Eq.11, each element (x_{ijk}) of the three-way array (X) corresponds to the multiplication of four values (a , b , c and g) and a small non-modeled part (e_{ijk}), the

noise. The values a , stores in **A** matrix, are related to the samples (equivalent to two-dimensional PCA scores), are the scores Tucker3 model. On the other hand, b and c , stored in **B** and **C** matrices, respectively, are the loadings including the model qualitative information. Unlike PCA (see Fig. 9B), Tucker3 models allow different amounts of factors in each mode (i.e., $P \neq Q \neq R$). A useful method to select the number of factors in Tucker3 is based on eigenvalues inspection of the matrices $\mathbf{X}_A$ ($I \times JK$), $\mathbf{X}_B$ ($J \times IK$) and $\mathbf{X}_C$ ($K \times IJ$), which are obtained with frontal planes of $\mathbf{X}$ ($I \times J \times K$).

The fourth value in Eq. 11 (g) corresponds to the loadings of interactions between the factors. Cross-interactions between factors are indicated by the presence of nonzero elements outside the **G** hyper diagonal (see Fig. 9B). This Tucker3 models characteristic makes it especially useful in the analysis of complex data [12].

4.2.3. PARAFAC

This model is based on the principle of ‘parallel proportional profiles’, i.e. that a set of common factors can be used to describe simultaneously the variation occurring in several matrices although with different weighting coefficients for each matrix.

The PARAFAC model assumes that the data array can be approximated by the low rank trilinear structure defined in Eq. (1). Therefore, the three-way array can be decomposed as a sum of triple outer product of vectors called loadings, which are not orthogonal (displayed in Fig. 9C). Interestingly enough, if the data follow this model, it uniquely reveals the underlying components, meaning that if data follow Beers law, then a PARAFAC model will retrieve the pure spectra even if the three-dimensional structure corresponds to a mixture of substances.

An interesting application of PARAFAC followed of LDA was presented by Muñoz de la Peña and co-workers [40] in a study to explore the potentiality of EEM

microscopy combined with multi-way chemometric analysis for the nondestructive forensic analysis of pairs of visually indistinguishable textile fibers. Figure 10 shows the excellent pairwise discrimination for AY17 and AY23 dyes on nylon-361 fibers. To facilitate the visualization of pairwise comparison, each plot includes the projections of the 95% confidence ellipses over the three planes defined by their corresponding axes. The prediction interval for the multivariate normal distribution yielded an ellipse consisting of $\mathbf{x}$ vectors satisfying the following equation:

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_k^2(p) \quad (12)$$

where $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is the covariance matrix and $\chi_k^2(p)$ is the quantile function for probability p of the chi-squared distribution with k degrees of freedom, where k is the dimension of the data. The axes are defined by the eigenvectors of the covariance matrix and the radius of each axis is equal to 2.796 times the square root of the corresponding eigenvalue. The value 2.796 is obtained from the square root of the Chi-Square distribution with three degrees of freedom and 95% confidence interval [94].

Figure 10. LDA CV scores for a three-components PARAFAC model, for 40 samples of the dyes AY17 and AY23 on nylon-361 fibers. The three dimensional projection of the 95% confidence ellipse of the data collected from each type of fiber is included to facilitate visualization of the obtained results [40].

4.2.4. MCR-ALS

Multivariate curve resolution-alternative least squares (MCR-ALS) is a bilinear decomposition method in compliance with a Beer's law, especially useful for resolution of unknown unresolved mixture data, such as those obtained in separative techniques [95,96]. Data recorded with the latter for a single sample are generally low rank bilinear. However, a three-way structure built with these matrices will be hardly low rank trilinear due to shift and warping of peaks. In this scene, the great advantage of MCR, especially to address classification problems, is the option of handling multiple data matrices simultaneously by building augmented matrices. An augmented matrix ($\mathbf{D}_{\text{aug}}$) is generated by arranging the instrumental response matrices in following ways: a) row-wise direction: one matrix below the other, or b) column-wise direction: one matrix next to the other. The direction of the augmentation must considerate the

instrumental mode which breaks the low rank trilinearity. For a single instrumental matrix $\mathbf{D}$ ($K \times J$), MCR can be mathematically expressed as:

$$\mathbf{D} = \mathbf{C} \mathbf{S}^T + \mathbf{E} \quad (13)$$

where, $\mathbf{C}$ ($K \times A$) and $\mathbf{S}^T$ ($A \times J$) are related to the concentration and spectral profiles respectively, and $\mathbf{E}$ ($K \times J$) is the fraction of the data not modeled by A factors. From a practical point of view, the inspection of the eigenvalues of matrix $\mathbf{D}$ is the most popular way to obtain the A value.

For an augmented matrix $\mathbf{D}_{\text{aug}}$, Eq. 13 assumes new dimensions, according to the direction of the augmentation. For example, Fig. 11 shows three instrumental matrices of unresolved mixtures of three components ($A = 3$) arranged in column-wise direction. The dimensions are the following: $\mathbf{D}_{\text{aug}}$ ($IJ \times K$) (for $I = 3$ samples), $\mathbf{C}_{\text{aug}}$ ($IJ \times 3$), $\mathbf{S}^T$ ($3 \times K$) and $\mathbf{E}_{\text{aug}}$ ($IJ \times K$).

Figure 11: A pictorial representation of the MCR decomposition.

An interesting literature example is the metabonomic study carried out for rambo and raf tomatoes cultivars treated with carbofuran pesticide [9]. The authors applied MCR-ALS to second-order data sets obtained by LC-MS. Samples were picked up during a 21 days period after treatment and analyzed by LC-MS in scan mode, along with the corresponding blank samples. Then, the data sets were modeled using column wise augmented matrices, and the evolutionary profiles as a function of the time after treatment were estimated for the metabolites present in both cultivars, as well as their corresponding pure spectra estimations. Using the areas under the evolutionary profiles (see Fig. 12), PLS-DA was used to check if it would be possible to classify samples by using the metabolites MCR estimations. Results showed that PLS-DA models for classification of treated or non-treated (blank) samples were the best ones obtained (98.44% of correct classifications for the validation set).

Figure 12. Evolutionary profiles of four components (one per region in which the chromatogram was divided for easy interpretation) through the eight sampling days in rambo treated (—), raf treated (—) and their respective non treated samples, rambo (----) and raf (----) [9].

4.3. Three-way algorithms

Up to day, there are not too much reports based on applications of strictly multi-way classification algorithms. Three methods can be mentioned as the most applied: a) multi-way soft independent modeling of class analogies (N-way SIMCA), b) partial least squares discriminant analysis (N-PLS-DA), and c) two-dimensional linear discriminant analysis (2D-LDA). They will be described below.

4.3.1. N-way SIMCA

The first application of a SIMCA-based approach with data decomposition using multi-way algorithms (Tucker3 in this case) was reported by Flåtten and co-authors [97] for quantification of pollution levels. On the other hand, a complete characterization from the statistical point of view of SIMCA methodologies for multi-way data has been introduced by Cocchi et al.[98]. Multi-way data can be decomposed by Tucker3 or PARAFAC, as shown in Fig. 9, for each target class of the training data. The choice to use PARAFAC or Tucker3 is made considering whether the data are low rank trilinear or not (see above suggested ways to determinate this condition). The best number of factors in each class can be accessed by sensitivity (*SEN*) and specificity (*SPE*) values attained for each model varying the number of factors from 1 to the highest number. After selecting the proper factor number for each class, classification rules are defined [99]. In Section 5, the mentioned figures of merit are commented.

Contrasting with other works, this approach is fully devoted to classification problems. A routine and user manual are available in <http://www.models.life.ku.dk/nsimca>. This strategy is of great importance for case studies involving adulteration, allowing the development of multi-way one-class models.

4.3.2. N-PLS-DA

The use of partial least squares for classification purposes (PLS-DA) has been well characterized by Barker and Rayens [100] and most recently by Brereton [101]. As was commented before, the basic idea consists in a training step in which the relation between the variables and sample codes is estimated from a set of training samples. Then, a prediction step in which the training parameters are used to estimate the codes for unknown samples. PLS-DA for first order data has been applied in several classification approaches in chemistry as well as other fields [102].

A relevant work detailing the theoretical aspects of the use of N-PLS-DA was presented by Ouertain *et al* [103] demonstrating a compact formulation of N-PLS, computational aspects about the calculation of scores, loadings and convergence of the algorithm in the discriminant analysis context. The N-PLS-DA model is given by the following equation:

$$p_{ijk} = \sum_{m=1}^M t_{im} w_{jm}^J w_{km}^K + e_{ijk} \quad (14)$$

where p_{ijk} is the signal measured for sample i at emission sensors j and k , w_{jm}^J and w_{km}^K are elements of two $\mathbf{W}$ loading matrices, and e_{ijk} is a residue not fitted by the model. The model finds the scores producing maximum covariance with the sample coded class (dependent variable), in a three-dimensional sense. Subsequently, the sample scores are used in a similar manner to Eq. (7) to predict unknown sample codes. The advantage of using N-PLS-DA over bidimensional regression is a stabilization of the decomposition involved in Eq. (14), which potentially gives increased interpretability and improved predictions.

Interestingly, Arancibia *et al.* introduced the idea of second-order advantage in classification problems by means of residual bilinearization (RBL). The presence of an

unexpected constituent in a test sample is detected by comparing the residual of this sample to the typical instrumental noise, estimated for the samples of the calibration set. Then, the scores of the test sample are modified so that the sample is allocated in its correct class [45]. In the later work, N-PLS-DA was applied training with the 20 same EEMs of oil denominated types 0, 1, 2, and 3 than those used for PLS-DA (see Fig. 11). Figure 13 shows the confidence intervals (computed in same way than for U-PLS-DA) for the oil classes and the predicted test samples. The comparison of the results obtained from the unfolded and the multidimensional PLS algorithms demonstrates the importance of maintaining the tridimensional structure of the data for the sample cluster resolution. Although the EEMs for the different oil types are very similar, the results reveal that that these data carry enough information to perform a successful supervised classification of samples [45].

Figure 13. Plot of the DN-PLS predicted vs nominal coded values for test samples: María Inés (blue crosses) and Medanito (green crosses) crude oils, and light (red crosses) and heavy (black crosses) lubricants. The geometrical symbols to the right of each group indicate the values of the corresponding training samples and were shifted for clarity. The gray regions indicate the confidence interval for each oil class [45].

4.3.3. Two-dimensional linear discriminant analysis

Two-dimensional linear discriminant analysis (2D-LDA) is an algorithm developed by Li and Yuan [104] for image feature extraction for classification issues. Recently, it has been implemented by Silva and co-authors for modeling chemical information in classification problems involving three-way spectral data [105]. The main difference between LDA and 2D-LDA is that the latter operates on matrices instead of vectors. A set of $\mathbf{X}$ ($J \times K$) sample matrices corresponding to the training set must be arranged to generate an augmented matrix $\mathbf{A}$ of dimension ($IJ \times K$) or ($IK \times J$) (see Fig.14). Then, the matrix of feature information $\mathbf{P}$ is obtained as:

$$\mathbf{P} = \mathbf{A} \mathbf{Z} \quad (15)$$

The $\mathbf{P}$ matrix contains the canonical variate (CV) scores. In addition, $\mathbf{Z}$ corresponds to the first A loadings of the matrix ($\mathbf{W}\mathbf{c}^{-1}\mathbf{B}\mathbf{c}$) obtained by PCA decomposition, in compliance with LDA. $\mathbf{B}\mathbf{c}$ and $\mathbf{W}\mathbf{c}$ matrices are in line with equations 9 and 10. A pictorial representation for 2D-LDA is shown in Fig. 14.

Figure 14. Pictorial representation of training step in 2D-LDA for two target classes and a prediction of a single unknown sample. In this illustrative case $A = 3$.

Unlike LDA, in 2D-LDA the matrix $\mathbf{A}$ is sized ($IJ \times K$ or $IK \times J$), thus $IJ > K$ or $IK > J$, avoiding previous steps of variables selection or data compression by PCA. In the prediction step, samples are assigned to a target class based on the similarity index, defined as the lowest mean Euclidean distance of the test sample from the training set. This stage is not present in the LDA, in which the samples are classified on the basis of a position with respect to the class boundary in CV scores space.

It should be indicated that unfolding, in any case $IJ \times K$ and $IK \times J$ would give different results when modeling chemical information. Thus, an alternative would be to choose the most discriminating mode as a practical solution to the problem.

5. Assessing the classification models

5.1. Confusion matrix

Classification results can be represented in a so-called confusion matrix, which is a square matrix with dimensions ($G \times G$), containing information about actual and predicted classifications done by a classifier that only assigns each sample to a single one of the available classes. A general representation of a confusion matrix is given in Fig. 15. In this table, G is the number of classes, each element n_{Gk} represents the number of samples belonging to class G and assigned to class k . The diagonal values (for example n_{GG}) correspond to the number of correct predictions, whereas the values outside the diagonal resemble misclassified samples. Thus, the classification performance can be assessed from the samples information contained in this confusion matrix.

		Predicted class				
		1	2	3	...	G
Actual class	1	n_{11}	n_{12}	n_{13}	...	n_{1G}
	2	n_{21}	n_{22}	n_{23}	...	n_{2G}
	3	n_{31}	n_{32}	n_{33}	...	n_{3G}
	...	...	...	...	...	...
	G	n_{G1}	n_{G2}	n_{G3}	...	n_{GG}

Figure 15. Confusion matrix for a dataset with G classes.

5.2. Figures of merit

In order to estimate the model fit, the well-known cross-validation procedure, also called internal validation, is the method commonly employed [106]. Otherwise, the performance or predictive ability of the model can be evaluated analyzing some figures of merit, which can be obtained by an external validation using an independent test set and building the confusion matrix.

In this context, the most common figures of merit for a classification model are: accuracy (AC), precision (P), sensitivity (S) and specificity (SP), which are of great usefulness for comparison of different models [107].

Each sample corresponding to internal and external validation sets is assigned to a default class. Then, it can be classified as true positive (TP), true negative (TN), false positive (FP) or false negative (FN). Being TP the number of positive samples that are correctly identified as positive samples, FN the number of positive samples that are

misclassified as negative samples, FP the number of negative samples that are incorrectly identified as positive samples, and TN the number of negative samples that are correctly identified as negative samples. Therefore, the figures of merit for a classifier using these parameters can be mathematically defined as it is shown in Table 1. Thus, S is the proportion of positives that are correctly identified; SP is the proportion of negatives that are correctly identified; AC characterizes the correct classification rate, i.e. the percentage of samples correctly classified; and P is the ratio between the number of samples correctly accepted and the total number of samples accepted by the same model [108, 109].

5.3. Area under ROC (Receiver Operating Characteristics) curves

On the other hand, a highly useful tool for evaluating model efficiency and making decisions about optimum models is the so-called Receiver Operating Characteristics (ROC) curves. They are usually obtained by plotting the sensitivity values against $(1 - \text{specificity})$ following alterations in the width of model boundaries (i.e. the confidence level). A diagonal line from the lower left corner to the upper right corner represents models that provide a random assignment of samples, while curves that lie above that line and tend towards the upper left corner indicate efficient models (Fig. 16). Such curves can be used both to compare different models and to identify the optimal compromise between sensitivity and specificity for a given model. The area under the curve (AUC) is often taken as a comprehensive scalar indicator of model performance. $AUC = 0.5$ means a poorly performing model with random sample assignment, while $AUC = 1$ indicates a model of maximum efficiency. Some caution has to be observed when ROC analysis is employed to evaluate model performances

under varying class distributions, since the outcomes depend on variations of true-positive and false-positive rates [110].

Figure 16. Graphical representation of ROC curves indicating the model performance presenting A) maximum ($AUC = 1$) B) medium ($0.5 < AUC < 1$) and C) poor ($AUC = 0.5$) efficiency.

6. Illustrative example

A current application will be explained and illustrated to furnish guidelines for the different steps in building multi-way classification models, including data pretreatments, and model validation. Previously described methods were applied on two data sets consisting in first-order and second-order fluorescence data with the aim of monitoring spoilage on mayonnaise as a quality control strategy [33].

The obtained models from these data sets were compared and an unswerving discussion concerning three-way method application advantages is performed. Consequently, this approach may be useful in cases where the information provided by first-order data is unsatisfactory and/or inadequate to achieve a good classification.

6.1. Data sets

Both data sets refer to storage time identification of homemade and commercial mayonnaise samples maintained at 37 °C in the following time range: A (0-15 h), B (16-39 h), C (40-63 h), D (64-87 h), and E (88 h-112 h). Occurred spectral changes on a

homemade mayonnaise samples during storage can be observed by visual inspection of EEMs in Fig. 17. It can be clearly appreciated the variations evidenced by mayonnaise components over the time. These assessments can be valued since fluorescence spectra provide information about the fluorescent molecules environment.

Figure 17. Contour plots of EEMs from homemade mayonnaise sample stored at 37 °C at the first minute of analysis, after 24, 48, and 87 h, respectively. Second-order (up) and first-order (bottom) fluorescence data (registered at maximum excitation wavelength of 295 in the range 300 to 600 nm) [33].

Mayonnaise sample description and the conditions under which the samples were collected are specified in [33]. In this case, only the lot stored at 37 °C containing 80 samples was analyzed. Then, the samples were grouped taking in account two different time periods for each range.

Fluorescence excitation spectra were recorded between 230 and 400 nm (increment 5 nm), and emission wavelengths from 300 to 600 nm, spaced by 0.5 nm interval in the emission domain. Thus, collected fluorescence matrices were combined in a three-dimensional array (80 samples $\times$ 302 emission wavelengths $\times$ 35 excitation wavelengths).

On the other hand, in order to obtain fluorescence first-order data, a vector was extracted of each data matrix from a slice of the three-way array $\mathbf{X}$ at maximum

excitation wavelength of 295 nm. Therefore, extracted fluorescence vectors were grouped in a two-dimensional array (80 samples $\times$ 302 emission wavelengths). Characteristic first- and second-order data signals of the mayonnaise samples are displayed in Fig. 18.

Figure 18. Characteristic mayonnaise fluorescence signal (A) EEM obtained from a single sample (B) emission at 295 nm acquired from a sample set.

In order to build classification models, the data set was partitioned into two samples groups employing the Kennard–Stone algorithm, which is a random selection approach for dataset splitting. This algorithm is aimed at covering the multidimensional space in a uniform manner by maximizing the Euclidean distances between the instrumental response vectors of the selected samples [111,112]. Thus, complete data set ($n = 80$) was split into training ($n = 60$, twelve of each class) and prediction ($n = 20$, four of each class) subsets, respectively. The training sample set was used for data modeling and internal validation procedure. The prediction sample set, which constitutes an independent set, was used to test the predictive ability of the models.

6.2. Preprocessing

EEM landscapes bears affected areas by first- and second-order Rayleigh scatter. The latter are artifacts of higher value than chemical signals and sometimes affect key part of them, regarding chemical information (see Section 3.1. *Scattering correction in luminescence matrices*). Thus, it is of paramount importance considering a preprocessing strategy in order to preserve the chemical information. For this reason, with the object of reducing the scattering effects, the EEMs were preprocessed replacing scattering areas with interpolated data [68]. Afterward, areas were cut and subsequently, replaced with interpolated values, as shown in Fig. 19.

Figure 19. An example of the EEM of one mayonnaise sample landscape: (A) EEM after scattering areas has been set to missing. (B) The same EEM after interpolation.

6.3. Data analysis: a comparative study between first and second-order data

In order to further assess the ability of both data sets (first- and second-order), supervised models were built from individual training sets employing partial least squares discriminant algorithms for first- (PLS-DA) and second-order data (N-PLS-DA), respectively. Taking into account the multiclass approach, the models were calculated applying PLS2 algorithm considering as many Y-variables as categories.

Thus, the samples were split into five classes (A, B, C, D and E), being each sample assigned to the class corresponding to the predicted **Y** column. The optimal latent variables (LV) number model was selected considering the minimum values of the classification error in cross-validation (using venetian-blind cross-validation). Thus, the optimal LVs for both models were fourteen and three for first and second-order data, respectively.

Afterward, data order (first and second) influence onto resulting models was evaluated, making a comparison of their discriminative ability. The obtained confusion matrix from attained fitting models according to the five different classes, are presented in Table 2. As it can be clearly observed, the multi-way approach leads to excellent classification rates in the training stage (almost 100 %), conversely to first-order data from which a bad classification has been achieved (lesser than 50 %). Based on these outcomes, it can be concluded that first-order data seem to be inappropriate to solve the proposed complex classification problem. In general terms, all the classes exhibited large error rates. Remarkably, for class B (range 40-63 h), not a single sample has been correctly assigned.

Subsequently, the achieved models were used to predict the independent four samples five test sets. The estimation of the classification performances for each class of both data sets are reported in Table 3. In addition, the attained global classification rates (AC) were 97, 91 and 100 for second-order data, for training set (*Train*), cross-validation (*CV*) and test set (*Pred*), respectively. On the other hand, for first-order data, the AC obtained were 48, 52 and 35.

As a conclusion, models built using second-order data and maintaining the three-way structure result highly specific and sensitive than those built with first-order data. Moreover, the N-PLS-DA method is more parsimonious since less number of LV were

necessary (3 against 14) to explain data variability, i.e. a large amount of PLS-DA LVs cannot accomplish a good mayonnaise samples classification. Besides, all samples were correctly assigned when analysing independent test sets with N-PLS-DA. Obviously, the obtained high error rate and poor figures of merit for PLS-DA model from samples prediction are expected insomuch as the first-order data analysis furnished a dreadful calibration stage.

The scores plots projecting training and prediction samples for the first three PLS factors of PLS-DA and N-PLS-DA models are displayed in Fig. 20A-B. As can be clearly seen, in the PLS-DA model all five classes are superimposed (fig. 20A). On the contrary, the N-PLS-DA model (Fig. 20B) shows a larger contribution of the three factors for discriminating the mayonnaise samples with respect to storage time variation is observed.

Figure 20. Projections of training and prediction samples scores on the models (A) PLS-DA, and (B) N-PLS-DA, depicting the first three PLS factors in three-dimensional score plots. Blue (class A), red (class B), green (class C), yellow (class D) and magenta (class E). The three dimensional projection of the 95% confidence ellipse of the data collected from each class is included to facilitate visualization of the obtained results.

Accordingly, comparing the intrinsic information that can provide both data sets (first- and second-order data), the emission spectra recorded from mayonnaise samples at a single excitation wavelength would not afford a complete characterization of the fluorescent compounds present in the samples owing to the chemical compounds fluoresce at different excitation wavelengths [67]. In this case, the second-order data modeled with a multi-way structure show an improvement in the classification quality,

which can be attributed to an increase of amount of provided data containing more information that leads to better samples characterization.

The later observation agree with Salvatore *et al.* who highlighted how beneficial it is to take into account the multi-way nature of the data array. In their practical and interesting work about classification methods of multi-way arrays, the authors focused and compared meticulously several applications of the different methods that can be applied to second-order data, and how they can impact in extracting the relevant information to model the class feature, distorting, compressing or losing information in the final outcomes [4].

7. Conclusion

Classification has been a challenging analytical concern along the time, addressed to solve a large number of discrimination issues in a wide variety of fields. In this context, application of second-order data modeling could supply interesting improvements as regards results attained when extremely complex systems should be classified (such as those found for instance in environmental, biological and food samples). Interestingly, until now, most of the literature reports have been focused in modeling a second-order data acquired from modern analytical techniques. It seems to be that few reports have been motivated their research on the need to increase the data order as a strategy for enhancing the quality of information extracted, and, consequently, to better solve a classification problem.

In this review, some of the challenging aspects of second-order data applied to classification problems have been outlined, discussing the main analytical platforms used for classification. One illustrative example allowed us to show how the ability in terms of discrimination power, in this particular case, can be improved by using second-

order data arranged in a three-way structure vs. first-order data. Finally, it should be commented that some of the presented methods, such as N-way SIMCA and N-PLS-DA are not limited to second order and can be applied to data array of any order.

Acknowledgements

The authors express their gratitude to CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP-2015 N° 0111) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project PICT 2014-0347) of Argentina, and Ministerio de Economía, Industria y Competitividad (Project CTQ2017) and Junta de Extremadura (Project GR18041) of Spain, for financially supporting this work. The authors are grateful PROPIT/Unifesspa for financial support.

References

- [1] T. Skov, A. H. Honoré, H. Max Jensen, T. Næs, S. B., Engelsen, Chemometrics in foodomics: handling data structures from multiple analytical platforms, Trends Anal. Chem. 60 (2014) 71–79.
- [2] J.A. Arancibia, P.C. Damiani, G.M. Escandar, G.A. Ibañez, A.C. Olivieri, A review on second- and third-order multivariate calibration applied to chromatographic data, J. Chromatog. B, 910 (2012) 22– 30.
- [3] G.M. Escandar, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, Second- and higher-order data generation and calibration: A tutorial, Anal. Chim. Acta 806 (2014) 8– 26.
- [4] E. Salvatore, M. Bevilacqua, R. Bro, F. Marini, M. Cocchi, Classification methods of multiway arrays as a basic tool for food PDO authentication, Compr. Anal. Chem. 60 (2013) 339-379.

- [5] P.S. Prata, G.L. Alexandrino, N.G.S. Mogollón, F. Augusto, Discriminating brazilian crude oils using comprehensive two-dimensional gas chromatography – mass spectrometry and multiway principal component analysis, *J. Chromatogr. A* 11 (2016) 99-106.
- [6] D. Ebrahimi, J. Li, D. Brynn Hibbert, Classification of weathered petroleum oils by multi-way analysis of gas chromatography–mass spectrometry data using PARAFAC2 parallel factor analysis, *J. Chromatogr. A* 1166 (2007) 163–170.
- [7] W. F. de Carvalho Rocha, M.M. Schantz, D.A. Sheen, P.M. Chu, K.A. Lippa, Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data. *Fuel* 197 (2017) 248–258.
- [8] I. Sánchez Pérez, M.J. Culzoni, G.G. Siano, M.D. Gil García, H.C. Goicoechea, M. Martínez Galera, Detection of Unintended Stress Effects Based on a Metabonomic Study in Tomato Fruits after Treatment with Carbofuran Pesticide. Capabilities of MCR-ALS Applied to LC-MS Three-Way Data Arrays, *Anal. Chem.* 81 (2009), 8335–8346.
- [9] G.G. Siano, I. Sánchez Pérez, M.D. Gil García, M. Martínez Galera, H.C. Goicoechea, Multivariate curve resolution modeling of liquid chromatography–mass spectrometry data in a comparative study of the different endogenous metabolites behavior in two tomato cultivars treated with carbofuran pesticide, *Talanta* 85 (2011) 264–275.
- [10] B. Schmidt, J.W. Jaroszewski, R. Bro, M. Witt, D. Stærk, Combining PARAFAC Analysis of HPLC-PDA Profiles and Structural Characterization Using HPLC-PDA-SPE-NMR-MS Experiments: Commercial Preparations of St. John’s Wort, *Anal. Chem.* 80 (2008) 1978-1987.

- [11] Yong Wang, Baohui Li, Yongnian Ni, Serge Kokot, Multi – wavelength high – performance liquid chromatography: An improved method for analysis of complex substances such as Radix Paeoniae herbs, *Chemometr. Intell. Lab. Syst.* 130 (2014) 159–165.
- [12] S.M. Azcarate, A. Araújo Gomes, L. Vera-Candioti, M.C. Ugulino de Araújo, J.M. Camiña, H.C. Goicoechea, Second-order capillary electrophoresis diode array detector data modeled with the Tucker3 algorithm: A novel strategy for Argentinean white wine discrimination respect to grape variety, *Electrophoresis* 37 (2016) 1902–1908.
- [13] R. Karoui, C. Blecker, Fluorescence Spectroscopy Measurement for Quality Assessment of Food Systems—a Review, *Food Bioprocess Technol.* 4 (2011) 364–386.
- [14] C.A. Stedmon, S. Markager, R. Bro, Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy, *Mar. Chem.* 82 (2003) 239–254.
- [15] K. Kumar, M. Tarai, A. K. Mishra, Unconventional steady-state fluorescence spectroscopy as an analytical technique for analyses of complex-multifluorophoric mixtures, *Trends Anal. Chem.* 97 (2017) 216–243.
- [16] S.M. Azcarate, A. Araújo Gomes, M.R. Alcaraz, M.C. Ugulino de Araújo, J.M. Camiña, H.C. Goicoechea, Modeling excitation–emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety, *Food Chem.* 184 (2015) 214–219.
- [17] S. Elcoroaristizabal, R.M. Callejón, J.M. Amigo, J.A. Ocaña-González, M.L. Morales, C. Ubeda, Fluorescence excitation–emission matrix spectroscopy as a tool for determining quality of sparkling wines, *Food Chem.* 206 (2016) 284–290.

- [18] C. Yin, H. li, C. Ding, H. Wang, Preliminary Investigation on Variety, Brewery and Vintage of Wines using Three-dimensional Fluorescence Spectroscopy, *Food Sci. Technol. Res.* 15 (2009) 27-38.
- [19] D. Airado-Rodriguez, T. Galeano-Diaz, I. Duran-Meras, J.P. Wold, Usefulness of fluorescence excitation-emission matrices in combination with PARAFAC, as fingerprints of red wines, *J. Agric. Food Chem.* 57 (2009) 1711–1720.
- [20] D. Airado-Rodriguez, I. Duran-Meras, T. Galeano-Diaz, J.P. Wold, Front-face fluorescence spectroscopy: a new tool for control in the wine industry, *J. Food Compos. Anal.* 24 (2011) 257–64.
- [21] C. Schueuermann, P. Silcock, P. Bremer, Front-face fluorescence spectroscopy in combination with parallel factor analysis for profiling of clonal and vineyard site differences in commercially produced Pinot Noir grape juices and wines, *J. Food Compos. Anal.* 66 (2018) 30-38.
- [22] J. Sádecká, M. Jakubíková, P. Májek, Fluorescence spectroscopy for discrimination of botrytized wines, *Food Control.* 88 (2018) 75–84.
- [23] M. Cabrera-Bañegil, M. del C. Hurtado-Sánchez, T. Galeano-Díaz, I. Durán-Merás, Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in red wine samples, *Food Chem.* 220 (2017) 168–176.
- [24] F. Guimet, J. Ferré, R. Boqué, F.X. Rius, Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation–emission matrix fluorescence spectroscopy, *Anal. Chim. Acta* 515 (2004) 75–85.
- [25] A. Kassouf, M. El Rakwe, H. Chebib, V. Ducruet, D.N. Rutledge, J. Maalouly, Independent components analysis coupled with 3D-front-face fluorescence

spectroscopy to study the interaction between plastic food packaging and olive oil, *Anal. Chim. Acta* 839 (2014) 14–25.

[26] P. Valderrama, P.H. Marçó, N. Locquet, F. Ammari, D.N. Rutledge, A procedure to facilitate the choice of the number of factors in multi-way data analysis applied to the natural samples: Application to monitoring the thermal degradation of oils using front-face fluorescence spectroscopy, *Chemometr. Intell. Lab. Syst.* 106 (2011) 166–172.

[27] F. Guimet, J. Ferré, R. Boqué, Rapid detection of olive–pomace oil adulteration in extra virgin olive oils from the protected denomination of origin “Siurana” using excitation–emission fluorescence spectroscopy and three-way methods of analysis, *Anal. Chim. Acta* 544 (2005) 143–152.

[28] F. Guimet, R. Boqué, J. Ferré, Cluster analysis applied to the exploratory analysis of commercial Spanish olive oils by means of excitation–emission fluorescence spectroscopy, *J. Agric. Food Chem.*, 52 (2004) 6673–6679.

[29] F. Guimet, R. Boqué, J. Ferré, Application of non-negative matrix factorization combined with Fisher’s linear discriminant analysis for classification of olive oil excitation-emission fluorescence spectra, *Chemom. and Intell. Lab. Syst.*, 81 (2006) 94–106.

[30] I. Durán Merás, J. Domínguez Manzano, D. Airado Rodríguez, A. Muñoz de la Peña, Detection and quantification of extra virgin olive oil adulteration by means of autofluorescence excitation-emission profiles combined with multi-way classification, *Talanta* 178 (2018) 751–762.

[31] R.M. Callejón, J.M. Amigo, E. Pairoc, S. Garmón, J.A. Ocaña, M.L. Morales, Classification of Sherry vinegars by combining multidimensional fluorescence, parafac and different classification approaches, *Talanta* 88 (2012) 456–462.

- [32] R. Ríos-Reina, S. Elcoroaristizabal, J.A. Ocaña-González, D.L. García-González, J.M. Amigo, R.M. Callejón, Characterization and authentication of spanish PDO wine vinegars using multidimensional fluorescence and chemometrics, *Food Chem.* 230 (2017) 108–116.
- [33] S.M. Azcarate, C.M. Teglia, F. Karp, J.M. Camiña, H.C. Goicoechea, A novel fast quality control strategy for monitoring spoilage on mayonnaise based on modeling second-order front-face fluorescence spectroscopy data, *Microchem. J.* 133 (2017) 182–187.
- [34] F. Ammari, L. Redjda, D.N. Rutledge, Detection of orange juice frauds using front-face fluorescence spectroscopy and Independent Components Analysis, *Food Chem.* 168 (2015) 211–217.
- [35] A. Hassoun, R. Karoui, Monitoring changes in whiting (*Merlangius merlangus*) fillets stored under modified atmosphere packaging by front face fluorescence spectroscopy and instrumental techniques, *Food Chem.* 200 (2016) 343–353.
- [36] A. Sahar, U. Rahman, A. Kondjoyan, S. Portanguen, E. Dufour, Monitoring of thermal changes in meat by synchronous fluorescence spectroscopy, *J. Food Eng.* 168 (2016) 160–165.
- [37] O. Monago-Maraña, T. Galeano-Díaz, A. Muñoz de la Peña, Chemometric discrimination between smoked and non-smoked paprika samples. Quantification of PAHs in smoked paprika by fluorescence-U-PLS/RBL, *Food Anal. Meth.*, 10 (2017) 1128–1137.
- [38] R. Karoui, E. Dufour, J.O. Bosset, J. De Baerdemaeker, The use of front face fluorescence spectroscopy to classify the botanical origin of honey samples produced in Switzerland, *Food Chem.* 101 (2006) 314–323

- [39] L. Lenhardt, R. Bro, I. Zekovic, T. Dramicanin, M.D. Dramicanin, Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey, *Food Chem.* 175 (2015) 284–291.
- [40] A. Muñoz de la Peña, N. Mujumdar, E.C. Heider, H.C. Goicoechea, D. Muñoz de la Peña, A.D. Campiglia, Nondestructive Total Excitation–Emission Fluorescence Microscopy Combined with Multi-Way Chemometric Analysis for Visually Indistinguishable Single Fiber Discrimination, *Anal. Chem.* 88 (2016) 2967–2975.
- [41] A.D. Campiglia, M. Rex, A. Muñoz de la Peña, H.C. Goicoechea, Excitation–emission matrix fluorescence spectroscopy combined with MCR-ALS as a tool for the forensic analysis of similar and dissimilar sets of textile fiber extracts. *Anal. Meth.* 8 (2016) 8314–8321.
- [42] F. Ammari, R. Bendoula, D. Jouan-Rimbaud Bouveresse, D.N. Rutledge, J.M. Roger, 3D front face solid-phase fluorescence spectroscopy combined with independent components analysis to characterize organic matter in model soils, *Talanta* 125 (2014) 146–152.
- [43] L. Hu, C. Yin, S. Ma, Z. Liu, Tracing the geographical origin of burdock root based on fluorescent components using multi-way chemometrics techniques, *Microchem. J.* 137 (2018) 456–463.
- [44] L. Hu, S. Ma, C. Yin, Discrimination of geographical origin and detection of adulteration of kudzu root by fluorescence spectroscopy coupled with multi-way pattern recognition, *Spectrochim. Acta A* 193 (2018) 87–94.
- [45] J.A. Arancibia, C.E. Boschetti, A.C. Olivieri, G.M. Escandar, Screening of oil samples on the basis of excitation–emission room- temperature phosphorescence data and multiway chemometric techniques. Introducing the second-order advantage in a classification study, *Anal. Chem.* 80 (2008) 2789–2798.

- [46] M. Insausti, A. de Araújo Gomes, J.M. Camiña, M.C. Ugulino de Araújo, B.S. Fernández Band, Fluorescent fingerprints of edible oils and biodiesel by means total synchronous fluorescence and Tucker3 modeling, *Spectrochim. Acta Mol. Biomol. Spectrosc.* 175 (2017) 185–190.
- [47] J. Sádecká, J. Tóthová, Fluorescence spectroscopy and chemometrics in the food classification - a Review, *Czech J. Food Sci.* Vol. 25 (2007) 159–173.
- [48] R. Karoui, C. Blecker, Fluorescence Spectroscopy Measurement for Quality Assessment of Food Systems-a Review, *Food Bioprocess Technol.* 4 (2011) 364–386.
- [49] A.M. Gómez-Caravaca, R.M. Maggio, L. Cerretani, Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review, *Anal. Chim. Acta.* 913 (2016) 1–21.
- [50] I. Reinholds, V. Bartkevics, I.C.J. Silvis, S.M. van Ruth, S. Esslinger, Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: A review, *J. Food Compos. Anal.*
- [51] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, *Food Control.* 86 (2018) 283–293.
- [52] L.A. Berrueta, R.M. Alonso-Salces, K. Héberger, Supervised pattern recognition in food analysis, *J. Chromatogr. A.* 1158 (2007) 196–214.
- [53] E. Borrás, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment - A review, *Anal. Chim. Acta.* 891 (2015) 1–14.
- [54] A. Smilde, R. Bro, P. Geladi, *Multi-way Analysis. Application in the Chemical Sciences*, WILEY, John Wiley & Sons, Ltd, England, 2004.

- [55] R. Leardi, C. Armanino, S. Lanteri, L. Albertonanza, Three-way principal component analysis for monitoring data from Venice lagoon, *J. Chemometrics* 14 (2000) 197–211.
- [56] A. Cifuentes, *Foodomics: Advanced Mass Spectrometry in Modern Food Science and Nutrition*, WILEY, John Wiley & Sons, Ltd, England, 2013.
- [57] J.M. Amigo, F. Marini, Chapter 7 - Multiway Methods, in: F. Marini (Ed.), *Data Handling in Science and Technology*, E-Publishing Inc., Amsterdam, 2013, pp. 283–309.
- [58] R. Bro (1998) Multi-way analysis in the food industry. Models, algorithms and applications, Doctoral Thesis, University of Amsterdam, The Netherlands.
- [59] A. Andrade-Eiroa, M. Canle, V. Cerdá, Environmental applications of excitation–emission spectrofluorimetry: an in-depth review II, *Appl. Spectrosc. Rev.* 48 (2013) 77–141.
- [60] M. Cabrera-Bañegil, D. Martín-Vertedro, E. Boselli, I. Durán-Merás, Control of olive cultivar irrigation by front-face fluorescence excitation-emission matrices in combination with PARAFAC, *J. Food Compos. Anal.*, 69 (2018) 189-196
- [61] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, 3rd ed., Springer Science + Business Media, New York, 2006.
- [62] R. Karoui, C. Blecker, Fluorescence Spectroscopy Measurement for Quality Assessment of Food Systems - a Review, *Food Bioprocess. Tech.* 4 (2011) 364–386.
- [63] M.E. Pacheco, L. Bruzzone, Synchronous fluorescence spectrometry: Conformational investigation or inner filter effect, *J. Lumin.* 137 (2013) 138–142.
- [64] K. Kumar, A.K. Mishra, Application of parallel factor analysis to total synchronous fluorescence spectrum of dilute multifluorophoric solutions: Addressing the issue of

lack of trilinearity in total synchronous fluorescence data set, *Anal. Chim. Acta* 755 (2012) 37-45.

[65] B. Khakimov, G. Gürdeniz, S.B. Engelsen, Trends in the application of chemometrics to foodomics studies, *Acta Aliment.* 44 (2015) 4–31.

[66] T. Skov, S.B. Engelsen, Chemometrics, mass spectrometry, and foodomics, in: A. Cifuentes (Ed.), *Foodomics: Advanced Mass Spectrometry in Modern Food Science and Nutrition*, John Wiley & Sons Inc., Hoboken, N.J., 2013, pp. 507-538.

[67] J. Christensen, L. Nørgaard, R. Bro, S. B. Engelsen, Multivariate Autofluorescence of Intact Food Systems, *Chem. Rev.* 106 (2006) 1979-1994.

[68] M. Bahram, R. Bro, C. Stedmon, A. Afkhami, Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation, *J. Chemometrics* 20 (2006) 99–105.

[69] S. Elcoroaristizabal, R. Bro, J.A. García, L. Alonso, PARAFAC models of fluorescence data with scattering: A comparative study, *Chemom. Intell. Lab. Sys.* 142 (2015) 124–130.

[70] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, *Chem. Rev.* 110 (2010) 4582–4605.

[71] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Oršić, Algorithms and tools for the preprocessing of LC-MS metabolomics data, *Chemom. Intell. Lab. Sys.* 108 (2011) 23–32.

[72] P.H.C. Eilers, A perfect smoother, *Anal. Chem.* 75 (2003) 3631–3636.

[73] M.N. Leger, A.G. Ryder, Comparison of derivative preprocessing and automated polynomial baseline correction method for classification and quantification of narcotics in solid mixtures, *Appl. Spectrosc.* 60 (2006) 182–93.

- [74] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, *Chemom. Intell. Lab. Syst.* 82 (2006) 59–65.
- [75] C.M. Teglia, P.M. Peltzer, S.N. Seib, M.J. Culzoni, H.C. Goicoechea, Simultaneous multi-residue determination of twenty one veterinary drugs in poultry litter by multivariate modeling of liquid chromatography with fluorescence and UV detection data, *Talanta* 167 (2017) 442-452.
- [76] P. H. C. Eilers, Parametric time warping. *Anal. Chem.* 76 (2004) 404-411.
- [77] J. Forshed, I. Schuppe-Koistinen, S. P. Jacobsson, Peak alignment of NMR signals by means of a genetic algorithm, *Anal. Chim. Acta* 487 (2003) 189-199.
- [78] G. Tomasi, F. Savorani, S. B. Engelsen, *icoshift*: An effective tool for the alignment of chromatographic data. *J. Chromatogr. A* 1218 (2011) 7832-7840.
- [79] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, *J. Chemometrics* 18 (2004) 231–41.
- [80] B. Schmidt, J.W. Jaroszewski, R. Bro, M. Witt, D. Stærk, Combining PARAFAC analysis of HPLC-PDA profiles and structural characterization using HPLC-PDA-SPE-NMR-MS experiments: commercial preparations of St. John's Wort, *Anal. Chem.* 80 (2008) 1978-1987.
- [81] C. M. Teglia, S. M. Azcarate, M. R. Alcaráz, H. C. Goicoechea, M. J. Culzoni, Exploiting the synergistic effect of concurrent data signals: low-level fusion of liquid chromatographic with dual detection data, *Talanta* <https://doi.org/10.1016/j.talanta.2018.04.090>.
- [82] P. Filzmoser, B. Walczak, What can go wrong at the data Normalization step for identification of biomarkers?, *J. Chromatogr. A* 1362 (2014) 194–205.
- [83] H. Martens, T. Naes, *Multivariate calibration*, Wiley, Chichester, UK, 1989.

- [84] E. K. Kemsley, A genetic algorithm (GA) approach to the calculation of canonical variates (CVs), *Trends Anal. Chem.* 1998, 17, 24-34.
- [85] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: A review of advanced and tailored applications and challenges, *Anal. Chim. Acta* 765 (2013) 28– 36.
- [86] M. Kompany-Zareha,, Y. Akhlaghia, R. Bro, Tucker core consistency for validation of restricted Tucker3 models, *Anal. Chim. Acta* 723 (2012) 18– 26.
- [87] D.N. Rutledge, D. Jouan-Rimbaud Bouveresse, Independent Components Analysis with the JADE algorithm, *Trends Anal. Chem.* 50 (2013) 22–32.
- [88] H. Jiang, L. Zhang, J. Xia, A new array decomposition method for multiway data analysis, *Chemom. Intell. Lab. Sys.* 101 (2010) 56–71.
- [89] J.M.F. ten Berge, Simplicity and typical rank of three-way arrays, with applications to Tucker-3 analysis with simple cores, *J. Chemometrics* 18 (2004) 17–21.
- [90] R. Henrion, N-way principal component analysis Theory, algorithms and applications, *Chemom. Intell. Lab. Sys.* 25 (1994) 1-23.
- [91] C.B.Y. Cordella, R. Leardi, D.N. Rutledge, Three-way principal component analysis applied to noodles sensory data analysis, *Chemom. Intell. Lab. Sys.* 106 (2011) 125–130.
- [92] H.A.L. Kiers, A. der Kinderen, A fast method for choosing the numbers of components in Tucker3 analysis, *Br. J. Math. Stat. Psychol.* 56 (2003) 119–125.
- [93] V. Pravdova, F. Estienne, B. Walczak 1, D.L. Massart, A robust version of the Tucker3 model, *Chemom. Intell. Lab. Sys.* 59 (2001) 75–88.
- [94] M. Slotani, Tolerance regions for a multivariate normal population *Ann. Ins. Stat. Math*, 16 (1964) 135-153.
- [95] S. C. Rutan, A. de Juan, R. Tauler. Introduction to Multivariate Curve Resolution. *Comprehensive Chemometrics*. Elsevier, (2009). Vol 3.

- [96] H. Parastar, R. Tauler. Multivariate Curve Resolution of Hyphenated and Multidimensional Chromatographic Measurements: A New Insight to Address Current Chromatographic Challenges. *Anal. Chem.* 86 (2014) 286–297.
- [97] G. Rune Flåten, B. Grung, O.M. Kvalheim, Quantification of pollution levels by multiway modelling, *J. Chemometrics* 18 (2004) 173–182.
- [98] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, *Chemom. Intell. Lab. Syst.* 106 (2011) 73–85.
- [99] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, *J. Chemometrics* 22 (2008) 601–609.
- [100] M. Barker, W. Rayens, Partial least square for discrimination, *J. Chemometrics* 17 (2003) 166–173.
- [101] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, *J. Chemometrics* 28 (2014) 213–225.
- [102] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding, *Anal. Chim. Acta* 879 (2015) 10–23.
- [103] S.S.Ouertani, G. Mazerolles, J. Boccard, S. Rudaz, M. Hanafi, Multi-way PLS for discrimination: compact form equivalent to the trilinear LLS2 procedure and its monotony convergence, *Chem. Intell. Lab. Syst.* 133 (2014) 25–32.
- [104] M. Li, B. Yuan, 2D-LDA: A statistical linear discriminant analysis for image matrix, *Pattern Recognit. Lett.* 26 (2005) 527–532.
- [105] A.C. da Silva a, S.F.C. Soares, M. Insausti, R.K.H. Galvao, B.S.F. Band, M.C.U. de Araújo, Two-dimensional linear discriminant analysis for classification of three-way chemical data, *Anal. Chim. Acta* 938 (2016) 53–62.

- [106] R.G. Brereton. Validation and optimization. In *Chemometrics for Pattern Recognition*. John Wiley & Sons, Ltd: Chichester, 2009; 311–391.
- [107] D. Szollosi, D.L. Denes, F. Firtha, Z. Kovacs, A. Fekete, Comparison of six multiclass classifiers by the use of different classification performance indicators, *J Chemometrics* 26 (2012) 76-84.
- [108] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, *Trends Anal. Chem.* 35 (2012) 74-86.
- [109] M. Forina, S. Lanteri, S. Rosso, Confidence intervals of the prediction ability and performance scores of classifications methods, *Chemom. Intell. Lab. Syst.* 57 (2001) 121-32.
- [110] Ballabio D., and Todeschini R. 2009. Multivariate classification for qualitative analysis. In *Infrared spectroscopy for food quality analysis and control*, ed. S. Da-Wen, 83-104, Amsterdam: Elsevier.
- [111] R.K. Harrop Galvao, M.C. Ugulino Araujo, G.E. José, M. J. Coelho Pontes, E.C. Silva, T.C. Bezerra Saldanha, A method for calibration and validation subset partitioning, *Talanta* 67 (2005) 736–740.
- [112] K.R. Kanduc, J. Zupan, N. Majcen, Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment, *Chemometr. Intell. Lab. Syst.*, 65 (2003) 221-229.

Figure captions

Figure 1. Several sample matrices (I) recorded for classification purposes, with dimension ($J \times K$) can be stacked in a three-way structure $X(I \times J \times K)$ (A), or appended, but maintaining I and unfolding the first and second instrumental modes in two different ways: augmented matrices $Xc(I \times JK)$ and $Xr(I \times KJ)$ can be built (B).

Figure 2: Eigenvalues for Xc (square-line) and Xr (circle-line) matrices corresponding to (A) EEMs and (B) CE-DAD data.

Figure 3. Typical EEM landscape employed for classification of Argentinean white wines according to the grape variety [16].

Figure 4. Typical time-wavelength matrix landscape obtained by CE/DAD for a wine sample, with dimensions of 421 elution times (from 1.17 to 7.39 minutes) $\times$ 71 wavelengths (from 227 to 367 nm) [12].

Figure 5. (A) EEM registered with Rayleigh scattering (first and second harmonic). (B) Zoom of the zone marked in Fig. 4(A) in which an area was selected with the object of eliminating the scattering effect. (C) Inserting missing values inside the scattering areas. (D) Replacing the subtracted scattering area with interpolated values.

Figure 6. LC-DAD original chromatogram (blue line), its computed baseline (green line) and the corrected chromatogram obtained by subtracting the background matrix (red line), all of them plotted at 200 nm [74].

Figure 7. Chromatogram before (A) and after COW alignment (B) [80].

Figure 8. Plot of the U-PLS-DA predicted vs nominal coded values for test samples: María Inés (blue crosses) and Medanito (green crosses) crude oils, and light (red crosses) and heavy (black crosses) lubricants. The geometrical symbols to the right of each group indicate the values of the corresponding training samples and were shifted for clarity. The gray regions indicate the confidence interval for each oil class (see text) [45].

Figure 9. Pictorial representation of PCA multi-way (A), Tucker3 (B) and PARAFAC models (C).

Figure 10. LDA CV scores (3 components PARAFAC model) for 40 samples of the dyes AY17 and AY23 on nylon-361 fibers. The three dimensional projection of the 95%

confidence ellipse of the data collected from each type of fiber is included to facilitate visualization of the obtained results [40].

Figure 11: A pictorial representation of the MCR decomposition.

Figure 12. Evolutionary profiles of four components (one per region in which the chromatogram was divided for easy interpretation) through the eight sampling days in rambo treated (—), raf treated (—) and their respective non treated samples, rambo (- - -) and raf (----) [9].

Figure 13. Plot of the DN-PLS predicted vs nominal coded values for test samples: María Inés (blue crosses) and Medanito (green crosses) crude oils, and light (red crosses) and heavy (black crosses) lubricants. The geometrical symbols to the right of each group indicate the values of the corresponding training samples and were shifted for clarity. The gray regions indicate the confidence interval for each oil class [45].

Figure 14. Pictorial representation of training step in 2D-LDA for two target classes and a prediction of a single unknown sample. In this illustrative case $A = 3$.

Figure 15. Confusion matrix for a dataset with G classes.

Figure 16. Graphical representation of ROC curves indicating the model performance presenting A) maximum ($AUC = 1$) B) medium ($0.5 > AUC < 1$) and C) poor ($AUC = 0.5$) efficiency.

Figure 17. Contour plots of EEMs from homemade mayonnaise sample stored at 37 °C at the first minute of analysis, after 24, 48, and 87 h, respectively. Second-order (up) and first-order (bottom) fluorescence data (registered at maximum excitation wavelength of 295 in the range 300 to 600 nm) [33].

Figure 18. Characteristic mayonnaise fluorescence signal (A) EEM obtained from a single sample (B) emission at 295 nm acquired from a sample set.

Figure 19. An example of the EEM of one mayonnaise sample landscape: (A) EEM after scattering areas has been set to missing. (B) The same EEM after interpolation.

Figure 20. Projections of training and prediction samples scores on the models (A) PLS-DA, and (B) N-PLS-DA, depicting the first three PLS factors in three-dimensional score plots. Blue (class A), red (class B), green (class C), yellow (class D) and magenta (class E). The three dimensional projection of the 95% confidence ellipse of the data collected from each class is included to facilitate visualization of the obtained results.

Table 1. Figures of merit applied in the classification model validation

Figures of merit	Equation ^a	Definition
Accuracy (AC)	$AC = (TP+TN)/(TP+TN+FP+FN)$	Percentage of samples correctly classified.
Precision (P)	$P = TP/(TP+FP)$	Probability that a positive decision will be correct.
Sensitivity (S)	$S = TP/(TP+FN)$	Samples correctly classified within the class.
Specificity (SP)	$SP = TN/(TN+FP)$	Samples correctly classified that do not belong to the class.

^a TP: true positive, TN: true negative, FP: false positive, FN: false negative.

Table 2. Confusion matrices corresponding to N-PLS-DA and PLS-DA mayonnaise classification results from training set.

Predicted by N-PLS-DA ^a					
Class	A	B	C	D	E
A	12	0	0	0	0
B	2	10	0	0	0
C	0	0	12	0	0
D	0	0	0	12	0
E	0	0	0	0	12

Predicted by PLS-DA ^a					
A	7	3	1	1	0
B	0	8	0	3	1
C	1	7	0	2	2
D	0	0	0	7	5
E	0	0	0	5	7

^a Number of samples predicted in each class of a set of 12 by cross-validation.

Table 3. Figures of merit of the individual models in training (Train) and prediction (Pred) stages, as a function of the fluorescence data order implemented.

Class	A		B		C		D		E	
	Train	Pred	Train	Pred	Train	Pred	Train	Pred	Train	Pred
N-PLS-DA										
<i>SP</i>	96	100	100	100	100	100	100	100	100	100
<i>S</i>	100	100	83	100	100	100	100	100	100	100
<i>P</i>	86	100	100	100	100	100	100	100	100	100
PLS-DA										
<i>SP</i>	90	100	79	75	98	94	77	88	83	63
<i>S</i>	58	25	67	50	0	0	58	50	58	50
<i>P</i>	88	100	44	33	0	0	39	50	47	25

Specificity (*SP*), sensitivity (*S*) and precision (*P*) of the individual class models.

Highlights

- 1) Challenging aspects of applying second-order data modelling in classification issues.
- 2) The main analytical platforms used for classification are briefly discussed.
- 3) Critical discussion of the advantages concerning the available algorithms.
- 4) One example to illustrate the high potentiality of higher-order modelling.