IRIS

INSTITUTIONAL RESEARCH INFORMATION SYSTEM
ARCHIVIO ISTITUZIONALE DEI PRODOTTI DELLA RICERCA

intestazione repositorydell'ateneo

Classification Methods of Multiway Arrays as a Basic Tool for Food PDO Authentication

Publisher copyright

(Article begins on next page)

18 September 2017

Chapter 14

C0070 # Classification Methods of Multiway Arrays as a Basic Tool for Food PDO Authentication

Elisa Salvatore*,†, Marta Bevilacqua†, Rasmus Bro‡, Federico Marini† and Marina Cocchi*

*Department of Chemistry, University of Modena and Reggio Emilia, Via Campi 183, Modena, Italy
†Department of Chemistry, University of Roma La Sapienza, Piazzale Aldo Moro, 5, Roma, Italy
‡Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg, Copenhagen, Denmark

s0005 ## 1 MULTIWAY METHODS IN FOOD AUTHENTICATION

p0005 The aim of this chapter is to provide awareness of the increasing applications of multiway methodology in the food authenticity context. The growing use of hyphenated analytical techniques in food characterization, such as non-destructive, information-rich techniques, generates high-order data arrays. Also, higher-order arrays are generated when food samples are characterized by measuring properties that vary with ageing, storage time and/or different

1

stages of processing. Here, we would like to stress, on the one hand, the advantages of using multiway tools for multiway data at both the explorative and recognition levels and, on the other, furnish some guidelines to tackle the issues connected with the practical applications of these methods, highlighting their feasibility.

p0010 The main features and advantages of the multiway approach are presented through some case study applications, all concerning regional food protected by designation of origin, focusing on presentation and interpretation of the results.

p0015 First, the authenticity context is explored. Au1

s0010 1.1 Authenticity Issues

p0020 An important branch in food research is the control of food authenticity. In this context, typical food products take a place of particular interest both for economic reasons and for their particular traditional process. In particular, the high commercial value of traditional foods is one of the main reasons for their counterfeiting and/or marketing of 'sounding like products', from other regions or countries. Such products are generally of inferior quality or adulterated.

p0025 The European Union (EU) has developed legal rules for defining and protecting the authenticity of products [1]. Under this PDO (protected designation of origin) system, a named food or drink registered at the European level is given legal protection against imitation throughout the EU. The PDO certification requires that *the product must be produced, processed and prepared in a particular geographical area*. Nowadays, there is also growing attention among consumers to high-quality food with a clear regional identity. Also, the producers have chosen to stress the product territoriality as a quality indicator in order to strengthen their marketing strategies.

p0030 Consequently, it emerges quite clearly that there is a need to find objective criteria in order to support the certification of authenticity and the real provenance of products. The European scientific community has focused many research efforts on the development of more objective analytical methodologies for food authenticity and traceability [2–4].

p0035 So analytical methods have to address two main issues: detecting possible product adulteration and assessing product authenticity. The latter task is particularly challenging since geographical origin, raw materials and transformation processes have to be taken into account. All these aspects introduce distinct sources of variability, leading to rather complex systems from the point of view of characterization and even more of data analysis. Last but not least, some typical foods, which are characterized by an artisanal procedure [5], have to be regarded from a completely different perspective from industrial products. In fact, by definition, while a well-codified production protocol is followed, more variability due to the small scale at which they are produced is introduced as compared to industrial processing in a plant. This means that quality cannot be described simply in terms of conformity

to a given set of features as it is usually done for industrial processing. This also implies that analytical methods and data analysis tools commonly used in food quality and process control have to be re-evaluated and modified to fit these new tasks.

s0015   1.2   Multiway Characterization in Food Authenticity Context

p0040   Recently, there has been a progressive change in the approach to the study of foodstuff. Foodstuff studies have evolved from development of dedicated methods for quantification of specific classes of constituents to the acquisition of a fingerprint. In other words, instrumental techniques are employed for rapid, non-destructive and non-selective sample characterization. This approach relies on data-driven discovery, that is, information is obtained by understanding the underlying relationships among variables as highlighted by data analysis *a posteriori* [6]. Authenticity encompasses various aspects: complete characterization of the product, identification of adulteration, control of compliance with the label and assessment of geographical origin. These aspects make it almost mandatory to use a fingerprint approach. In fact, the identification of specific markers for authenticity proof, for example, a quantitative determination of some of the identified components, may surely help to discriminate products although it has some limitations. More promising is the use of models that rely on chemometrics and consider the contemporary contribution of multiple effects. In fact, while frauds may be committed by altering the amount of the single components, it is unlikely that a whole instrumental profile could be artificially imitated.

p0045   Second-order analytical techniques, such as gas chromatography–mass spectrometry (GC–MS), LC–MS, HPLC-DAD and fluorescence emission/excitation spectroscopy, are often used in authenticity studies [7–9]. In particular, fluorescence spectroscopy is emerging as a competitive technique in the field of characterization and classification of intact food such as honey, cheese, wine and meat [10,11]. Second-order data also arise when the studied phenomena vary over time, location or processing conditions. Thus, multiway data analysis is gaining more and more interest in food authenticity applications, and it has been demonstrated that data analysis can be more effective when using multiway methods on multiway data compared to unfolding procedures, that is, rearranging the multiway data into a two-way matrix structure. Unfolding methods may result in more complex models with an associated risk of poor predictive ability. Unfolding may also be less efficient in terms of capturing and interpreting the underlying structure in a data set. On the contrary, the use of multiway methods potentially simplifies the interpretation of the results and provides more adequate and robust models using relatively few parameters.

p0050   Truly multiway classification methods have been developed only recently [12–14]. In this chapter, we describe the main ones and compare them with the approach of first unfolding and then performing two-way classification.

p0055 Moreover, considering that the typical questions in food traceability, authenticity and quality control are "Can sample X, stated of class *A*, really belong to class *A*?" or "Does sample X follow specifications?", class-modelling techniques such as N-SIMCA [12] are discussed with respect to discriminant ones.

s0020 2 METHODS

p0060 Multiway data arrays typically arise when a set of samples has been character- Au2 ized by two distinct sets of variables such as chemical profiles and sensory analysis, when a set of samples has been characterized by a set of variables taken at different occasions such as at a different times [15], different sampling sites [16] and different targets, when samples are characterized by 2D or 3D data such as GC–MS, HPLC–UV, 2D and 3D NMR spectroscopy and so on [8,17]. Historically, such data have been treated by unfolding methods, that is, rearranged in a two-way data matrix and treated by bilinear models. The results have sometimes been rearranged into the original multiway structure for interpretation. Nowadays, there is an established set of chemometric multiway methods and algorithms [12,13,18,19] that are able to directlyhandle the multiway nature of these data sets.

p0065 In the food area, multiway problems arise in a number of ways, ranging from storage/ageing (samples × variables × time) to sensory analysis (samples × attributes × judges), batch data (batches × time × variables) and, of course, handling data from hyphenated analytical technique systems.

p0070 The most used multiway decomposition methods are parallel factor analysis (PARAFAC) and Tucker3, and we limit our methodological description to these in Sections 2.2.1 and 2.2.2, respectively. Both PARAFAC and Tucker3 methods accomplish data compression and allow explorative analysis of multiway data arrays. Moreover, they serve as a basis for classification tools such as N-SIMCA and NPLS (Section 2.3).

p0075 Section 2 is organized as follows: Section 2.1 illustrates the unfolding plus bilinear classifier approach, Section 2.2 covers the approach using bilinear classifiers on the outcome of multiway decomposition and finally, Section 2.3 illustrates truly multiway classification methods.

s0025 2.1 Unfolding + Two-Way Classification

p0080 When dealing with a multiway data structure with the aim of classification, the simpler way to transfer the standard chemometric classification technique is to rearrange the multiway data array into a data matrix that can be processed by the classical chemometric two-way classifiers.

p0085 This starting step is commonly called 'unfolding' or, in assonance with the term 'vectorizing' indicating the process of reshaping a matrix in a vector,

'matricizing' and is performed by concatenating two-way array (matrices) extracted from the data array.

p0090   This process can be performed in different ways. A three-way data array can be rearranged in six different types of two-way matrices. In detail, defining $\underline{\mathbf{X}}(I \times J \times K)$, the three-way array of size $(I \times J \times K)$, the following two-way 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 $\mathbf{X}(K \times JI)$.

p0095   However, when dealing with classification issues, that is, with problems in which samples of different origins have to be assigned to different groups on the basis of measured signals, it is obvious that the first mode of the data array, namely, the 'sample dimension', has to be preserved. Therefore, only $\mathbf{X}(I \times KJ)$ and $\mathbf{X}(I \times JK)$ unfolding can be performed. In both cases, one speaks of 'row-wise unfolding' and the result is a matrix having as many rows as the number of samples (i.e. where each row represents the complete signal measured on an individual), while the number of columns is equal to the product of the other two dimensions. In particular, the former produces a matrix in the columns of which the original index k runs fastest and j the slowest, while the latter brings a configuration in which the index j runs fastest and k slowest. When, for instance, a data set from hyphenated chromatography (e.g. GC–MS) is considered, the previous statement is translated into the possibility of matricizing the array by juxtaposing one another for each sample, either the chromatograms recorded at the different m/z or the mass spectra registered at the various retention times (as illustrated in Figure 1); in the context of classification, both these approaches are equally valid.

p0100   It must be pointed out that the unfolding strategy leads to the loss of the so-called second-order advantage that is inherent in certain three-way models. This property, however, is mostly of interest in calibration problems, and the advantage of unfolding is that it allows for employing all the many classification methods available for two-way data. Unfolding the array can present some drawbacks, resulting from the fact that a very large data matrix is often obtained upon unfolding. Instead of having J plus K variables, the unfolded
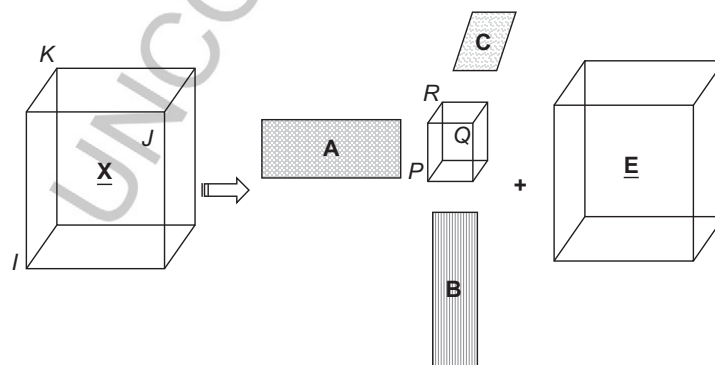
f0005   **FIGURE 1**   Unfolding procedure applied to both the WINE and EVOO data set.

matrix will have JK variables, which is typically much larger. When the number of irrelevant or noisy variables becomes large with respect to that of the really meaningful ones, numerical problems in the computation of the models can occur or, even if a solution for the problem is found, it can be unstable or unreliable. The more serious these issues, the higher the extent of the noise affecting the modelling. Moreover, interpretation of the results in terms of the portions of the signal being more relevant in the definition of the classification model becomes more difficult (if not almost impossible).

p0105 Having these considerations in mind, after the initial unfolding step is carried out as explained, the matrices can be processed by the standard two-way classification methods. In particular, in the remainder of this section, partial least squares-discriminant analysis (PLS-DA) and soft independent modelling of class analogies (SIMCA) are described in detail, as examples of the most frequently used methods, to conduct discriminant and modelling classification, respectively, in the case of large matrices.

s0030 ### 2.1.1 Partial Least Squares-Discriminant Analysis

p0110 PLS-DA [20,21] represents an example of the so-called discriminant classification methods. Discriminant classifiers divide the multidimensional space of the variables in as many regions as the number of given categories, that is, of classes for which there are training samples available. This approach has the direct consequence that whenever a new sample has to be classified, depending on the portion of space it falls in, it will be attributed to one and only one class. In this framework, the peculiarity of PLS-DA is that the classification model is built based upon the PLS algorithm, as the name suggests. What allows the use of a calibration method such as PLS to cope with classification issues is that the qualitative information about class-belonging of the samples can be coded into a quantitative binary-valued matrix of responses $\mathbf{Y}$.

p0115 To account for the necessity of coding the information on class-belonging, the matrix $\mathbf{Y}$ contains as many columns as the number of given categories and as many rows as the number of samples; in particular, the row corresponding to each individual will have a 1 in the position corresponding to its true class and zeros elsewhere (or will have minus one, which is equivalent). For instance, if the problem involves five classes and the jth sample belongs to class 2, then it will be described by the row vector

$$\mathbf{y}'_j = [0 \quad 1 \quad 0 \quad 0 \quad 0]. \tag{1}$$

p0120 On the other hand, if the kth sample comes from class 4, it will be coded as

$$\mathbf{y}'_k = [0 \quad 0 \quad 0 \quad 1 \quad 0]. \tag{2}$$

p0125 Accordingly, the classification problem can be re formulated as finding the best regression model linking the experimental data measured on the sample $\mathbf{X}$ to the binary-coded dummy matrix $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{XB} \tag{3}$$

p0130   **B** being the matrix of regression coefficients. As the name suggests, in the case of PLS-DA, the PLS algorithm is used to calculate the regression model in Equation (3), which makes the model applicable also to the cases where the predictor matrix is ill-conditioned (highly correlated variables and/or high variable to samples ratio, which is the common case of unfolded matrices). Indeed, the peculiarity of the PLS algorithm is that it looks for a low-dimensional representation of both the *X*- and *Y*-spaces so that the corresponding scores have the maximum covariance. Mathematically, this statement can be formulated as:

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}^T + \mathbf{E}_X \\ \mathbf{Y} &= \mathbf{UQ}^T + \mathbf{E}_Y \\ \mathbf{U} &= \mathbf{TC} \end{aligned} \tag{4}$$

where **T**, **U**, **P** and **Q** are the *X* and *Y* scores and loadings, respectively, and **C** is the matrix collecting the coefficients of the so-called inner relation, that is, the regression model relating **T** and **U**. The regression coefficient matrix **B** in Equation (3) is then calculated by combining the relations in Equation (4). This regression coefficient matrix allows prediction of the *Y* values for unknown samples $\mathbf{X}_{new}$ according to:

$$\hat{\mathbf{Y}}_{new} = \mathbf{X}_{new}\mathbf{B} \tag{5}$$

where the hat indicates predicted values. As the matrix $\hat{\mathbf{Y}}_{new}$ containing the prediction can assume real values and not only ones and zeros (or minus one, if it is codified in this way), classification of the samples is accomplished by assigning the individuals to the category corresponding to the highest value of the predicted response. For instance, in the case of a five class problem, as the one already described in Equations (1) and (2), if the predicted vector of responses for an unknown sample, $\hat{y}_{unk}$, is

$$\hat{y}_{unk} = \begin{bmatrix} -0.01 & 0.08 & 0.89 & -0.04 & 0.13 \end{bmatrix}, \tag{6}$$

then the sample is assigned to class 3.

s0035 2.1.2 Soft Independent Modelling of Class Analogies

p0135   As stated in Section 2.1.1, discriminant classification techniques provide a unique assignation of the samples to one and only one of the categories represented in the training set. As a consequence, this approach does not seem optimal for dealing with problems in which new classes are continuously emerging (as in the case of food traceability, where the number of products with a protected denomination of origin is increasing with time) or, for instance, when one is interested in a single category and the other is loosely defined as everything not belonging to that class. In all these cases, a different

approach to classification can be employed, which is commonly termed class modelling. Class-modelling techniques focus on capturing the similarities between members of the same class rather than on discriminating between individuals from different categories. Indeed, they model one class at a time so that the answer that one can get is whether a sample is accepted by that specific category or not. When more than one class is modelled, three different situations can occur: an individual can be accepted by a single class, by more than one class (confused samples) or by no category at all.

p0140   In this context, SIMCA represents probably the most widely used class-modelling technique, especially for dealing with wide and/or highly correlated data matrices [22–24]. The basic assumption of SIMCA is that the similarity between samples coming from a particular class can be captured by a principal component model of opportune dimensionality so that 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. In more detail, to build the model of a particular category, say class A, a PCA model of opportune dimensionality is computed based on the data collected on the training samples coming from that class, $\mathbf{X}_A$, according to:

$$\mathbf{X}_A = \mathbf{T}_A \mathbf{P}_A^T + \mathbf{E}_A. \tag{7}$$

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, that is, the portion of the total variability in $\mathbf{X}_A$ not accounted for by the PC model. Once the PC decomposition is computed, it is possible to define the distance of the kth sample to the model of the class, $d_{k,A}$ as:

$$d_{k,A} = \sqrt{\left(\left(\mathrm{OD}_{k,A} \right)^2 + \left(\mathrm{SD}_{k,A} \right)^2 \right)} \tag{8}$$

where $\mathrm{OD}_{k,A}$ and $\mathrm{SD}_{k,A}$ represent the orthogonal and the score distances of the individual to the model of class A, respectively. Orthogonal distance is a measure of the distance of a sample to the PC subspace and is connected to the extent of the residuals, while score distance indicates how far the individual is from the training objects of the class in the PC space, and is connected to the value of the score vector. Over the years, based on these principles, many different ways of defining the terms in Equation (8) have been proposed, and some of them have been evaluated and compared in the case of the multiway extension of the method [12] (see Section 2.3.1). In the case of the two-way implementation considered in this chapter, only the criterion borrowed from multivariate statistical process control was used. In particular, the two statistics T^2 and Q, which represent, respectively, the squared Mahalanobis distance of a sample to the centre of the score space and the sum of the squared residuals, are introduced to account for the score and orthogonal distance terms in Equation (8). Accordingly, the distance of the kth sample to the model of the class is expressed as:

$$d_{k,A} = \sqrt{\left(\left(T^2_{\mathrm{red}_{k,A}} \right)^2 + \left(Q_{\mathrm{red}_{k,A}} \right)^2 \right)} \tag{9}$$

where the subscript 'red' indicates that the variables are normalized by diving each term for the 95th percentile of the corresponding distribution under the null hypothesis. By definition, since each of the two terms in Equation (9) 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.

s0040 2.2 Multiway Compression + Classification of Scores

p0145 One approach for developing a classification method for multiway data, without recourse to unfolding of the data, is to decompose the data array by a multiway decomposition technique and then use the sample scores (or mode one loading, in a more general terminology) as a new set of variables. This new set of variables can afterwards be subjected to two-way classification methods such as LDA, SIMCA and PLS-DA. The first step, the data compression, takes full advantage of the multiway nature of data, while the second step of building classification rules is disconnected from the multiway model. Hence, it will not be possible to have a direct interpretation of the raw data in terms of their contributions to the class separation. Moreover, the multiway decomposition step considers all the categories together. Hence, preprocessing and choice of the number of components is done at this global level. Thus, even if SIMCA is used at a second stage, it cannot be considered a true class-modelling method. In fact, in the first step, that is, decomposition of the multiway array, both data pretreatments (if any) and model dimensionality are applied to the whole data.

p0150   At the second stage, application of a two-way classifier, data pretreatment such as centring and scaling can be applied, and in this case, they apply to scores, so they may serve, for example, to compensate for different amounts of variance accounted for by the different PARAFAC/Tucker factors. Then, in order to assess the optimal dimensionality of classification (PLS-DA) or class-modelling (SIMCA) models, the same tools as used in two-way classification can be used, such as classification rate in cross-validation.

p0155   To understand the role of the original sets of variables, the relevance of the derived factors has first to be evaluated. For example, suppose that scores on factors 1 and 3 show the highest discriminant capability, then going back to the decomposition model it has to be evaluated, in case of PARAFAC, which variables have high loading values on factors 1 and 3 on the respective mode. It is evident that interpretation in terms of the original variables is not straightforward.

p0160 Application of discriminant two-way classifiers to PARAFAC scores have been reported for classifying olive oils [25] and vinegar [9], while SIMCA on PARAFAC scores has been applied to environmental and geological data [26–28].

p0165 In the next two sections, the PARAFAC and Tucker3 methods are illustrated since these are the bases for the derivation of multiway classification methods.

s0045 ## 2.2.1 PARAFAC

p0170 The method was independently introduced in 1970 by Harshman [29] and Caroll and Chang [30]. It is based on the principle of 'parallel proportional profiles', that is, that a set of common factors can be used to describe simultaneously the variation occurring in several matrices albeit with different weighting coefficients for each matrix.

p0175 The PARAFAC assumes that the data array can be approximated by a low-rank trilinear structure. This means that the three-way array can be decomposed as a sum of triple outer product of vectors. The three sets of vectors are called loadings. Given a three-way array $\underline{\mathbf{X}}$ of dimension $I \times J \times K$, with elements x_{ijk}, the PARAFAC model can be expressed as follows:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{10}$$

p0180 $\mathbf{A}(I \times F)$ with element a_{if} is the first mode score (or loadings) matrix, $\mathbf{B}$ $(J \times F)$ with element b_{jf} and $\mathbf{C}(K \times F)$ with element c_{kf} are the second and the third modes loadings, respectively. F is the number of components used in the PARAFAC model; e_{ijk} is a residual term containing all the unexplained variation (the not-modelled part of the data array $\underline{\mathbf{X}}$).

p0185 The extracted components are not orthogonal in a PARAFAC model. PARAFAC allows for non-orthogonal components, which may seem like a disadvantage. However, PARAFAC also allows for so-called unique models. This means that if the data follow the PARAFAC model, PARAFAC is able to uniquely uncover the underlying components. For example, if data follow Beers law, then a PARAFAC model will be able to estimate the pure spectra even from a mixture. In, for example, a PCA model of similar two-way data, it would only be possible to find some abstract linear combination of the pure spectra.

p0190 This uniqueness property renders PARAFAC especially suited for hyphenated techniques such as HPLC-DAD, or for excitation/emission fluorescence spectroscopic data.

p0195 In general, we may see PARAFAC as recovering the profile corresponding to each unique phenomenon that is generating a variance source in the analyzed data, for example, a time profile, if, for instance, the same variables are measured for the same sample at different harvesting seasons, or in food processing if food samples are measured over maturation time, ageing and so on.

p0200   The PARAFAC model is not always appropriate. If the data set does not have a low-rank trilinear structure, then PARAFAC may provide a bad approximation or it may even be numerically impossible to fit. In such a case, the Tucker3 (see next) model may be a better choice.

p0205   To have a preliminary evaluation of how many factors could be worth considering in the subsequent classification step, the model dimensionality may be explored in terms of fit (evaluated by entity of model residuals), core consistency [13,18,19] and model stability (since best fit solution is searched through a minimization algorithm, local minima may be encountered, thus restarting the search and comparing similarity of solutions ensure a more stable model). Split-half analysis also serves the same purpose: the data set is divided into two parts in samples direction and a PARAFAC model is separately fitted on each split part. If the model is valid and the right number of components is chosen, a very similar model should be obtained with the same loadings profile in each mode as in the not-split solution.

s0050 2.2.2   Tucker3

p0210 The Tucker3 method, first introduced by Tucker in 1964 [31], can be considered an extension of principal component analysis to higher-order arrays [13]. Considering a three-way array $\underline{\mathbf{X}}$, of dimension $I \times J \times K$, it is decomposed into orthonormal triplets of loadings vectors where the number of extracted components can be different for the three modes. The decomposition is illustrated in Figure 2 and it is expressed as
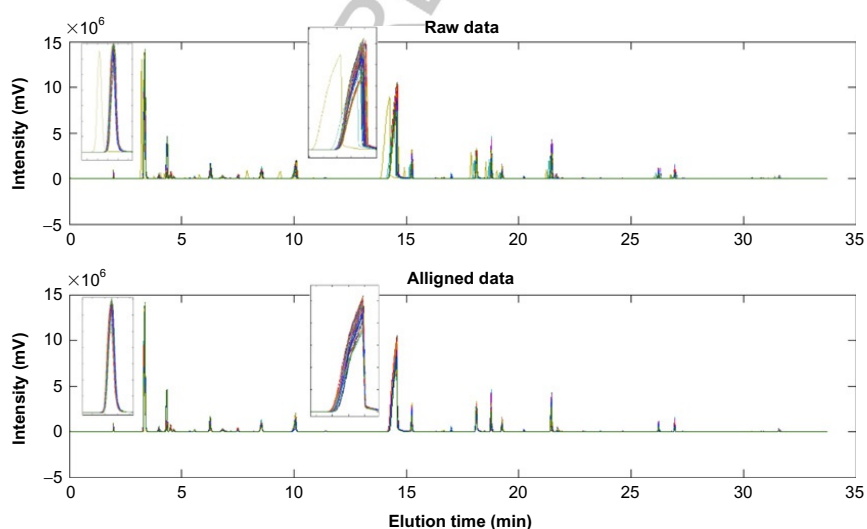
f0010   **FIGURE 2**   Tucker decomposition scheme. (For colour version of this figure, the reader is referred to the online version of this chapter.)

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

p0215   The extracted components are characterized by three sets of loadings: $\mathbf{A}(I \times P)$, for mode 1, also called scores in analogy with PCA; $\mathbf{B}(J \times Q)$ for mode 2; and $\mathbf{C}(K \times R)$ for mode 3; plus a residual term e_{ijk}. P, Q and R are the number of components extracted for each mode. Thus, the model of the original data is the weighted sum of outer products between components in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. The array $\underline{\mathbf{G}}$, of dimension $P \times Q \times R$, with elements g_{pqr}, is called the core array and represents the value by which the single component product is weighted. Therefore, the value and the sign of each core element give information about the entity of the interaction among the components of the different modes. The squared elements of the core matrix are proportional to the variation explained by the combination of the components corresponding to their indices, that is, if g_{112} is the largest core element, special attention in interpreting the model has to be given to the interaction between component 1 of mode 1, component 1 of mode 2 and component 2 of mode 3 [13,18,32,33].

p0220   The Tucker3 model does not provide a unique solution in the way PARAFAC does. It is possible to define different solutions $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{G}}$ that decompose the data array $\underline{\mathbf{X}}$ with identical fit. This is called 'rotational ambiguity' and as in PCA, to obtain an identified solution, orthogonality of loading components is imposed in the Tucker3 model. A Tucker3 model on multiway data achieves data compression and allows feature extraction and exploring data trend, much as PCA does for two-way data.

p0225   A preliminary evaluation of the model dimensionality, that is, the number of latent factors to be retained in each mode, can be done by inspection of a plot similar to the scree-plot in PCA, obtained by plotting the total number of factors, summed over each mode, against the variance explained by the model and looking to the best compromise among fit and parsimony [18].

p0230   As for a PCA model, cross-validation may also be used to assess the model complexity for both decomposition methods, PARAFAC and Tucker3, by leaving out segments of data in the array and imputing their values as missing elements, in order to estimate the variance explained in cross-validation [34].

p0235   There are no '*true*' rules for the choice of the best multiway model to use. Of course, it can be influenced by *a priori* knowledge of the structure of data set, but such knowledge is often not available. In practice, the simplest model, among those giving similar (validated) fit, has to be chosen for a given data set. PARAFAC is usually preferred because of its uniqueness in spectroscopic and calibration applications, while Tucker3 is sometimes preferred in explorative data analysis because of factor orthogonality and simpler numerical algorithms.

s0055 ## 2.3 Multiway Classification Methods

p0240 Application of truly multiway classification methods, especially in the food area, is still limited. Multiway partial least squares discriminant analysis

(NPLS-DA) has been used in food authentication [35] and very recently an approach based on dissimilarity representation has been reported [36]. In that case, the classification is accomplished on the dissimilarity matrix, that is, a square matrix reporting the degree of similarity/dissimilarity (calculated by a suitable elaboration of sample distances in the multiway domain). This approach does not allow model interpretation in terms of the set of variables/ spectral features used to characterize the samples.

p0245   However, the multiway equivalent of PLS-DA, namely NPLS-DA, and SIMCA, namely N-SIMCA, are available and are described below.

s0060 ## 2.3.1 N-SIMCA

p0250 The SIMCA method for two-way data set has been described in Section 2.1.2. However, it is worth recalling here that two main approaches exist to build class models in the SIMCA approach [12]:

o0005 **i.** original SIMCA, where the reference class variance is defined based on squared residuals for the calibration set. Then, for each projected sample, the distance to the class is defined by weighting the distance to the class model (orthogonal distance based on squared residuals) and the distance inside the class model (scores distance, which, depending on the authors, is evaluated componentwise as the distance from the class boundaries plus a confidence range, or it is the Mahalonobis distance from the centre of the PC space). Sample acceptance is evaluated by means of an F-test where the distance to that class of a projected sample (e.g. belonging to evaluation/test set or to classes different from the modelled one) is compared with the class variance. If the test is passed, the object is assigned to that class;

o0010 **ii.** alternative SIMCA (i.e. the version implemented in the PLS Toolbox [37]) where statistics, hence limits, for orthogonal distance (here referred to as Q) and scores distance (here referred to as T^2, calculated as Mahalanobis distance) are calculated on the calibration sample, independently for each category, by using two different reference statistical distributions: Hotelling-T^2 to obtain the T^2_{lim} and χ^2 to obtain the Q_{lim} [38]. This comes from the multivariate statistical process control context. The classification rule is then based on the reduced distance from the class model, as explained in Section 2.1.2, Equations (8 and 9).

p0265 Hence, a projected sample is accepted by the class model if its reduced distance is equal or less than the square root of two. Some authors [39] proposed to weight the two distances in Equation (9) differently.

p0270   Both frameworks have been extended to the multiway case to obtain a truly multiway class-modelling algorithm for multiway data arrays, namely, N-SIMCA [12].

p0275   In analogy, with the two-way SIMCA, the data of each class are modelled separately by using a multiway decomposition method, for example,

PARAFAC or TUCKER3, on the calibration set. One model is obtained for each class. Orthogonal distance (from class model) is based on squared residuals from PARAFAC or Tucker3 model and score distances (distance in model space) on model scores (mode 1 loadings). To extend the original SIMCA approach to the multiway case, the most critical step is the evaluation of the degrees of freedom to be used in the class variance estimation and in the F-test. The interested reader is referred to the original article [12]. When extending the alternative SIMCA approach, the first decision has been that of using, as corresponding to score distances, the mode one samples leverage values ($\mathbf{H} = \text{diag}[\mathbf{T}(\mathbf{TT}^{\text{T}})^{-1}\mathbf{T}^{\text{T}}]$). As a consequence, to obtain the leverage limit for the calibration set, reference statistics for leverage have been considered, implementing both suggestions formulated by Forina [40], here referred to as $H\text{lim}_{\text{fit}}$ and Pomerentsev [41], here referred to as $H\text{lim}_{\text{fit}}(AP)$.

p0280 For both original and alternative SIMCA approaches, class boundaries based on evaluation in leave-one-out cross-validation have also been implemented. In original SIMCA, this means that reference class variance has been estimated on residual values for left-out samples in the cross-validation loop (this will be referred to as original SIMCA(CV) classification criterion); in alternative SIMCA, this leads to calculation of the H and Q values for left-out samples in the cross-validation loop (H_{CV}, Q_{CV}) and their limits $H\text{lim}_{\text{CV}}$ and $Q\text{lim}_{\text{CV}}$ by using the 95‰ of the respective set of values. Moreover, the Pomerentsev statistics [41] have been evaluated by using cross-validated estimations of sample leverages and residuals.

p0285 The classification ability of N-SIMCA models is evaluated by the specificity (number of samples belonging to different classes correctly refused by the class model, expressed as percentage) and sensitivity (number of samples belonging to a class correctly accepted by their class model, expressed as percentage) in calibration (training set) and cross-validation according to the different class limits definition, that is, the different classification criteria defined above. To assess the optimal model dimensionality (number of factors for PARAFAC and number of factors in each variable mode for Tucker3), the best efficiency (geometric mean of sensitivity and specificity) in cross-validation is considered.

p0290 Once the optimal complexity, and hence the final category, models are [Au3] chosen, the results may be represented in terms of H/Q plots for each class; in the case of alternative SIMCA the indication of the respective class boundaries corresponding to the different statistics or cross-validated reference limits and in the case of original SIMCA, by means of Coomans plots [22,42] reporting the distance to a given class versus another one.

p0295 As in two-way SIMCA, pretreatments, in this case multiway centring and scaling [43], are applied separately for each class. Pretreatments, such as alignment, baseline/background correction and normalization, can be performed initially and usually on the whole data set.

p0300 As far as model interpretation is concerned, PARAFAC or Tucker3 loadings on variable modes may be inspected and interpreted. In the case of

Tucker3 models, it has to be stressed that interpretations across several models have to take into account the most relevant core array terms, which indicate the interactions among factors of different modes to be considered.

s0065 ### 2.3.2 NPLS-DA

p0305 The PLS regression extension to higher-order arrays is NPLS [13,44]. It was first developed as a PARAFAC-like model of $\underline{X}$, and it was shown that the method could be easily extended to any desired order for both $\underline{X}$ and $\underline{Y}$ matrices. This method was further elaborated and lastly improved with respect to residual analyses by introducing a core array in the model of $\underline{X}$ [45,46].

p0310 　　Considering an $\underline{X}$ array of dimension $I \times J \times K$, the NPLS model is obtained by modelling $\underline{X}$ as in Tucker3 decomposition:

$$\mathbf{X} = \mathbf{T}\mathbf{G}_X \left(\mathbf{W}^K \mathbf{W}^J \right)^{\mathbf{T}} + \underline{\mathbf{E}}_X \tag{12}$$

where $\mathbf{X}$ is the $\underline{X}$ array unfolded to an $I \times JK$ matrix, $\mathbf{T}$ holds the first mode scores (sample mode), $\mathbf{W}^J$ and $\mathbf{W}^K$ are the second and the third mode weights, respectively. The symbol $\otimes$ denotes the Kronecker product [13].

p0315 　　$\underline{G}_X$ is the core array of size $F \times F \times F$, where F is the number of NPLS components (factors) and it is defined by

$$\mathbf{G}_X = \mathbf{T} + \mathbf{X}\left(\left(\mathbf{W}^K \right)^+ \otimes \left(\mathbf{W}^J \right)^+ \right) \tag{13}$$

p0320 　　The superscript '+' means Moore-Penrose pseudo-inverse.

p0325 　　The dependent variable block in NPLS-DA is a matrix $\mathbf{Y}$ holding the class information, that is, for each category a y-variable is defined as in Equations (1) and (2) (Section 2.1.2); in this case, we used the notation one/minus one to indicate inclusion in class membership or not. The $\mathbf{Y}$ array is modelled as in the two-way case, by [Au4]

$$\mathbf{Y} = \mathbf{U}\mathbf{Q} + \mathbf{E}_Y \tag{14}$$

Where $\mathbf{U}$ holds the $\mathbf{Y}$-scores and $\mathbf{Q}$ is the loadings matrix.

p0330 　　$\underline{\mathbf{E}}_X$ and $\mathbf{E}_Y$ are $\underline{X}$ and $\mathbf{Y}$ residual arrays, respectively. In analogy with the two-way PLS algorithm, the weights are determined such that the scores obtained from the $\underline{X}$ decomposition ($\mathbf{T}$) have maximum covariance with the scores obtained from $\mathbf{Y}$ decomposition (inner relation: $\mathbf{U} = \mathbf{T}\mathbf{C}$). Regression coefficients that apply directly to $\underline{X}$ may also be derived [47,48] and used to predict the $\mathbf{Y}$ for new samples as in Equation (5) (Section 2.1.2).

p0335 　　As two-way PLS, the actual NPLS algorithm is sequential and the optimal model dimensionality can be assessed by cross-validation and classification rate in cross-validation, in the case of NPLS-DA.

p0340 　　The class assignment rule adopted here is based on Y values recalculated (calibration samples) and/or predicted (cross-validation and test set) by the model, once the number of NPLS components has been chosen. The sample

is assigned to the class for which it gets the highest value of the corresponding *y*-variable, as in Equation (9) for the two-way case (Section 2.1.2).

p0345 Assessment of the most discriminant **X** variables may be guided by inspection of NPLS weights, which, separately for each component and mode, show the most influent variables (signal features) in the NPLS model of a given *y*-variable, since each *y*-variable is related to one class. Also, regression coefficients, which summarize the variable contribution to the overall extracted components, can be represented as a landscape map for each *y* and interpreted. Moreover, very recently [49], the variable influence on projection (VIP) parameter, used in bilinear PLS-DA to assess the salient discriminant **X**-variables, has been extended to NPLS.

s0070 3 CASE STUDIES

p0350 The previously described methods are compared on two data sets with the objective of assessing the authenticity of the studied products, table wines and extra virgin olive oils (EVOOs), respectively. Both data sets consist of GC–MS data. Often, such data are reduced by summing over the *m/z* direction obtaining the total ion current, TIC, signal. This approach may be sub-optimal in cases where the information about the mass fragmentation profile may help the discrimination of the products of different origin. The applications will be illustrated to furnish guidelines for the different steps in building multiway classification models such as data pretreatments, choice of model dimensionality and validation.

s0075 3.1 Discrimination of Table Wines

s0080 3.1.1 Data Set

p0355 The data set *Wine* refers to the geographical identification of wine samples produced from the same grape (100% Cabernet Sauvignon) but harvested in different geographical areas (South America, Australia, South Africa).

p0360 The samples have been analyzed by the head space sampling technique coupled with GC–MS; for each spectrum, a scan (*m/z*: 5–204) measured at 2700 elution time-points is obtained, providing as a result a three-way array of dimensionality $70 \times 2700 \times 200$. The data set is described in Ref. [50] and it is available for 42 samples of the 72 considered by us, which include further samples collected under the same conditions by the same authors.

p0365 The data set is split into two subsets: one set is used to build the model (training set, 46 samples consisting of 20 South American, 13 Australian, 13 South American) and the other is projected in order to test the predictive ability of the models (test set, 24 samples comprising 10 South American, 7 Australian and 7 South American). The sets have been selected by using the Duplex algorithm [51], which has been implemented to obtain an effective and balanced assignment of data objects in training and test sets, evaluating

the mutual distances between pairs of points. Here, the Duplex algorithm is applied classwise with a 2:1 training/test splitting ratio on the data matrix containing the TIC chromatograms, that is, after compression of the third dimension of the data set by summing all m/z contributions for each time point to obtain the TIC.

p0370   The data set has been analyzed by unfolding approaches (unfolding the array as $\underline{\mathbf{X}}_{i,j,k} \rightarrow \mathbf{X}_{(I,JK)}$) and multiway approaches, as well as class-modelling and discriminant techniques.

s0085 ### 3.1.2 Preprocessing

p0375   Some pretreatments such as baseline, centring and misalignment removal are applied in the same way, regardless of whether data are unfolded or retained as three-way. Other approaches such as scaling require different considerations in the two cases.

p0380   The pretreatments applied to the raw data are described in detail below.

p0385   *Baseline removal*: Baseline has been corrected by using the asymmetric least square algorithm developed by Eilers [52,53]. This method has the advantage of not assigning a predefined shape to the baseline/background effect, permitting correction also for irregular baseline. Moreover, it avoids introducing negative traits. This is a common drawback when subtracting a polynomial trend from a signal.

p0390   *Misalignment correction*: Irregularity in the reproducibility of elution profiles (time direction, second mode of data array) is corrected with an algorithm for peak alignment: Interval Correlation Optimized shifting algorithm (*i*cohift) [54]. This method is based on Fast Fourier Transform calculation and allows speedy alignment of large data sets. It works on two-way data matrices, so to apply it on three-way data, the data were first converted into a matrix considering the TIC signals, that is, samples versus total ion count chromatogram. After the alignment of the TIC data (Figure 3), the displacement scheme applied for each sample is reapplied to each mass value (each point in m/z direction) [55]. In this way, alignment of the entire 2D-landscape is accomplished.

p0395   *Removal of uninformative variables*: Retention times and masses with a variance near zero are removed to focus on relevant information. The removal of
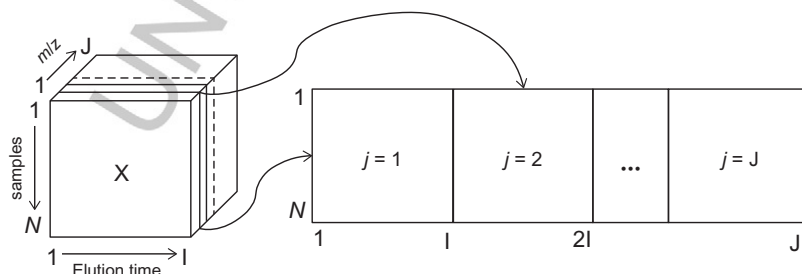
f0015   **FIGURE 3** WINE data set—Raw TIC chromatograms and aligned ones.

these times and masses is applied separately for the three-way and the unfolded data. Obviously, the deleted times and masses are not exactly the same for the two data structures, so there are some differences in the reduction of the data set for the two approaches. For the data array, the final dimension is $70 \times 684 \times 99$, while for the unfolded data, details are given in Section 3.1.4.

p0400 *Centring*: Both data sets were column mean centred in the case of unfolded data and centred across sample mode, that is, mode 1, in the case of three-way data array.

p0405 *Scaling*: For both two-way and multiway classification approaches, several scaling methods were tested, namely, weighting by the inverse of standard deviation (we will refer to this as std-scaling), pareto scaling (weighting by the inverse of square root of standard deviation) [56] and block-scaling (to equal block variance, where each block corresponds to a given signal region in the retention time and/or *m/z* directions) [57]. Moreover, comparison with unscaled data was also considered.

s0090 ### 3.1.3 Data Analysis: Unfolded Data

p0410 The simpler classification approaches, consisting in unfolding the data array, $\underline{\mathbf{X}}_{I,J,K}$, to a matrix, $\mathbf{X}_{(I,JK)}$, and then applying standard two-way classification techniques is described first.

p0415 The training and test data sets, which initially were of size $46 \times 2700 \times 200$ and $24 \times 2700 \times 200$, respectively, were unfolded row-wise by putting the chromatograms at each different *m/z* for the same sample, one after another (as illustrated in Figure 1).

p0420 Successively, as described in the previous paragraphs, the columns of the unfolded matrices corresponding to variables having zero or almost zero variance for the training samples were deleted, leading to training and test matrices of final size $46 \times 21,350$ and $24 \times 21,350$, respectively.

s0095 #### 3.1.3.1 Unfolded PLS-DA

p0425 PLS-DA has been applied to the training set and the effect of different types of preprocessing (mean centring, auto-scaling and pareto scaling) on the resulting models was evaluated and compared. In particular, models were calculated following a multi-class approach (PLS2 algorithm, as many *Y*-variables as categories, i.e. South American, Australian and South African) and assigning each sample to the class corresponding to the predicted **Y** column for which the sample gets the highest value.

p0430 The optimal complexity of the models, that is, the optimal number of latent variables, was selected by looking at the classification rate in cross-validation (where six randomly selected segments were considered).

p0435 Looking at the results reported in Table 1, it can be observed that each of the preprocessing approaches leads to very high correct classification rates in calibration, while in cross-validation, such results were not obtained, in

TABLE 1 WINE Data Set—PLS-DA Classification After Unfolding

Preprocessing	LV	Class 1 (South America)			Class 2 (Australia)			Class 3 (South Africa)		
		Cal	CV	Pred	Cal	CV	Pred	Cal	CV	Pred
Mean	7	95	85	90	100	69	43	92	69	86
Auto	7	100	95	80	100	100	86	100	69	86
Pareto	9	100	80	90	100	85	71	100	69	100

Correct classification rates for the three discriminated categories in calibration (Cal), cross-validation (CV) and external validation (Pred) as a function of the different preprocessing. LV indicates the number of PLS components.

particular, for the third class (South Africa). Based on these outcomes, the optimal preprocessing seems to be auto-scaling, which results in correct classification rates higher than 80% for the external test set samples and for all the three classes.

s0100 ### 3.1.3.2 Unfolded SIMCA

p0440 SIMCA analysis has also been carried out on the unfolded data set, using as classification rule the reduced distance Equation (9), described in Section 2.1.3. Analogous to what was already done for PLS-DA, the effect of the different preprocessing on the final results was evaluated and compared. The optimal complexity of each category model, that is, the number of principal components to be retained for the definition of the class space, was chosen on the basis of a sixfold cross-validation procedure by taking the dimensionality corresponding to the maximum value of the geometric average of sensitivity and specificity, named efficiency [40,42]. Once the individual category models were built, it was possible to compute the sensitivity (percentage of correctly recognized samples belonging to the modelled class) and specificity (percentage of correctly rejected samples belonging to different classes) values in validation by projecting on the class-space the test samples from the same class and all the other samples from the other two classes, respectively. The results of SIMCA modelling on the WINE data set after unfolding are summarized in Table 2.

p0445 Because in SIMCA, as in all class-modelling techniques, each category is modelled independently, in principle, it is possible to choose a different preprocessing for each class. If this is done, and the criterion for optimality is defined as the maximum efficiency for the category, then, by observing the results in Table 2, according to CV results, it is possible to see that the best data pretreatments seem to be mean centring for the first class (South American) and pareto scaling for the other two (Australian and South African). However, if a single preprocessing, resulting in good performances for all classes, is sought, auto-scaling seems to be the best choice. In Figure 4, the unfolded SIMCA classification results are shown for the three classes by plotting the reduced distances Hotelling-T2 (distance inside class model) versus Q (distance from the class model), that is, the respective distances divided by the class limits; thus, projected samples are recognized as belonging to the modelled class if they fall below the square root of two semicircles.

s0105 ### 3.1.3.3 PLS-DA and SIMCA Classification on Tucker3 Scores

p0450 As discussed in the theoretical section, another way of performing classification on multiway arrays is to use standard two-way classifiers on the scores obtained by applying multiway decomposition techniques to the data arrays. This approach involves two stages: at first, multi-linear techniques such as PARAFAC or TUCKER3 are used to model the three-way array.
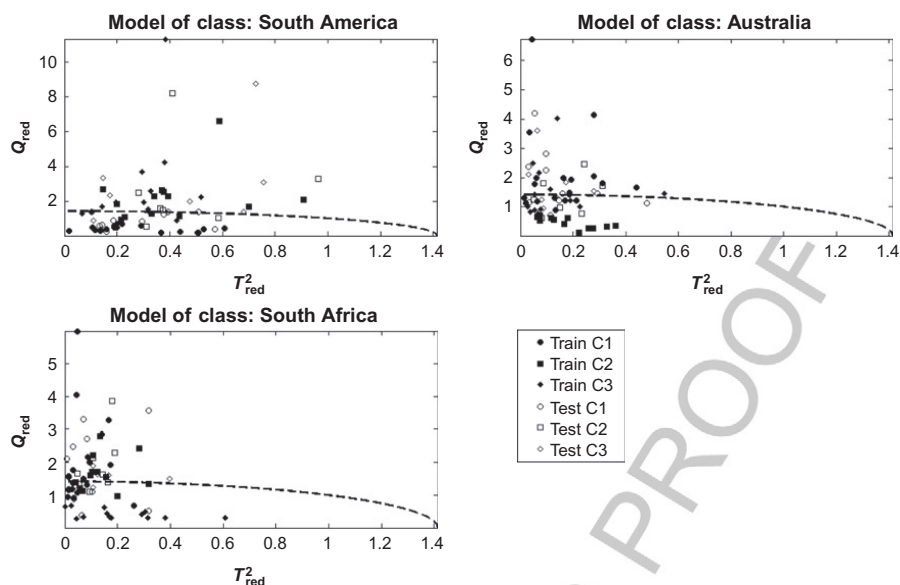
TABLE 2 WINE Data Set—SIMCA Modelling on Unfolded Matrices

		Class 1 (South American)				Class 2 (Australian)				Class 3 (South African)			
		LV	*Cal*	*CV*	*Pred*	*LV*	*Cal*	*CV*	*Pred*	*LV*	*Cal*	*CV*	*Pred*
Mean	SENS	5	100	75	80	4	100	77	43	3	100	46	72
	SPEC		69	74	79		48	53	47		30	47	47
Auto	SENS	3	100	75	90	3	100	77	71	3	100	54	43
	SPEC		50	58	51		48	51	65		73	80	88
Pareto	SENS	4	100	75	100	4	100	77	57	3	100	69	43
	SPEC		65	66	79		54	62	59		51	64	65

Sensitivity and specificity of the individual class models in calibration (Cal), cross-validation (CV) and external validation (Pred), as a function of the preprocessing adopted. LV indicates the number of PCA components for each class model.

t0010

f0020 **FIGURE 4** WINE data set—reduced T^2 versus reduced Q plot of SIMCA models, reported in Table 2 (unfolded data). Top left: South American (model on mean centred data); Top right: Australian (model on paretoscaled data) and South African (model on paretoscaled data). Dotted semicircle corresponds to the class boundary, that is, square root of two, according to alternative SIMCA approach.

Subsequently, the loadings of the sample mode (scores), which are arranged in a two-way matrix, are used as new variables and processed by standard classification techniques.

p0455   Particularly for the WINE data set, Tucker3 decomposition was used to extract scores. The data array was pretreated as described in Section 3.1.3, and as far as centring and scaling are concerned, the array was centred across mode 1 and pareto scaling was applied for mode 2 and mode 3; this choice is explained in Section 3.1.5. Models with dimensionality, that is, number of factors on each mode, ranging from [1 1 1] to [10 10 10], were explored, and considering the best balance between explained variance and model dimensionality, a model corresponding to eight factors in each mode was chosen and the corresponding scores for the eight factors of the first mode (samples mode) taken as input for the standard classification tool, that is, discriminant (PLS-DA) and modelling (SIMCA) techniques.

p0460   The results obtained by applying PLS-DA on the Tucker3 scores are reported in Table 3. Analogous to what was already described in Section 3.1.3.1, in this case also the choice of the optimal number of latent variables in the PLS models was made on the basis of a cross-validation procedure with six segments. By looking at the values of the classification rates reported in Table 3, it is possible to observe that the best results were obtained by using mean centring as preprocessing with respect to the outcomes of applying PLS-DA on the unfolded

t0015

TABLE 3 WINE Data Set—PLS-DA Classification on Tucker3 Scores									
		Class 1 (South America)			Class 2 (Australia)			Class 3 (South Africa)		
	LV	Cal	CV	Pred	Cal	CV	Pred	Cal	CV	Pred
Mean	5	95	95	80	61	46	71	92	61	86
Auto	2	95	90	80	61	46	57	85	61	86
Pareto	3	95	85	80	54	46	57	85	69	86

Correct classification rates for the three discriminated categories in calibration (Cal), cross-validation(CV) and external validation (Pred) as a function of the different preprocessing. LV indicates the number of PLS components.

matrices, while on the South African samples, the predictive ability remained the same, and validation samples from Australia were better predicted. On the other hand, the results on South American samples are slightly worse.

p0465   Then, on the same data set resulting from Tucker decomposition, SIMCA was also applied. In particular, class models were built on differently pretreated category data, and the optimal complexity was chosen, based on the results of a sixfold cross-validation procedure.

p0470   As can be seen from the values in Table 4, analogous to what was already observed in Section 3.1.3.2, pareto scaling provides a better compromise between sensitivity and specificity for the second and the third classes (Australian and South African), while for the first one, auto-scaling seems to be better. Anyway, as the outcomes of applying pareto scaling on the first class are not significantly worse than those obtained using auto-scaling, it could be used for all classes if a unique data preprocessing is desired. Comparing the results in Table 3 with those reported in Table 4, it is possible to affirm that the SIMCA models built on the Tucker scores appear in general to have higher sensitivity and specificity values than the corresponding ones built on the unfolded data matrices.

s0110   ### 3.1.4   N-SIMCA

p0475   Dealing with multiway data, it must be underlined that the scaling pretreatment can be performed in each variable mode (two modes in this case) and there can be differences in the results depending on the order in which each pretreatment is applied [13,43]. To choose the best pretreatment, an explorative data analysis step has been carried out in order to evaluate the possible pretreatments and order of application. In this case, as we are dealing with GC–MS and strict trilinearity cannot be assumed, Tucker3 has been used as decomposition technique. Thus, the different pretreatments were studied through screening Tucker3 models by inspecting them in mode 1 scores plot

TABLE 4 WINE Data Set—SIMCA Modelling on Tucker3 Scores

|		Class 1 (South American)				Class 2 (Australian)				Class 3 (South African)			
		PC	Cal	CV	Pred	PC	Cal	CV	Pred	PC	Cal	CV	Pred
Mean	SENS	1	85	70	100	2	69	38	57	1	92	38	71
	SPEC		39	69	29		58	79	71		48	67	59
Auto	SENS	6	100	85	100	5	100	92	100	5	100	54	100
	SPEC		58	81	57		30	42	35		24	79	41
Pareto	SENS	4	100	75	100	4	100	77	86	3	100	61	57
	SPEC		69	78	64		42	60	47		58	77	65

Sensitivity and specificity of the individual class models in calibration (Cal), cross-validation (CV) and external validation (Pred), as a function of the preprocessing adopted. LV indicates the number of PCA components for each class model.

t0020

(samples mode), which combination gives better separation of the classes. The centring and scaling combinations tested, considering as well the order in which the pretreatment is performed, were the following:

u0005   mean centring, std-scaling, std-scaling for mode 1, 2 and 3, respectively. Both order: 3 2 1 and 2 3 1 were considered;

u0010   mean centring, pareto scaling, pareto scaling; order: 3 2 1 and 2 3 1;

u0015   mean centring, no pretreatment, pareto scaling; order: 3 2 1 and 2 3 1;

u0020   mean centring, pareto scaling, no pretreatment; order: 3 2 1 and 2 3 1.

p0500   Among all the scaling procedures, by visual inspection of the plot of mode 1 scores, the pretreatment, mean centring (mode 1), pareto scaling (mode 2) and pareto scaling (mode 3) in the order 2 3 1, resulted in the best visual appearance with respect to class separation. In Figure 5, the effect of data pretreatment is shown for the landscape of one of the samples, while the mode 1 scores plot obtained by the Tucker3 screening model is shown in Figure 6 (left plot). It can be seen that while all the three classes overlap, class 1 has lower dispersion with respect to class 3 and 2, which has the highest. For comparison, in the right part of Figure 6, the scores plot for the first two PCs of PCA of the unfolded data is also shown. The class separation is even less clear, indicating that in destroying the three-way structure of data some information is actually missing.

p0505   Classes are indicated by different symbols as shown in the legend. Empty symbols are for test set projected samples.

p0510   In order to reduce the number of N-SIMCA models to calculate, a preliminary assessment of the Tucker3 model dimensionality to be explored has been evaluated separately on the training set of each class. As an example, the results for the Australian category are reported in Figure 7. The plot shows
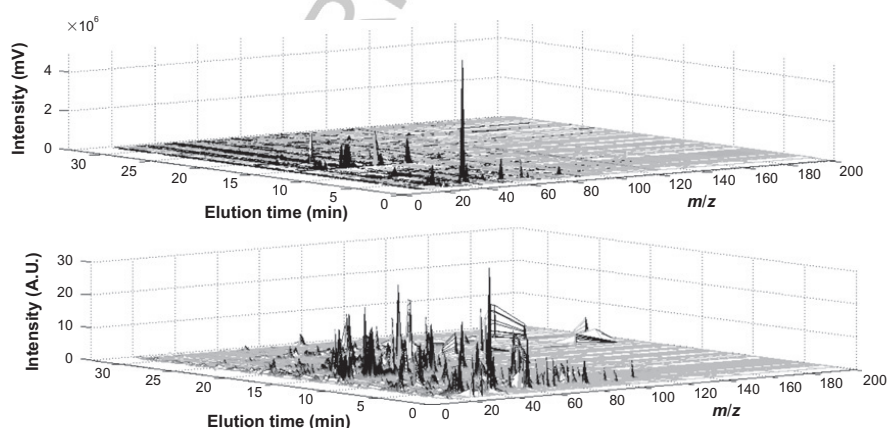
f0025   **FIGURE 5**   WINE data set—As an example, the GC–MS data for sample one of the class South American is shown before (top figure) and after scaling (bottom figure), pareto scaling within modes 2 and 3.
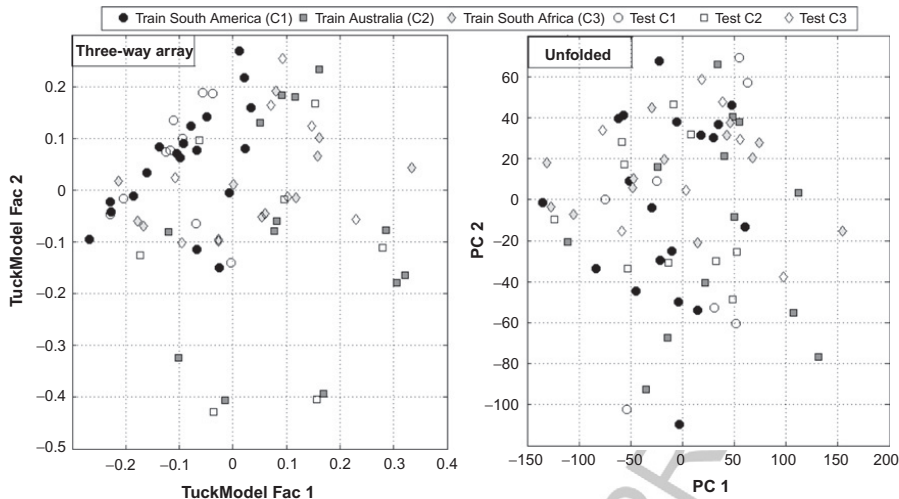
f0030   **FIGURE 6**   WINE data set—Left plot: Factor 1 versus Factor 2 scores plot of the preliminary Tucker3 model calculated on the whole calibration set. Test set is also projected on the model; Right plot: PC1 versus PC2 scores plot of the PCA model calculated on unfolded data matrix.
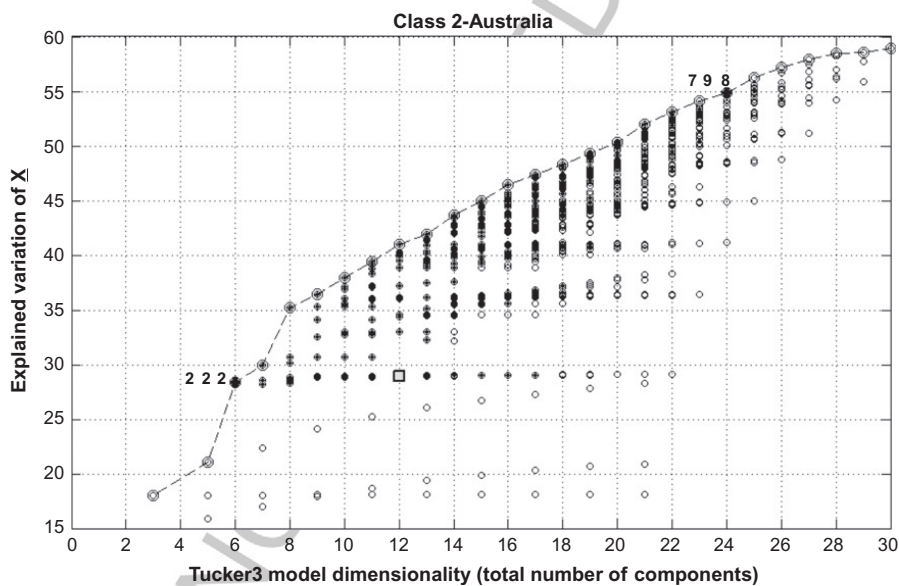
f0035   **FIGURE 7**   WINE data set—Explained $\underline{X}$ variance as function of Tucker3 models dimensionality, for the class Australian. The explored models in N-SIMCA are indicated by black filled circles and are all feasible combinations from [2 3 2] to [7 9 8]. The best performing model, according to efficiency criterion for this class, was the grey square, corresponding to [6 2 4].

the explained model variance versus the sum of the number of factors in each mode. The combinations corresponding to the highest explained variance, for the same sum, are labelled and connected by a dotted line. Considering the variance trend and by inspecting the analogous plots for the other two

categories, factors combinations from [3 2 2] to [6 6 6] were explored for South American and from [2 3 2] to [7 9 8] (corresponding to filled circles in Figure 7) for the other two classes.

p0515   In assessing what an appropriate N-SIMCA model is, local models are calculated for the data of each class. Once these models are calculated, the best factor combinations for each class (i.e. the N-SIMCA model dimensionality) are chosen considering the best efficiency (geometric mean of sensitivity and specificity for each class) in cross-validation. As an example, Figure 8 shows the values obtained for class 2 (Australian). The different classification criteria, described in detail in Section 3.2.1, are identified as curves of different colours. For each criterion, the dimensionality corresponding to the highest efficiency is shown in the figure. Sensitivity and specificity for the N-SIMCA models corresponding to these combinations of factors were evaluated, and the best performing model corresponds to class reference limits estimated in fit for both leverage and residuals ($H\lim_{\text{fit}}$, $Q\lim_{\text{fit}}$) and to the combination of factors [6 3 2] in each of the three modes, respectively.

p0520   When evaluating the model corresponding to the highest CV-efficiency in the original SIMCA approach, it behaves with no sensitivity at all for the test set samples.

p0525   Table 5 reports, for each class, the values of sensitivity and specificity in fit (SENS_Cal, SPEC_Cal), CV (SENS_CV, SPEC_CV) and prediction (SENS_Pred, SPEC_Pred) obtained for the chosen criteria and best
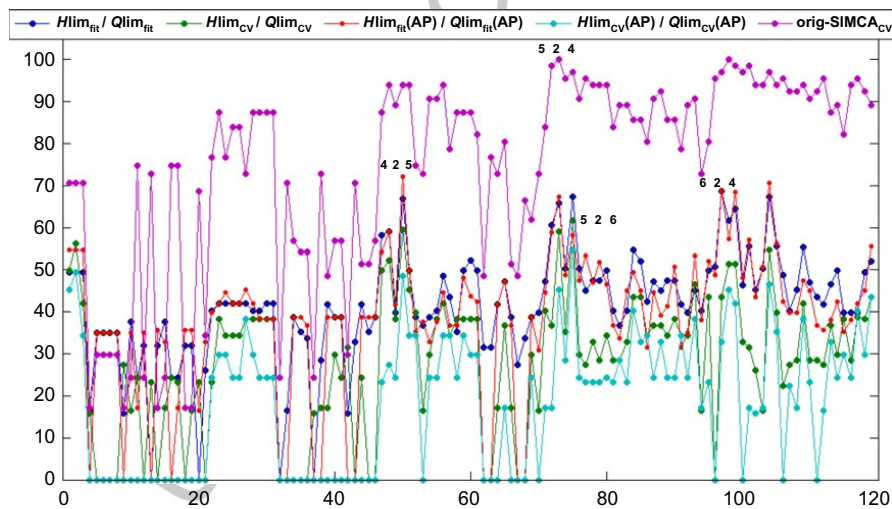
f0040   **FIGURE 8**   WINE data set—N-SIMCA models for class 2 (Australian). Values of the efficiency (estimated in CV) in function of the number of combinations of factors tested (for clarity not all explored combinations are shown). Different lines and symbols refer to different classification criteria, as reported in the legend. The best combinations for each criterion are shown. In particular, the [6 2 4] model emerged as the best one. (For colour version of this figure, the reader is referred to the online version of this chapter.)

t0025

TABLE 5 WINE Data Set—N-SIMCA

$H\text{lim}_{fit}/Q\text{lim}_{fit}$	Class 1 (South America) [6 3 2]	Class 2 (Australia) [6 2 4]	Class 3 (South Africa) [5 2 4]
SENS_Cal	95	100	100
SPEC_Cal	88	76	70
SENS_CV	65	62	69
SPEC_CV	88	76	70
SENS_Pred	60	86	71
SPEC_Pred	79	71	65

Sensitivity and specificity in calibration (cal), cross-validation (CV) and for the external test set (pred) for each category. The model was chosen according to best efficiency in cross-validation and the best performing classification criterion was that for which class assignation was accomplished on the basis of $H\text{lim}_{fit}$ and $Q\text{lim}_{fit}$ boundaries. The model dimensionality is reported in brackets for each class.

combinations of factors. These were selected according to higher efficiency, estimated in 'leave-one-out' cross-validation for fit-based classification rules, being the CV-based classification criteria already calculated on projection of samples taken out in cross-validation run.

p0530 As it may be expected from the sample distribution and class overlap observed by explorative Tucker3 model (Figure 5), class 1, showing the lowest dispersion, has the highest specificity value.

p0535 In Figure 9, *H* versus *Q* plots are shown for each class (training, CV and test samples) with the class acceptance limits calculated according to the different classification criteria in the alternative SIMCA framework. For all classes, criteria in fit lead to best performing N-SIMCA models, if a best compromise between sensitivity and specificity is sought. As the classes are rather heterogeneous, cross-validated samples (grey-filled symbols) are rather spread; thus class limits in CV (dashed line in Figure 9) are larger and as a consequence specificity decreases. There is little difference when limits are calculated according to Pomarentsev criteria (grey lines in Figure 9).

p0540 It is possible to see which the chromatographic and mass spectral regions contributing to the category models are by looking at the loadings plots of mode 2 and 3, respectively. For each class model, exploration of the core array showed 1 1 1 as the most relevant term , so in Figure 10, the plot of the first factor for mode 2 and 3 for each class is reported. In particular, in the top figure, the mode 2 (Retention time) factor one loadings for the three classes are superimposed. Some Retention time regions are relevant for all classes but some are specific for specific classes. The mode 3 (masses) factor one loadings plot for each class is shown separately and also here some mass
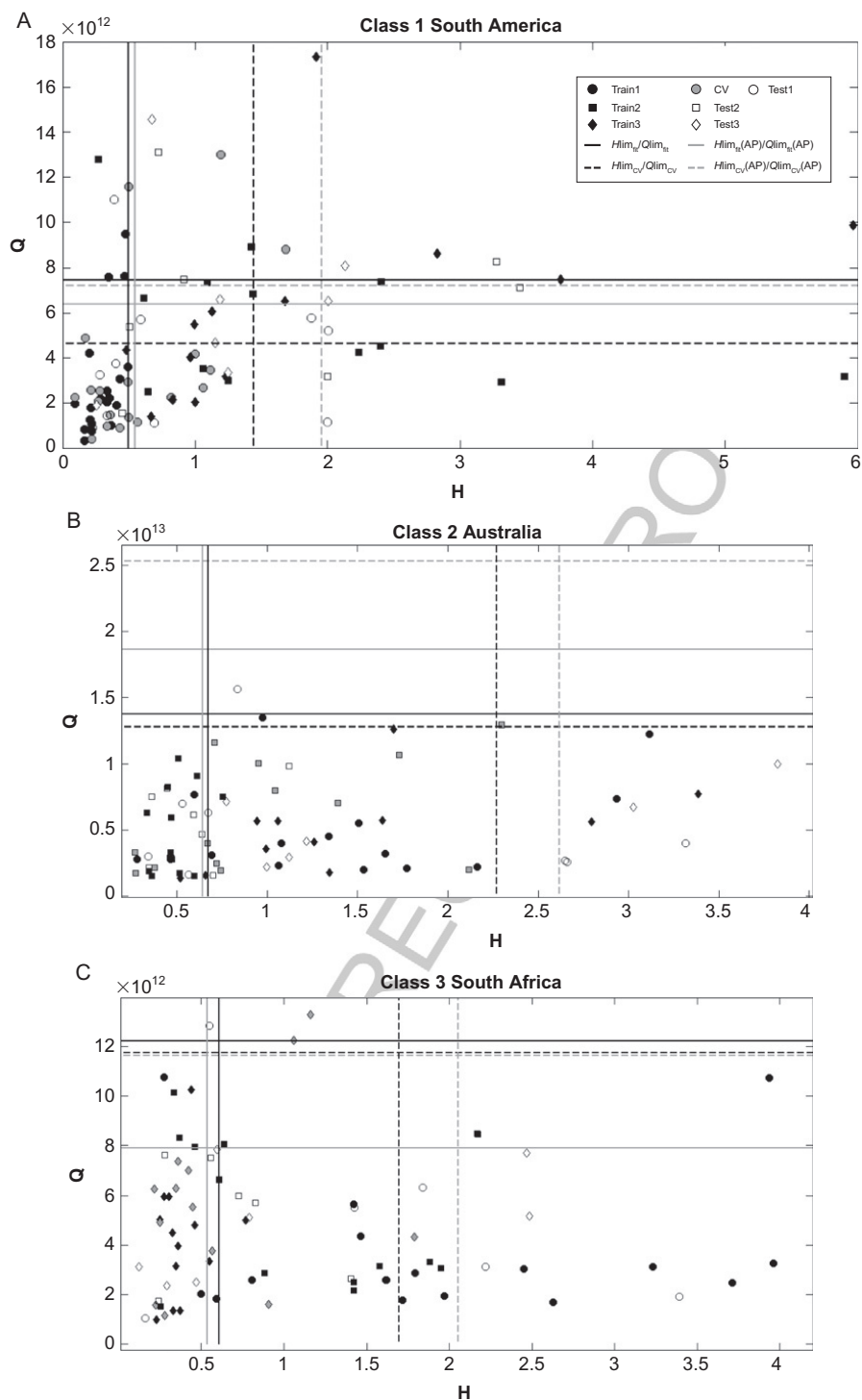
f0045 **FIGURE 9** WINE data set—N-SIMCA models. *H* versus *Q* plots with the class boundaries corresponding to the different classification criteria shown with horizontal and vertical lines as indicated in the legend. (A) Class1: South American; (B) Class 2: Australian; (C) Class 3: South African. The samples belonging to different classes are shown by different symbols: class 1, circles; class 2 squares; class 3, diamonds. Solid black is for calibration, solid grey for cross-validated and empty for the test set samples.
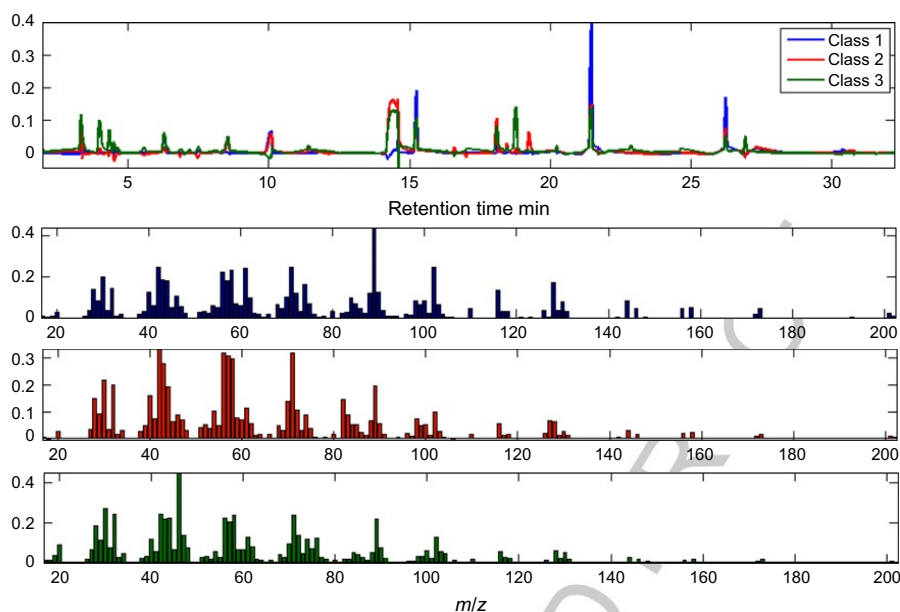
f0050 **FIGURE 10** WINE data set—N-SIMCA models. First row plot: second mode, Factor 1 loadings superimposed for the three classes; second to fourth rows plot: third mode Factor 1 loadings for each of three classes in the order: class 1, class 2 and class 3. (For colour version of this figure, the reader is referred to the online version of this chapter.)

pattern is common and some are specific. A detailed discussion in terms of chemical components corresponding to these spectral features is beyond our present aims, but it shows that loadings can be interpreted as in two-way data analysis.

s0115 ### 3.1.5 NPLS-DA

p0545 NPLS-DA has been performed using the same data preprocessing described in Section 3.1.5. A five latent variables (LV) model has been chosen considering the minimum values of the classification error in cross-validation (using six-fold venetian-blind cross-validation). The objects are assigned to the class for which the corresponding value of the dummy *y*-variable, as predicted by the model is higher. Table 6 reports the classification rates for training set (Cal), in cross-validation (CV) and for the test set (Pred) for the three classes.

p0550 In Figure 11, the predicted dummy *y*-variable versus samples for each class are plotted with different lines. Results are satisfactory despite lower specificity, considering the starting class overlaps.

s0120 ### 3.1.6 Final Remarks

p0555 The values obtained for the different approaches reflect the initial situation observed in the screening analysis of data reported in the Tucker3 scores plot

t0030

TABLE 6 WINE Data Set—NPLS-DA

NPLS-DA LV = 5	Class 1 (South American)	Class 2 (Australian)	Class 3 (South African)
Cal	100	100	92
CV	65	54	67
Pred	90	86	86

Correct classification rates for the three discriminated categories in calibration (Cal), cross-validation (CV) and external validation (Pred). LV indicates the number of NPLS components.
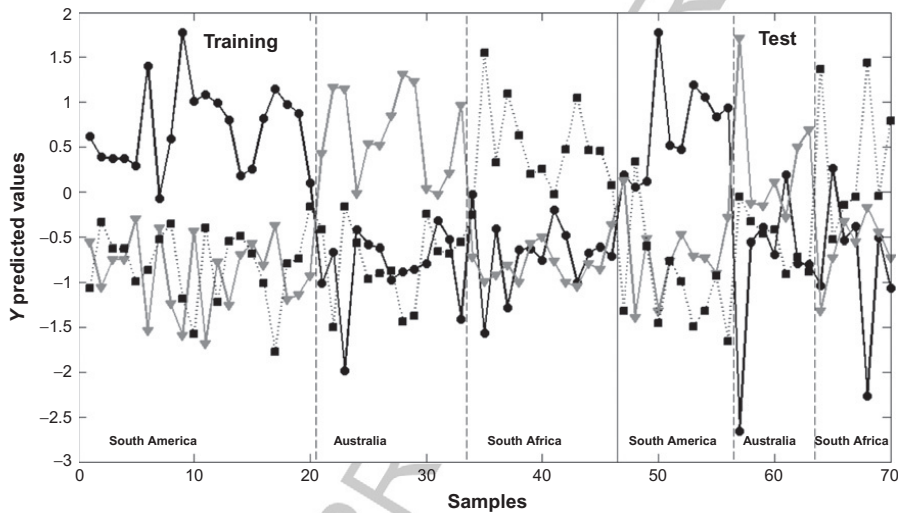
f0055   **FIGURE 11** WINE data set—NPLS-DA results. The *Y* values recalculated (training set) and predicted (test set) by the NPLS-DA model are shown. The South American class corresponds to the black line with solid circles, Australian to grey line with triangles and South African to dotted line with squares.

of Section 3.1.5 (Figure 6). The three classes appear overlapping and quite disperse in the Factor 1 versus Factor 2 score space. The situation is almost the same even if score plots for the subsequent factors of the screening model are considered (results not shown). Table 7 summarizes the results obtained for both discriminant and class-modelling methods with the different approaches: unfolding, tucker scores followed by two-way classification and multiway classification. For the discriminant methods (Unfolded PLS-DA, Tucker scores + PLS-DA and NPLS-DA), the percent of correct classification in calibration and prediction is reported for each class. The performance of the methods is comparable except for the Tucker scores + PLS-DA approach, which shows lower values in the discrimination of class 2, Australian (C2).

t0035

TABLE 7 WINE Data Set—Comparison of the Best Performing Model for Each Classification Approach

Method	South American		Australian		South African	
	Cal	*Pred*	*Cal*	*Pred*	*Cal*	*Pred*
Unfold PLS-DA	Auto-scaling; LV=7					
	100	80	100	86	100	86
Tucker scores+PLS-DA	Mean centre; LV=5					
	95	80	61	71	92	86
NPLS-DA	[Mean Centre; Pareto; Pareto] LV=5					
	100	90	100	86	92	86
Unfold SIMCA	Auto; PC=3		Auto; PC=3		Auto; PC=3	
SENS	100	90	100	71	100	43
SPEC	50	51	48	65	73	88
Tucker scores+SIMCA	Pareto; PC=4		Pareto; PC=4		Pareto; PC=3	
SENS	100	100	100	86	100	57
SPEC	69	64	42	47	58	65
N-SIMCA	[mnc; psc; psc] Factors=[6 3 2]		[mnc; psc; psc] Factors=[6 2 4]		[mnc; psc; psc] Factors=[5 2 4]	
SENS	95	60	100	86	100	71
SPEC	88	79	76	71	70	65

The NPLS-DA method is more parsimonious, five LV instead of seven, with respect to unfolded PLS-DA, and gives better prediction results for the first class, that is, the most homogenous one. Moreover, the chromatographic and m/z regions contributing most to the NPLS-DA model, hence more discriminant, are easily interpretable in terms of VIP. These are reported in Figure 12, one for each *y*-variable that in this case correspond to a given class. As far as chromatographic regions are concerned (Figure 12A), some peaks which seem characteristic only for the first and the second class are highlighted; this could be discussed together with VIP on the masses mode (Figure 12B) that shows relevant masses for each category. Also NPLS regression coefficients landscape may be obtained and interpreted. As an example, in Figure 12C the regression coefficient maps for the second and third classes are reported. The important VIP regions at about 5 and 18 min (elution time) are now seen as showing also different and more relevant
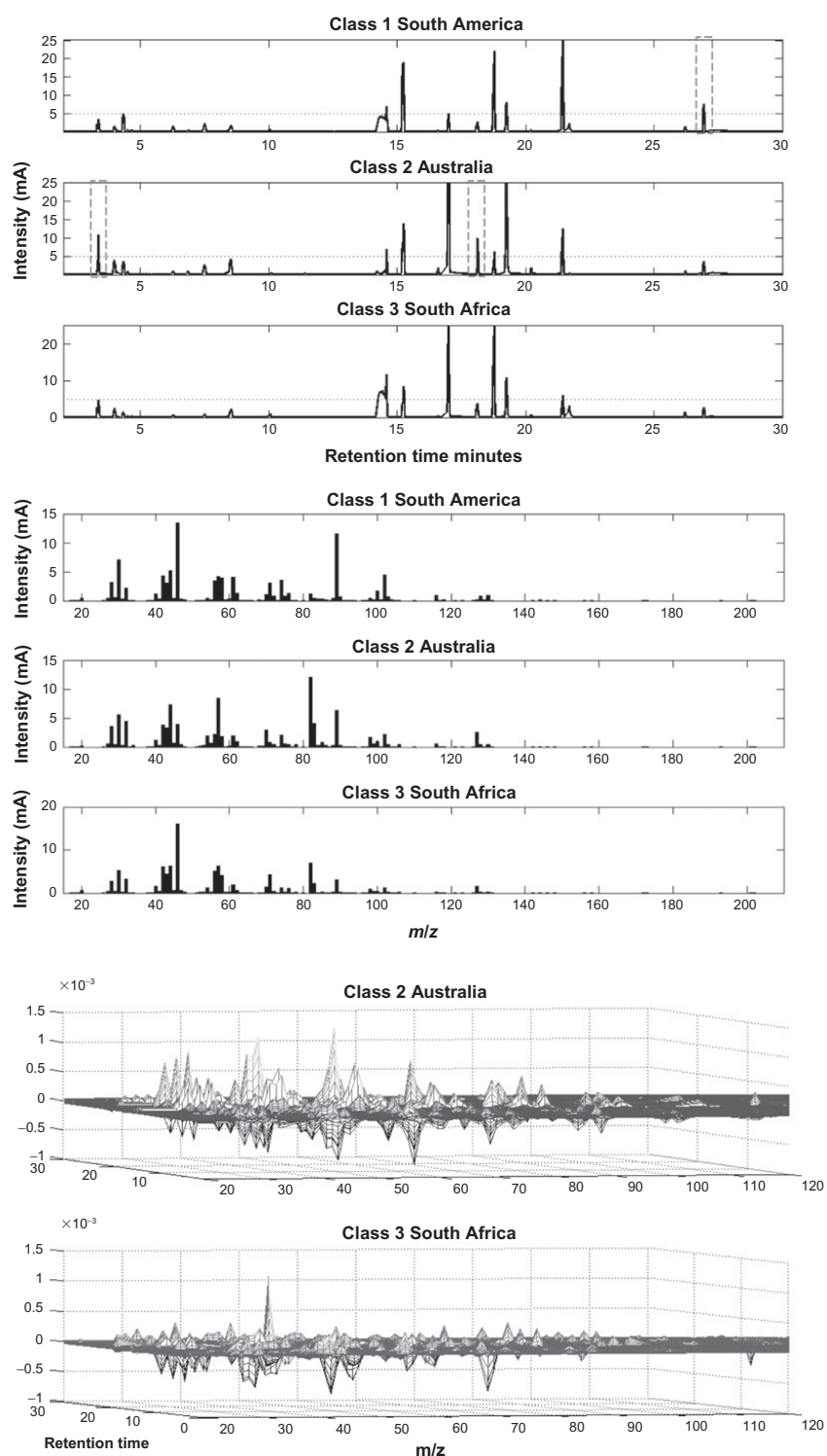
f0060 **FIGURE 12** WINE data set—NPLS-DA results. VIP plot for the three classes, for (A) mode 2 elution profiles and (B) mode 3 spectra profile *m/z*. (C) 2D map of NPLS-DA regression coefficients for class 2 and 3.

regression coefficient values with respect to class. This will be extremely difficult in the unfolding case given the very high number of variables in the second dimension.

p0560 Regarding class modelling, that is, SIMCA method, sensitivity and specificity (not distinguished by single classes but considered altogether) values are reported in the bottom part of Table 9. It is evident that the class-modelling approach, in this case, suffers from excessive dispersion of the three classes when compared to a discriminant one that resulted in this case performing more. The SIMCA model, on unfolded data, presents poor values of specificity for all the classes, unless, as in the case of the third class (South African), specificity is higher but at the expense of sensitivity. The Tucker3 scores+SIMCA model also does not assure specificity, performs very badly on class 3 and seems slightly worse than the unfolding approach.

p0565 Models calculated with N-SIMCA with respect to the SIMCA models previously discussed for unfolded data and Tucker3 scores present improved values of specificity for all classes for both calibration and prediction sets, equal sensitivity for the class 2 and improved sensitivity for class 3 in prediction; only in the case of class 1, test set sensitivity is lower. Moreover, N-SIMCA models always show a better balance between sensitivity and specificity.

p0570 In conclusion, multiway models even when performing similarly in classification, as in the case of discriminant approach, offer a better interpretability of the results, and in the case of class modelling perform better, highlighting that taking into account the multiway structure allows improvement of class characterization. To decompose the data set by a multiway method, such as Tucker3, and then applying classification tools in general, furnishes worse results than unfolding; this may be explained for the PLS-DA model by the fact that at the stage of selecting the dimensionality of the Tucker3 model classification capability is not taken into account in an analogous manner as when principal component regression is compared with PLS for bilinear data; when unfolding, the multiway data structure is distorted, but overall data set variation is better retained since all variables in both second and third mode are still used.

s0125 3.2 Discrimination of EVOO

s0130 3.2.1 Data Set

p0575 This data set concerns the classification of EVOO. The samples of six different olive cultivars were considered for the study, and the aim was to model the class Liguria Taggiasca against the others, since this variety has been designed by PDO certification and represents one of the most estimated EVOO [58]. The other samples belong to other cultivars of other different Mediterranean areas (Apulia in Italy, Greece, Spain and Tunis). The EVOO samples have been analyzed [12] by Head Space Solid Phase Extraction coupled with Gas Chromatography Mass Spectrometry (i.e. HS-SPME-/GC–MS)

in order to characterize the samples according to the volatile fraction that is the most characteristic for the evaluation of aroma.

p0580　The chromatograms were acquired for a total time of about 67 min, but the first 3 min and the last 10 min in the retention mode were cut because there were no peaks at all. The acquired mass range was 35–250 a.m.u. Thus, the data constitute a three-way array: 73 (samples) $\times$ 1514 (Retention times) $\times$ 216 (m/z). In Ref. [12], a preliminary selection of the masses to be considered was done, while here the whole range of masses has been considered. After the elimination of variables with almost null variance, both for the chromatographic and mass spectrum directions, the final three-way array was of dimensionality $73 \times 1039 \times 161$. As already pointed out for WINE data set, the deleted times and masses are not exactly the same for the unfolded data for which the details are given in Section 3.2.3.

p0585　As explained in the case of the Wine data set, the Duplex algorithm has been used on TIC signals to split the samples into training and test sets.

p0590　The analysis of this data follows the same guidelines as adopted with the WINE data set by comparing the different classification methods on both two-way unfolded and three-way array data sets.

p0595　Unfolding is accomplished row-wise in this case too, and then PLS-DA and SIMCA are applied on the unfolded matrix.

p0600　The Tucker3 scores used for the subsequent application of PLS-DA and SIMCA analyses are taken from a [10 10 10] Tucker3 model. This model was selected according to the best compromise among explained variance and number of factors, as evaluated by looking at explained variance as a function of model complexity plot, where all feasible component combinations from [1 1 1] to [12 12 12] were explored. The choice of the optimal complexity of the classification models for both PLS-DA and SIMCA was made, according to cross-validation, as described in Sections 3.1.4 and 3.1.5 for the WINE data set. ┤Au6├

s0135　3.2.2　Preprocessing

p0605　For this data set also, baseline was corrected by asymmetric least squares, and chromatograms were aligned, as described in Section 3.1.3. As far as scaling is concerned, with this data set also, different kinds of scaling were evaluated and will be reported in the respective section for unfolded and three-way arrangements.

s0140　3.2.3　Unfolded Data & Classification on Tucker3 Scores

p0610　In order to use the standard two-way classification and class-modelling methods, the data array has been rearranged by row-wise unfolding. Then, all the columns corresponding to variables having zero or almost zero variance on the training samples were removed from the unfolded data matrices to be used

for model building and validation, whose final size results were $42 \times 49{,}644$ and $31 \times 49{,}644$, respectively.

p0615   At first, classification was accomplished through the use of PLS-DA model dimensionality choice according to minimum classification error in cross-validation. As it was done for the WINE data set, three different preprocessing strategies (mean centring, auto-scaling and pareto scaling) were evaluated and compared. The results of PLS-DA modelling are summarized in Table 8.

p0620   Inspection of the results shows that a perfect discrimination between the two classes can be obtained, irrespective of the selected pretreatment. However, when considering the model complexity, auto-scaling results in the highest parsimony, as only two latent variables are necessary. PLS-DA on Tucker3 scores also provides very good classification rates (Table 8, bottom part), irrespective of the pretreatment chosen. All the models appear to be rather parsimonious as only one latent variable is needed.

p0625   However, considering that in general, dealing with the authenticity issue, the focus is on a single category, as in this case where the aim is to have a model to assess ligurian EVOO, adopting a class-modelling approach may be more appropriate since it is not always necessary for a different category to be modelled. Thus, a class-modelling approach by means of the SIMCA algorithm is also performed.

p0630   The same classification rule used in the case of the WINE data set, that is, the combination of T^2 and Q statistics to define a reduced class distance, is

t0040 **TABLE 8** EVOO Data Set—PLS-DA Classification After Unfolding and PLS-DA Classification on Tucker3 Scores

		Liguria		
	LV	*Cal*	*CV*	*Pred*
Unfold PLS-DA				
Mean centring	4	100	100	100
Auto-scaling	2	100	100	100
Paretoscaling	4	100	100	100
Tucker scores + PLS-DA				
Mean centring	1	100	92	100
Auto-scaling	1	100	92	100
Pareto scaling	1	100	92	100

Correct classification rates for the three discriminated categories in calibration (Cal), cross-validation (CV) and external validation (Pred) as a function of the different preprocessing. LV indicates the number of PLS components.

used for the definition of the class space. The optimal complexity of the models is selected as the one leading to the maximum efficiency in cross-validation. The results are reported in Table 9.

p0635   In analyzing the results in Table 9, it has to be stressed that, differently than in the case of WINE, which is a true multi-class situation, here the problem is asymmetric as the second class is actually a non-class, including all the samples coming from origin different from Liguria. Thus, only the SIMCA model built for the category Liguria has a real interest, and it is presented. All the pretreatments lead to very similar results, with perfect specificity but poor sensitivity for the validation set.

p0640   When using SIMCA on Tucker scores, the best results are obtained using mean centring. This model is the only one showing at the same time acceptable sensitivity and very good specificity. Indeed, the other two pretreatments lead to a slightly better sensitivity but a significantly poorer specificity. With respect to the results obtained on the unfolded data set, while sensitivity decreased for the calibration set, training and test set performance were similar.

s0145   ### 3.2.4   Multiway Classification: N-SIMCA & NPLS-DA

p0645   Considering the non-trilinearity of the chromatographic data, Tucker3 decomposition is preferred also for the decomposition of this data set. The data

t0045   **TABLE 9** EVOO Data Set—SIMCA Modelling on Unfolded Matrices and Sensitivity and Specificity of the Individual Class Models in Calibration (Cal), Cross-Validation (CV) and External Validation (Pred), as a Function of the Preprocessing Adopted

| | | Liguria | | | | | |
| | | Cal | | CV | | Pred | |
	LV	SENS	SPEC	SENS	SPEC	SENS	SPEC
Unfold SIMCA							
Mean	2	100	100	92	100	56	95
Auto	1	100	100	54	100	56	100
Pareto	1	92	100	85	100	56	100
Tucker scores + SIMCA							
Mean	1	77	96	62	93	81	91
Auto	3	100	59	92	78	100	27
Pareto	2	100	59	92	38	100	2

LV indicates the number of PCA components for each class model.

analysis strategy was the same as described for the WINE data set. First several pretreatments were evaluated on the training set by explorative Tucker3 decomposition, looking at the sample scores. Most promising results, in term of less overlap of the Liguria with the other classes, are obtained by pretreating the data set with mean centring in the first mode, block-scaling in the second mode and std-scaling in the third mode, with the order 3 2 1. Then, a preliminary inspection of model dimensionality was done by looking at the variance explained by several combinations of components: Tucker3 models have been fit varying the number of factors from [1 1 1] to [10 20 20] for the three modes, respectively. From inspection of explained variance versus total number of factors, we focused on the factor range [7 8 8]–[10 9 9] and run N-SIMCA with these settings. Finally, considering the efficiency of the model in cross-validation, the best performing classification criteria were relative to alternative SIMCA framework for both fit and CV criteria; for both criteria, the combination leading to higher efficiency in CV corresponds to factors [9 9 8]. In Table 10 are reported the values of sensitivity and specificity for the Liguria class comparing the criteria in fit and cross-validation for the classification rule based on Pomerantsev limits definition for H and Q statistics, $H\lim_{\text{fit}}(AP)$, $Q\lim_{\text{fit}}(AP)$, best performance is obtained by fit criteria, as also shown in Figure 13.

p0650   The NPLS-DA model dimensionality has been chosen according to minimum classification error in fivefold venetian blind cross-validation, corresponding to three LV. Results obtained for the training and test set are reported in Table 11 and Figure 14, which show the values of Y predicted for each class. The model shows perfect sensitivity and specificity as well in prediction. The most influent signal regions are highlighted in the VIP plot,

t0050   **TABLE 10** EVOO Data Set—N-SIMCA Modelling on the Three-Way Matrix

Liguria	$H_{\text{fit}}/Q_{\text{fit}}$ AP [9 9 8]	H_CV/Q_CV AP [9 9 8]
SENS_Cal	92	100
SPEC_Cal	100	96
SENS_CV	92	92
SPEC_CV	96	97
SENS_Pred	100	100
SPEC_Pred	100	77

Sensitivity and specificity in calibration (cal), cross-validation (CV) and for the external test set (pred) for the class liguria. The model was chosen according to best efficiency in cross-validation. The best performing classification criteria are reported, that is, those for which class assignation was accomplished on the basis of $H\lim_{\text{fit}}$ (AP) and $Q\lim_{\text{fit}}$ (AP) boundaries, both in calibration and cross-validation. The model dimensionality is reported in brackets for each class.
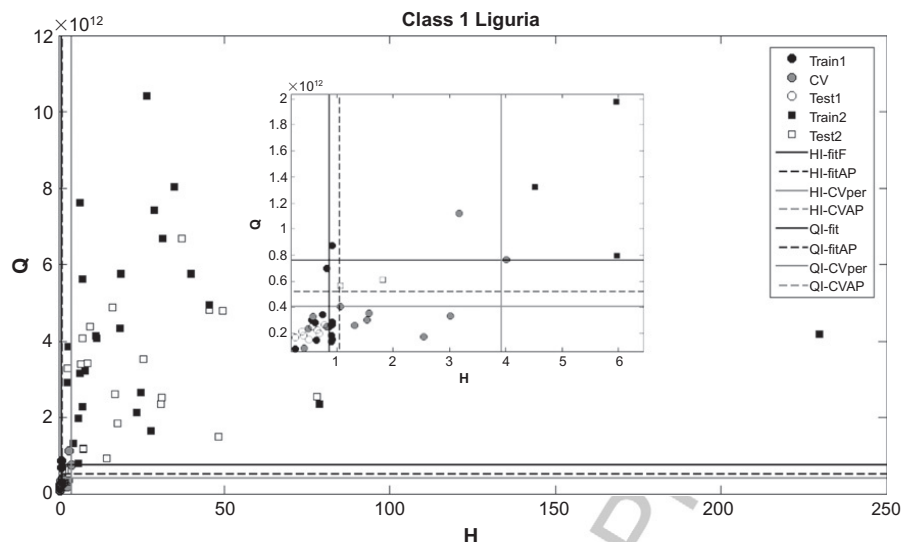
f0065 **FIGURE 13** EVOO data set—N-SIMCA results. *H* versus *Q* plot for class Liguria, Tucker3 model with combination factors [9 9 8].

t0055

TABLE 11 EVOO Data Set—NPLS-DA

EVOO	Liguria LV 7
Cal	100
CV	100
Pred	100

Rate of correct classification for calibration (Cal), cross-validation (CV) and for the external test set (Pred). LV indicates the number of NPLS-DA components.

Figure 15. As an example, the relevant contribution of retention time of about 46 min together with mass patterns 53, 68, 79, 93, 107, 121 can be attributed to limonene; by interpreting jointly the VIP plots in the two mode some information can be gathered on the volatile pattern of the ligurian EVOO.

s0150 *3.2.5 Final Remarks*

p0655 Comparing all the results for this data set, it is clear that in both cases, PLS- [Au7] DA on unfolded data and PLS-DA on Tucker3 scores, optimal discrimination of Ligurian EVOO is obtained. It has to be noticed that PLS-DA model on Tucker scores uses only one LV—in fact 100% of Sensitivity and Specificity both for training and test is reached. The same optimal results are obtained with NPLS-DA. Class modelling, using SIMCA, furnishes a model with no
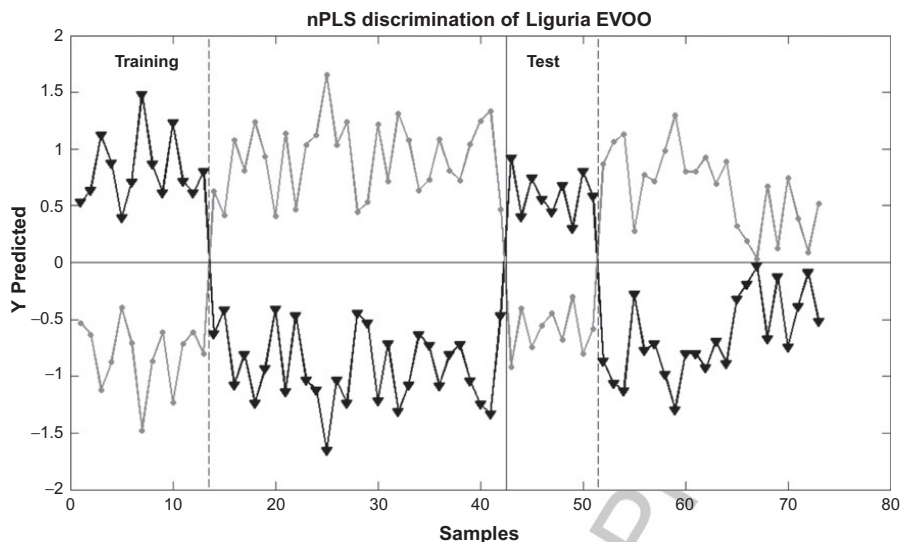
f0070 **FIGURE 14** EVOO data set—NPLS-DA results. Values of predicted *Y* for the training and test sets. Black line with triangles refers to Liguria class and grey line with points, to non-ligurian samples.
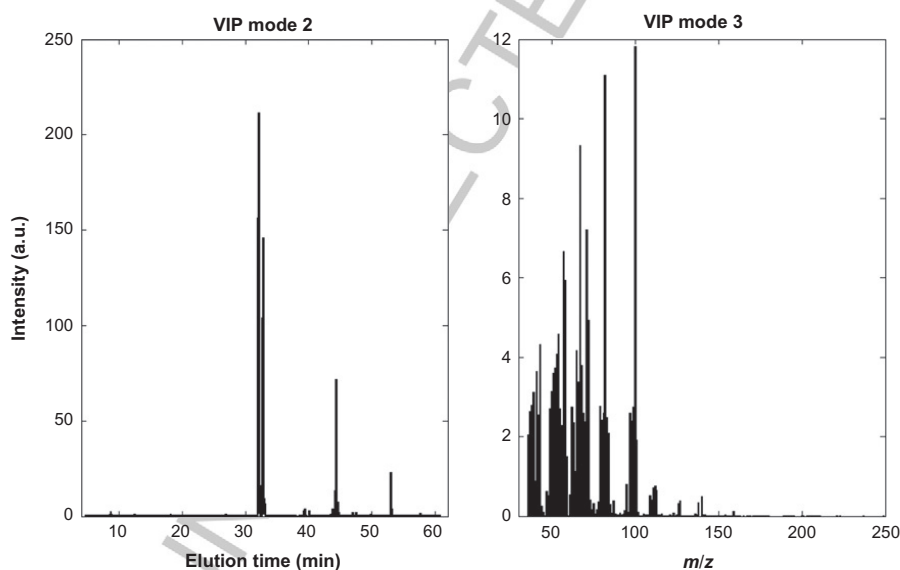
f0075 **FIGURE 15** EVOO data set—NPLS-DA. VIPs plots for second and third modes.

sensitivity for the test set when applied to unfolded data. When applied on Tucker3 scores, sensitivity increases but remains unsatisfactory. Optimal results are obtained by using N-SIMCA. This highlights how beneficial it is to take into account the multiway nature of the data array extracting the relevant information to model the class feature without distorting, compressing or

losing information. In authentication tasks, this asymmetric one class situation is very common, and the possibility to use a class-modelling approach is very relevant and it has become feasible with N-SIMCA.

s0155 4 CONCLUSIONS

p0660 The time and methods are now mature to take full advantage of the multiway structure of data when dealing with discrimination and class-modelling problems. In particular, we have illustrated some guidelines starting from data preprocessing, going through explorative multiway data analysis, that is, Tucker3 (in case of non-strictly trilinear data) or PARAFAC (data following trilinearity) to evaluate the feasibility of the classification tasks, the most suitable preprocessing and the range of model dimensionality (number of factors or combination of factors) to consider in the classification task. Then, NPLS-DA and N-SIMCA have been illustrated and compared, with particular focus on the choice of model dimensionality based on classification ability (either as minimal classification error or as higher efficiency) in internal validation and on model interpretability by inspection of loadings plots, regression coefficients map and VIP values.

p0665 This latter aspect, even when performance of models, obtained by bilinear classifiers on unfolded/compressed data, could be similar, points in favour of multiway classification methods.

REFERENCES

[1] EC regulations 2081/92 and 1898/06.

[2] TRACE: "Tracing Food Commodities in Europe", No. FP6-2003-FOOD-2-A 006942 (2005–2009).

[3] Lees M, editor. Food authenticity and traceability. Cambridge: Woodhead Publishing; 2003.

[4] Sun DW, editor. Modern techniques for food authentication. London, UK: Academic Press/Elsevier; 2008.

[5] Cocchi M, Bro R, Durante C, Manzini D, Marchetti A, Saccani F, et al. Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. Food Qual Prefer 2006;17:419.

[6] Munck L, Nørgaard L, Engelsen SB, Bro R, Andersson CA. Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. Chemom Intell Lab Syst 1998;44:31.

[7] Cocchi M, Durante C, Grandi M, Manzini D, Marchetti A. Three-way principal component analysis of the volatile fraction by HS-SPME/GC of aceto balsamico tradizionale of modena. Talanta 2008;74:547.

[8] Pereira AC, Reis MS, Saraiva PM, Marques JC. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD. Chemom Intell Lab Syst 2011;105:43.

[9] Callejón RM, Amigo JM, Pairo E, Garmónd S, Ocana JA, Morales ML. Classification of Sherry vinegars by combining multidimensional fluorescence, PARAFAC and different classification approaches. Talanta 2012;88:456.

[10] Christensen J, Nørgaard L, Bro R, Engelsen SB. Multivariate autofluorescence of intact food systems. Chem Rev 2006;106:1979.

[11] Andersen CM, Mortensen G. Fluorescence spectroscopy: a rapid tool for analyzing dairy products. J Agric Food Chem 2008;56:720.

[12] Durante C, Bro R, Cocchi M. A classification tool for N-way array based on SIMCA methodology. Chemom Intell Lab Syst 2011;106:73.

[13] Smilde A, Bro R, Geladi P. Multi-way analysis with applications. In: Multiway analysis in the chemical sciences. United Kingdom: Wiley; 2004. p. 35–45. Chapter 3.1.

[14] Castro C, Manetti C. A multiway approach to analyze metabonomic data: a study of maize seeds development. Anal Bioanal Chem 2007;371:194.

[15] Durante C, Cocchi M, Grandi M, Marchetti A, Bro R. Application of N-PLS to gas chromatographic and sensory of traditional balsamic vinegars of modena. Chemom Intell Lab Syst 2006;83:54.

[16] Leardi R, Armanino C, Lanteri S, Albertonanza L. Three-way principal component analysis for monitoring data from Venice lagoon. J Chemom 2000;14:197.

[17] Yilmaz A, Nyberg NT, Jaroszewski JW. Metabolic profiling based on two-dimensional J-resolved ^{1}H NMR data and parallel factor analysis. Anal Chem 2011;83:8278.

[18] Kroonenberg PM. Applied multiway data analysis. New York: Wiley; 2008.

[19] Andersson CA, Bro R. The N-way toolbox for MATLAB. Chemom Intell Lab Syst 2000;52:1.

[20] Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 2001;58:109.

[21] Barker M, Rayens W. Partial least squares for discrimination. J Chemom 2003;17:166.

[22] Wold S. Pattern recognition by means of disjoint principal components models. Pattern Recogn 1976;8:127.

[23] Wold S, Sjostrom M. SIMCA: a method for analyzing chemical data in terms of similarity and analogy. In: Kowalski B, editor. Chemometrics: theory and application. ACS symposium series, vol. 52. Washington, DC: American Chemical Society; 1977. p. 243–82.

[24] Maesschalck RD, Caldolfi A, Massart DL, Heuerding S. Decision criteria for soft independent modelling of class analogy applied to NIR. Chemom Intell Lab Syst 1999;47:65.

[25] Guimet F, Boque R, Ferre J. Application of non-negative matrix factorization combined with Fisher's linear discriminant analysis for classification of olive oil excitation–emission fluorescence spectra. Chemom Intell Lab Syst 2006;81:94.

[26] Hall GJ, Kenny JE. Estuarine water classification using EEM spectroscopy and PARAFAC-SIMCA. Anal Chim Acta 2007;581:118.

[27] Khosravi A, Melendez J, Colomer J. Classification of sags gathered in distribution substations based on multiway principal component analysis. Electr Power Syst Res 2009;79:144.

[28] Renard N, Bourennane S. Dimensionality reduction based on tensor modeling for classification methods. IEEE Trans Geosci Electron 2009;47:1123.

[29] Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis. UCLA Working Papers in Phonetics 1970;16:1. [Au8]

[30] Caroll JB, Chang J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart–Young' decomposition. Psychometrika 1970;35:283.

[31] Tucker LR. The extension of factor analysis to three-dimensional matrices. In: Frederiksen N, Gulliksen H, editors. Contributions to mathematical psychology. New York: Holt, Rinehart & Winston; 1964. p. 110–82. [Au9]

[32] Henrion R. Body diagonalization of core matrices in three-way principal component analysis: theoretical bounds and simulation. J Chemom 1993;7:477.

[33] Henrion R, Andersson CA. A new criterion for simple-structure transformation of core arrays in N-way principal component analysis. Chemom Intell Lab Syst 1999;47:189.

[34] Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem 2008;390:241.

[35] Guimet F, Ferré J, Boqué R. Rapid detection of olive-pomace oil adulteration in extra virgin olive oils from the protected denomination of origin Siuriana using excitation-emission fluorescence spectroscopy and three-way methods of analysis. Anal Chim Acta 2005;544:143.

[36] Porro-Munoz D, Duin RPW, Talavera I, Orozco-Alzate M. Classification of three-way data by the dissimilarity representation. Signal Process 2011;91:2520.

[37] Wise B, Gallagher N, Bro R, Shaver J, Windig W, Koch R. PLS_Toolbox manual 4.0. Wenatchee, USA: Eigenvector Research, Inc.; 2009.

[38] Wise BM, Gallagher NB, Watts S, Butler D, White Jr D, Barna G. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. J Chemom 1999;13:379.

[39] Branden KV, Hubert M. Robust classification in high dimensions based on the SIMCA method. Chemom Intell Lab Syst 2005;79:10.

[40] Forina M, Casale M, Oliveri P, Lanteri S. CAIMAN brothers: a family of powerful classification and class modeling techniques. Chemom Intell Lab Syst 2009;96:239.

[41] Pomerantsev AL. Acceptance areas for multivariate classification derived by projection methods. J Chemom 2008;22:601.

[42] Forina M, Oliveri P, Lanteri S, Casale M. Class-modeling techniques, classic and new, for old and new problems. Chemom Intell Lab Syst 2008;93:132.

[43] Bro R, Smilde AK. Centering and scaling in component analysis. J Chemom 2003;17:16.

[44] Bro R. Multi-way calibration. Multi-linear PLS. J Chemom 1996;10:47.

[45] Smilde AK. Comments on multilinear PLS. J Chemom 1997;11:367.

[46] De Jong S. Short communication regression coefficients in multilinear PLS. J Chemom 1998;12:77.

[47] Nilsson J, Jong S, Smilde AK. Multiway calibration in 3D QSAR. J Chemom 1997;11:511.

[48] De Jong S. Regression coefficients in multilinear PLS. J Chemom 1998;12:77.

[49] S. Favilla, M. Cocchi. Assessing features relevance in NPLS models by VIP. Chemom Intell Lab Syst 2012; submitted. [Au10]

[50] Skov T, Balabio D, Bro R. Multiblock variance partitioning. A new approach for comparing variation in multiple data blocks, Anal Chim Acta 2008;615:18.http://www.models.kvl.dk/Wine_GCMS_FTIR.

[51] Snee RD. Validation and regression models: methods and examples. Technometrics 1977;19:415.

[52] Eilers PHC. Parametric time warping. Anal Chem 2004;76:404.

[53] Boelens HFM, Dijkstra RJ, Eilers PHC, Fitzpatrick F, Westerhuis JA. New background correction method for liquid chromatography with diode array detection, infrared spectroscopy detection and Raman spectroscopy detection. J Chromatogr A 2004;1057:21.

[54] Tomasi G, Savorani F, Engelsen SB. icoshift: an effective tool for the alignment of chromatographic data. J Chromatogr A 2011;1218:7832.

[55] Savorani F, Tommasi G, Engelsen SB. icoshift: a versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson 2010;202:190.

[56] Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. Scaling. In: Umetrics, editor, Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS). 1999, p. 213–225.

[57] Wold S, Johansson E, Cocchi M. PLS: partial least squares projections to latent structures. In: Kubinyi H, editor. 3D QSAR in drug design: theory, methods and applications. Leiden: ESCOM Science Publishers; 1993. ISBN 90-72199-14-6. p. 523–50.

[58] Zunin P, Boggia R, Lanteri S, Leardi R, De Andreis R, Evangelisti P. J Chromatogr A 2004;1023:271.

Non-Print Items

Abstract

Food chain traceability, identification of adulteration and the control of labelling compliance are areas that require evaluation of foodstuff in its entirety. More and more researchers are investigating the possibility of using multidimensional or hyphenated techniques for fingerprinting of food products. However, these techniques produce data structures that are multidimensional as well and that require proper chemometric approaches for data processing (multiway data analysis).

In this chapter, the state-of-the-art approaches for the classification of multiway data is discussed theoretically and compared with the case studies coming from the food authenticity context, such as the traceability of extra virgin olive oils of protected denomination of origin and table wines.

Keywords: Multiway classification; Traceability; Food authentication; N-PLSDA; N-SIMCA