2. METHODS

2.1 Background: Two-way methods

2.1.1 Data exploration

In this section Principal Component Analysis (PCA), which is the most widespread and known exploration method in multivariate analysis, is briefly described. Some concepts and aspects are clarified and explained, these information will be of fundamental importance also for understanding other techniques described later in the text.

Considering the most generalized situation for a two-dimensional data set, this can be represented by a data matrix, denoted by **X**; the *N* rows in the table are termed 'objects', these, in our case, corresponds to chemical samples, but can be every kind of experiment (e.g. patients for medical data); the *K* columns are termed 'variables' and comprise the measurements made on the 'objects'.

PCA (Principal Component Analysis) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [1-3]. The number of principal components at maximum can be equal to the number of original variables (if the data matrix is of full rank, in that case PCA is an invariant orthogonal transformation) but in general is lower, and the aim is to retain only few significant components (in this case data reduction is accomplished and noise is filtered out). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it should be orthogonal to (i.e., uncorrelated with) the preceding components.

Main aims of PCA is that of reducing the dimensionality of the data without sacrificing (too much) the accuracy and maintaining as much as possible the variation contained in the data set, since the information of many

independent variables is summarized in a smaller set of derived variables, known with different names: latent variables, principal components or eigenvectors. These new variables are calculated as a linear combination of the original variables [4,5]. Expressing the equations in terms of matrices yields to:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}' + \mathbf{E} \; [\text{Eq. } 1].$$

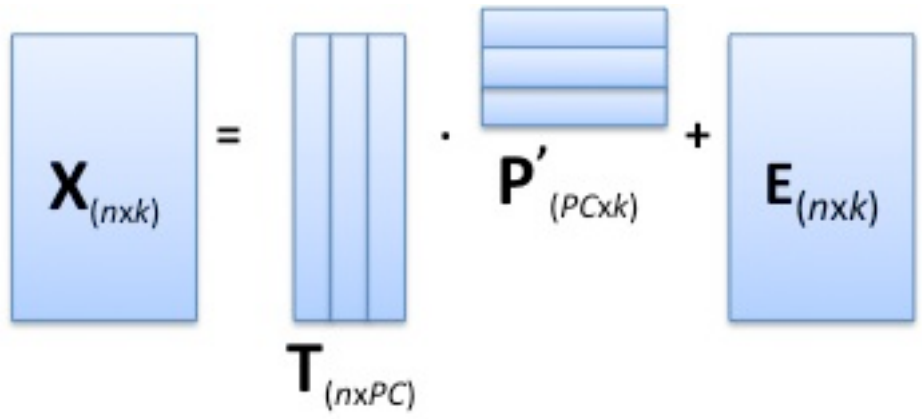

Figure 1: Principal component decomposition of a matrix.

This means that the original data are decomposed in two matrices: **T** which is called scores matrix, and holds the new latent variables, i.e. the coordinates of data points (objects) in the PCA space; and **P** which is called the loadings matrix, and contains the weights with which each original variables contributes to the linear combination (a_{nj}).

The plot of the columns of the matrix **T** gives a picture of the dominants 'object patterns' of the **X** and analogously, plotting the rows of **P'** shows the complementary 'variables patterns'; data can be more easily interpreted with a plot. In Equation 1, **E** is the residuals matrix; these are the difference between the original data matrix **X** and the reconstructed one ($\mathbf{T} \cdot \mathbf{P}'$)**.**

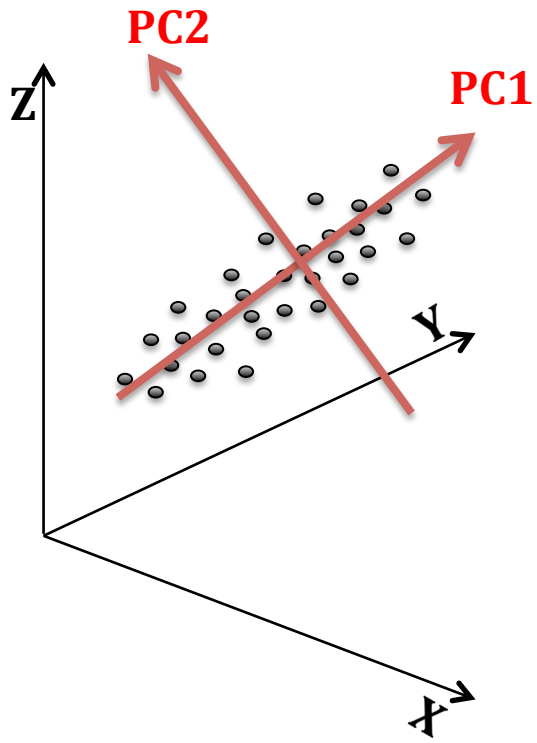

Figure 2: Geometric interpretation of principal component analysis.

Generally the aim of PCA is to capture the maximum information in the simplest model data set; hence the choice of the principal components number is oriented in the selection of the minimum number of components necessary to explain all the variability of the data. From a geometrical point of view a data matrix of dimension *N* (objects) x *K* (variables) corresponds to *N* objects, projected in a *K*-dimensional space, clearly is impossible to visualize a space with dimensionality higher than three, but from a mathematical point of view there is not difference and the geometrical concepts like points, lines, planes, angles and distances have the same properties in the *K*-dimensional space and the 3-dimensional space. When calculating the principal components the same *N* objects will be projected in an *A*-dimensional space (where *A* is the number of PCs), this space will clearly have a lower dimensionality, because most of the variance of the data will be capture from the first principal components. The main advantage

from an explorative point of view, is that data can now be visualized with few, two or three-dimensional, scatter plots.

The PCA calculation is a maximization problem (retaining maximum data variance) under constraints (orthogonality and normalization of scores and loadings vectors) and can be solved by eigenvectors/eigenvalues decomposition [6]. This can be done using different algorithms, such as Eigenvalue Decomposition (EVD), Power method, singular values decomposition (SVD) and NIPALS [7,8]. In chemometrics software usually are implemented SVD and/or NIPALS, and graphical interfaces are provided.

PCA is sensitive to the relative scaling of the original variables, but the different pretreatments, which can be applied to data, will be discussed in Section 2.4.

The minimum number of latent variables to be selected can be chosen considering the variance explained by each principal component (scree plot [9]). Total explained variance increases with the number of components, but if the increment between two consecutive components is not significant, information is not modeled, but only noise; another or complementary solution can be that of analyzing the residuals of the model; these should be unstructured when information is exhausted. Moreover cross-validation approaches can be used for the selection of the correct number of components for the system, but this argument is affronted in a more specific manner, further in Section 2.5.1.

Once the number of components is set a PCA model is obtained. Geometrically this model is an hyper plane of A-dimensions; two distances can be defined for each object (sample) the distance from the hyper plane (from the PC model), which is expressed by the sum of squared residuals for that objects and is also called Q, and the distance from the origin of the hyper plane (distance within the model space), which is expressed by the sum of squares of scores and is called T^2 (it is the Mahalonobis distance if scores are scaled to unit variance, which is usually done) [10].

For a j^{th} object with row value x_j, Q_i is calculated by:

$$Q_i = \mathbf{e_j} - \mathbf{e_j^T} \text{ [Eq. 2]}.$$

This value represents a measure of the variation of the data outside the PCA model and it is the Euclidean distance of the data point from the plane formed by the PCs model.

T^2 is the sum of the normalized squared scores, known as Hotelling's T^2 statistic [11]. It is defined as:

$$T_j^2 = t_j \boldsymbol{\lambda}^{-1} t_j^T \text{ [Eq. 3]}.$$

where t_j is the j^{th} row of the $\mathbf{T_A}$ matrix of scores vectors from the PCA model and $\boldsymbol{\lambda}^{-1}$ is the diagonal matrix containing the inverse of the eigenvalues associated with the A eingenvectors (principal components) retained in the model. T^2 is a measure of the distance from the multivariate mean, i.e. the distance between the projection of the object in the PCs space and the center or origin of the PCs model. In other words, T^2 gives information about the variation inside the class model.

T^2 and Q and associate statistics, namely Hotelling – T^2 for T^2 and c–statistics for Q, can be used to depict possible outliers and extreme behaving samples. Moreover, if the modeled objects are similar, i.e. belong to the same population, e.g. are of the same category, as in classification tasks, statistics limits at a given confidence level (usually the 95%) for T^2 and Q are used to define the space of that category.

2.1.2 Regression analysis

Classical regression (univariate calibration) analysis and multivariate regression/calibration are here described, the first has been used, in this thesis, for the quantification of the analytes resolved by the MCR methodology (see Section 4.1); while the second is the starting point for understanding its related classification method.

2.1.2.1 Univariate calibration

In chemistry calibration is based on the comparison of two sets of values: the concentration of standards (reference compounds at known concentration values, carefully decided and prepared in laboratory), representing the dependent variables, are graphically reported versus the instrumental response, the independent variable. Each instrument has its personal answer, and different kinds of curves can occur, but generally for small intervals of concentration the parameters of the function of calibration curves are approximated with a line which best fit the data and it is calculated by least squares regression. If x_i and y_i represent the data, the algorithm calculate the function f which minimize the square sum of the Euclidean distance:

$$\min_i \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \text{ [Eq. 4].}$$

For the determination of a line y=*b*x+*a*, two parameters must be calculated (*a* and *b*) with the following equations:

$$b = \frac{N\sum(x_i y_i) - \sum x_i \sum y_i}{N\sum(x_i^2) - (\sum x_i)^2} \text{ [Eq. 5]}$$

$$a = \frac{\sum y_i \sum(x_i^2) - \sum x_i \sum(x_i y_i)}{N\sum(x_i^2) - (\sum x_i)^2} \text{ [Eq. 6].}$$

The least square regression can be extended to more variables in Multiple Linear Regression (MLR) [12].

2.1.2.2 Multivariate Calibration (PLS)

MLR presents important limitations, for instance, in case of rank deficiency (when the number of variable is higher than the number of the objects) or when the variables are collinear (a high correlation exists between absorbances at adjacent wavelengths), the solution produced is unstable in the sense that perturbations of the magnitude of the experimental noise

cause a method to produce different results and overfit the calibration data, thus reducing its applicability in the prediction of new data [13].

Principal Component Regression (PCR) [14,15] and Partial Least Squares Regression (PLS-R) overcome the collinearity problem, the first is not treated in this thesis.

Partial least square regression (PLS) consists in a regression between the scores of **X** matrix and **Y**, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. The PLS model can be considered as consisting of outer relation (**X** and **Y** block individually) and a inner relation (linking both blocks) [16]. The outer relation for the **X** block (analogously to PCA) is:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \sum t_h p_h' + \mathbf{E} \text{ [Eq. 7]}.$$

In the same way it can be built the relation for the Y block:

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F}^* \sum u_h \, q_h' + \mathbf{F}^* \text{ [Eq. 8]}.$$

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:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathbf{T}}+\mathbf{E}_{\mathbf{X}} \text{ [Eq. 9]}$$

$$\mathbf{Y}=\mathbf{U}\mathbf{Q}^{\mathbf{T}}+\mathbf{E}_{\mathbf{Y}} \text{ [Eq. 10]}$$

$$\mathbf{U}=\mathbf{T}\mathbf{C} \text{ [Eq. 11]}$$

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,

i.e. the regression model relating **T** and **U**. The regression coefficient matrix **B** in Equation 12 is then calculated by combining the relations in Equation 11. This regression coefficient matrix allows prediction of the **Y**-values for unknown samples $\mathbf{X}_{new}$ according to:

$$\mathbf{Y}_{new}=\mathbf{X}_{new}\mathbf{B} \quad \text{[Eq. 12]}$$

where the hat indicates predicted values.

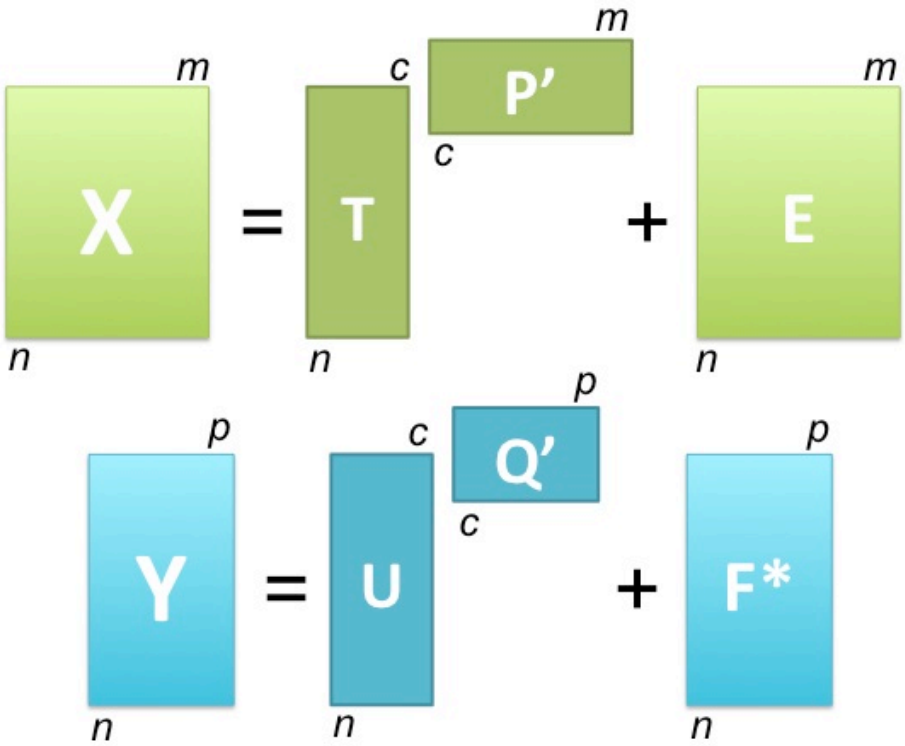

Figure 3: Schematic representation of a PLS regression for X and Y block.

The important part of any regression is its use in predicting the dependent block from the independent one, this is done by decomposing the **X** block and building up the **Y** one using the values calculated for the calibration (*p', q', w', b*).

The number of component to be used is a very important property of a PLS model: if the underlying model for the relation between **X** and **Y** is a linear

model, the number of component needed to describe is equal to the model dimensionality, but non-linear model require extra components to describe non linearity. Best components number can be selected with more than one method, but the most widespread calculates a statistic for lack of prediction accuracy called PRESS (prediction residual sum of squares) and plotting the number of components versus this value the minimum of the curve is selected. Another value can be also considered the RMSECV (root mean square error in cross-validation) [17,18].

2.1.3 Classification

A significant part of the applications of chemometric techniques in analytical chemistry falls in the general framework of pattern recognition, i.e. the classification of objects in groups based on the results of a series of measurements. Pure classification differs from class modeling for some aspects, mainly it is oriented in discriminating among the different groups and operate dividing the hyperspace in as many regions as the number of classes. As a consequence if a sample falls in the region of space corresponding to a particular category, it is classified as belonging to that category: in this way, each sample is always assigned to one and only one class [19].

On the other hand, class modeling techniques represent a different approach to pattern recognition, as they focus on modeling the analogies among the elements of a class rather than on discriminating among the different categories. In class- modeling each category is modeled separately.

With respect to pure classification techniques, class modeling tools offer main advantages: it is possible to identify samples, which don't belong to none of the examined categories. These samples could be outliers or members of a new class not considered in the experiments; moreover, as each category is modeled separately, the model can be update, adding classes without recalculating the already existing class models.

2.1.3.1 Discriminant methods

Discriminant classifiers divide the multidimensional space of the variables in as many regions as the number of given categories, i.e., of classes for which there are training samples available. This approach has as direct consequence that a new sample, depending on the portion of space it falls in, will be attributed to one and only one class. Partial Least Squares – Discriminant Analysis (PLS-DA) represents an example of discriminant classification methods and it has been demonstrated to converge to linear discriminant analysis if the number of PLS latent variables is equal to the number of descriptors (x-variables) [20]. It is a Partial Least Square regression in which the responses matrix **Y** contains the class membership information in a binary coded form, in other words as many columns as the number of given categories and as many rows as the number of samples are defined where, the row corresponding to each individual will have a 1 in the position corresponding to its true class and zeros, or -1 elsewhere.

Accordingly, the classification problem can be reformulated as finding the best regression model linking the experimental data measured on the samples **X** to the binary-coded dummy matrix **Y**:

$$\mathbf{Y} = \mathbf{XB} \text{ [Eq. 13]}.$$

being **B** the matrix of regression coefficients. As the name suggests, in the case of Partial Least Squares-Discriminant Analysis, the PLS algorithm is used to calculate the regression model in Equation 13, 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 for unfolded matrices). As the matrix $\widehat{\mathbf{Y}}_{\mathbf{new}}$ containing the predicted values can assume real values and not only ones and zeros, a classification rule to assign the samples to a given category has to be defined. In general, there are two approaches a “true” discriminant one where classification is accomplished by assigning the samples to the category corresponding to the highest value of the predicted dummy response, e.g. if

the classification problem regards three classes samples belonging to class one will be codified as [1 0 0] (or [1 −1 −1] if the other codification is used), samples belonging to class two will be codified as [0 1 0] (or [−1 1 −1]) and so on.

A second approach is based on the choice of a class threshold for each category, i.e. a value for each dummy y above which the sample is assigned to the class and viceversa if it is under. The threshold is chosen considering the best compromise between *sensitivity* and *specificity* of the class model, usually estimated in cross-validation. Sensitivity and Specificity are defined as:

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \quad [\text{Eq. } 14]$$

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \quad [\text{Eq. } 15].$$

In general, positive corresponds to identified and negative to rejected. Therefore: true positive corresponds to correctly identified, false positive incorrectly identified, true negative correctly rejected and false negative incorrectly rejected. The same concepts are also of common use in class modeling methods.

In order to evaluate the contribute of each descriptor (x-variables) in the discriminant model, the variable importance in prediction (VIP's) parameter may be considered. The VIP score of a predictor, first published in [21], is a summary of the importance for the projections to find h latent variables. The VIP scores for the j^{th} variable can be calculated by Equation 16. On the other hand, since the average of squared VIP scores equals 1, 'greater than one rule' is generally used as a criterion for variable selection.

$$VIP_j = \sqrt{p \sum_{k=1}^{h} \left(SS(b_k \boldsymbol{t}_k) \left(\boldsymbol{w}_{jk} / \|\boldsymbol{w}_k\| \right)^2 \right) \Big/ \sum_{k=1}^{h} SS(b_k \boldsymbol{t}_k)} \quad \text{[Eq. 16]}$$

$$SS(b_k \boldsymbol{t}_k) = b_k^2 \boldsymbol{t}_k^t \boldsymbol{t}_k \text{ [Eq. 17]}.$$

2.1.3.2 Class modeling

Class modeling techniques focus on capturing the similarities between members of the same class rather than on discriminating between individuals from different categories. Class modeling methods assign the samples in this way: when unknown samples are projected, three different situations can occur: in the first case the unknown object falls in the boundaries of a specific class, in this case the object is correctly assigned to that class; in the case in which the unknown object does not fit in any class boundaries it is said to be 'unknown', maybe this object can be an outlier or it can belongs to a new class which has not been modeled. Furthermore another situation can appear, the same object can fit at the same time the parameters for more than one class, in this case it is said to be 'confused'.

SIMCA (Soft Independent Modeling of Class Analogies) is a supervised, class modeling method, whose first version was developed by S. Wold [22] in 1976. It is based on the calculation of disjoint PCA models for each class and the construction of the relative class boundaries. PCA is separately performed on the objects belonging to each class/category, providing a scores-T and loadings-P matrices for each class, that can be summarized in a different dimension by means of a different number of PCs or latent factors (<< of the number of variables). Hence, it is possible to classify and at the same time obtaining information about the properties of each analyzed class. This method is focused on a soft modeling concept in fact, two or more groups or classes can overlap (and hence are 'soft'), i.e. an object can belong to one, more or neither classes (on the contrary the concept of hard modeling implies that an object can belong to only one class, such as in discriminant analysis). This is quite in accordance with many situations in

chemistry and foodstuff authentication, where one object could fit into more then one class simultaneously. Besides, since a distinct PCA is carried out for each class, giving separate class models, it is possible to include independently another different class to the exiting models without need to change the already existing ones (Independent modeling of classes). The classification rule is based on an F-test comparing the distance of a "new" object to the class model with the class residual variance; if the test is passed the object is assigned to that class.

The distance of a "new" object to the class model is estimated by two contributions the distance from the PCA subspace defining the class, called orthogonal distance (OD) and the distance to the boundaries (SD) of the class model from the point where the "new" object is projected on the PCA space. The Orthogonal Distance (OD) represents the Euclidean distance of an observation (a sample) to the PCA subspace. For an object *p*, which is projected in a PCA model of a class *q* ($\hat{p}$ is its projection on the model), the orthogonal distance is calculated by the following Equation:

$$OD_p^{(q)} = s_p^{(q)} = \sqrt{\frac{\sum_i e_{ip}^2}{(M-A)}} \quad [\text{Eq. } 18].$$

In the equation M is the number of variables and A are the component selected for the PCA model for class *q*, e_{ip} are the residuals for the object *p*:

$$e_{ip} = \left(\hat{P}_i - P_i\right) [\text{Eq. } 19].$$

The distance in PCA space, scores distance, SD is defined as:

$$SD_p^{(q)} = \sqrt{t_p \Lambda^{-1} t_p} \quad [\text{Eq. } 20]$$

where t_p are is the *p*-th row of the matrix of scores $\mathbf{T}_A$ in the PCA model of class *q* and $\mathbf{\Lambda}^{-1}$ is the diagonal matrix with the inverse of the eigenvalues referred to the principal component retained in the model.
The assignation of the new object to the class is then accomplished by considering a linear combination of SD and OD $d_p^{(q)}$, which is defined as:

$$d_p^{(q)} = \sqrt{s_p^{(q)} + \sum_a \Phi_a^2 \, (t_a - \vartheta_{a,lim}^{(q)})^2} \text{ [Eq. 21].}$$

where a=1…A_q,; t_a are the scores of the new object and $\vartheta_{a,lim}^{(q)}$ are the scores boundaries for the training set of class *q*; the term Φ_a^2 is a correction factor needed to have both terms on the same scale and $\Phi_a = s_p^{(q)}/s_{\vartheta_a}^{(q)}$ where the denominator is the standard deviation of the training set scores for that PCs
This value is compared with the total residual deviation (RSD) of class *q* representing the variance of the training set objects of class *q*:

$$s_0^{(q)} = \sqrt{\frac{\sum_i \sum_k e_{ik}^2}{(N - A - 1)(M - A)}} \text{ [Eq. 22].}$$

where *M* is the number of variables, *N* the number of the calibration objects of class q, *i* and *k* are their respective index, A is the number of principal component retained and e are the residuals for the training set. The F test compares the ratio $(s_p^{(q)})^2/(d_p^{(q)})^2$, if the observed value is smaller of the critical one the new observation *p* belong to the class *q*. If the F-test is passed for more classes, the object is assigned for the one for which $s_p^{(q)}$ or $d_p^{(q)}$ is smallest.

Original SIMCA (the one developed by Wold) can be schematized with the following Figure 4, which explains the geometrical implication.

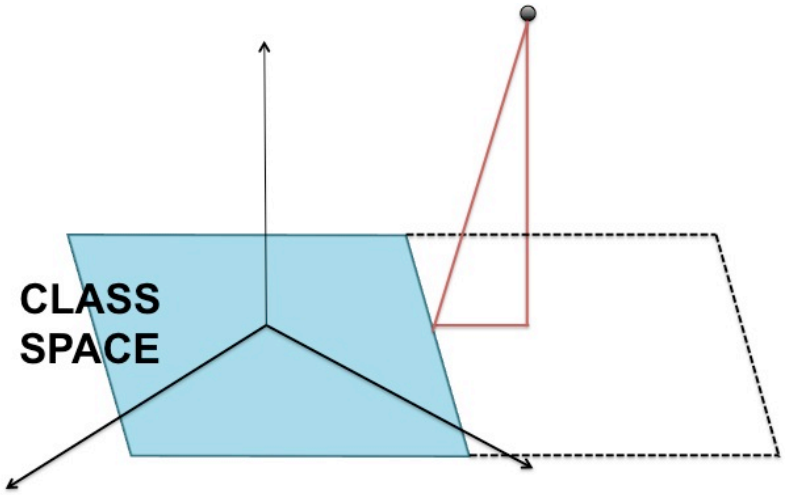

Figure 4: Schematic representation of the original SIMCA developed by Wold, class space is represented by the cyano area.

In the figure $s_p^{(q)}$ and $d_p^{(q)}$ are represented with segments, if the object projection falls in the class space, only $s_p^{(q)}$ is considered, otherwise $d_p^{(q)}$ is also computed always deciding the assignation on the base of an F-test.

During the years, new versions of SIMCA were proposed by several authors [22-27]; main differences were introduced in the definition of the boundary distance [26-32], taking the distance of the projection of the new observation *p* from the center of PCA space instead of the boundaries, i.e. considering the Mahalonobis distance as scores distance, and adjusting the degree of freedom used in Equation 22.

Moreover, a different approach is used in the version implemented in the PLS-toolbox where T^2 and Q are taken as SD and OD, respectively and the calculation of the class boundaries is based on the statistics for these distributions: Hotelling-T^2 to obtain the T^2lim and $\chi 2$ for the Q_{lim}; these kind of statistics coming from the process control have become part of chemometric analysis and they were previously explained in the Section 2.2.1.

In this case a new object is assigned to a class if:

$$\sqrt{\left(\frac{Q}{Q_{lim}}\right)^2 + \left(\frac{T^2}{T^2_{lim}}\right)^2} \leq \sqrt{2} \quad [\text{Eq. } 23].$$

When Equation 23 is true for more than one class, the new object is assigned to the class for which this parameter is lower. This distance measure gives equal weighting to distance in the model space (T^2) and residual space (Q). For the difference between original SIMCA and alternative SIMCA see Appendix I.

2.2 Multiset methods

2.2.1 Multivariate Curve Resolution

In this section multivariate curve resolution (MCR) techniques will be described [33], in particular in this thesis we refer always to the alternating least square MCR. As stated in Chapter 1 this technique make use of a multiset arrangement of the data, where even if finally the structure obtained is a matrix, data related to a single experiment do not lose their original shape of bidimensional landscapes.

2.2.1.1. Alternating least squares

Curve resolution methods were studied by different authors, using different techniques for the calculation, for example some are based on the direct calculation of the two matrices **C** and **S** passing through the intermediate rotation matrix **R**. Other approaches are based on iterative calculations [34–38]. In particular here is applied the alternating least square method (ALS) [36].

This method, like all the other iterative ones, represent directly the values of the matrices without using an initial singular value decomposition, the implementation is based on the iterative calculation of **S** and **C** using the following Equations:

$$\mathbf{S^T} = \mathbf{CD} = \mathbf{C^T C^{-1} C^T D} \text{ [Eq. 24]}$$

$$\mathbf{C} = \mathbf{D}(\mathbf{S^T})^+ = \mathbf{D S S^T S^{-1}} \text{ [Eq. 25].}$$

Matrices marked with the sign ^ indicates an estimated value and $^+$ represent the pseudoinverse. All the iterative method require an initial estimations of either the concentration or the profile spectra matrix to initiate the algorithm, then the two equation are used to iterate the calculation until some convergence criterion are respected and satisfied,

in particular it is stopped if the difference of fit between two consecutive iterations is below a predefined threshold value (0.01% change in standard deviation), alternatively the calculation is topped if the number of iteration exceeded a predefine values (50 default setting) [39].
Percent of variance explained Equation 26 and percent of lack of fit Equation 27 are analyzed analogously to a PCA model for the evaluation of the resolution obtained.

$$expl.var.\% = \frac{\sum d_{ij}^2 - \sum e_{ij}^2}{\sum d_{ij}^2} \times 100 \text{ [Eq. 26]}.$$

$$lof(PCA)\% = \sqrt{\frac{\sum \left(d_{ij} - d_{ij}^*\right)^2}{\sum d_{ij}^2}} \times 100 \text{ [Eq. 27]}.$$

In these equations d_{ij} represents the elements of matrix of experimental data, e_{ij} are the residuals obtained from input elements and MCR-ALS reproduction and d_{ij}^* are the reproduced data using a PCA model with the same number of component of the MCR model. These values help to understand how accurately the matrices obtained fit the original data.

2.2.1.2 Bilinear structure of data

Multivariate Curve Resolution (MCR) is the name, that identify a group of techniques whose aim is to describe the correct underlying contribute of a data set. It can be considered as an extension of the Lambert Beer's Law to higher order as represented in Equation 28, in which the absorbance of a mixture of two compounds is observed at wavelength λ

$$A_\lambda = \varepsilon_{\lambda,X} b c_X + \varepsilon_{\lambda,Y} b c_Y \text{ [Eq. 28]}.$$

where $\varepsilon_{l,x}$ and $\varepsilon_{l,y}$ are the molar absorptivity coefficients at wavelength l of x and y species, respectively; b is the path length of the

spectrophotometric cell and c_x, c_y are the values of concentration of each absorbing species. The most common example of application is represented by an HPLC-DAD analysis of coeluted components, but the same concept can be extended to any analytical method that yields linear additive responses, for example fluorescence, spectroscopy, chromatography and voltammetry. MCR methods discern the individual contribution to the mixture response also when spectral or calibration information is not available, under certain specific circumstances. Considering the Equation 28 for a series of different wavelength we obtain a more generic format:

$$\begin{aligned} d_1 &= c_1 s_{1,1} + c_2 s_{1,2} + \cdots + c_N s_{1,N} \\ d_2 &= c_1 s_{2,1} + c_2 s_{2,2} + \cdots + c_N s_{2,N} \\ &\vdots \\ d_J &= c_1 s_{J,1} + c_2 s_{J,2} + \cdots + c_N s_{J,N} \end{aligned} \qquad [\text{Eq. } 29]$$

$$d_j = \sum_{n=1}^{N} c_n \, s_{j,n} \; [\text{Eq. } 30].$$

In these equations d_j is the absorbance (or more generally an instrumental response) at the j^{th} wavelength, c_n represents the concentration of the nth species in the mixture and $s_{j,n}$ represents the molar absorptivity (or the appropriate instrumental sensitivity factor) at the j^{th} wavelength for the n^{th} species in the mixture. Converting the expression in vector form this becomes:

$$\mathbf{d} = \mathbf{c}\mathbf{S}^{\mathrm{T}} \; [\text{Eq. } 31].$$

MCR techniques are used for second order, where the linear model for the response is given as

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \; [\mathrm{Eq.}\, 32],$$

where matrix **D** (*I* x *J*) contains the individual absorbance or instrumental response for *I* different mixtures measured at *J* different wavelengths, the **C** (*I* x *NC*) matrix represents the concentrations of the *NC* different species in the *I* different mixtures of the compounds, and E(*I* x *J*) represents the error contribution. This model follows a bilinear structure and can be represented also by Figure 5.

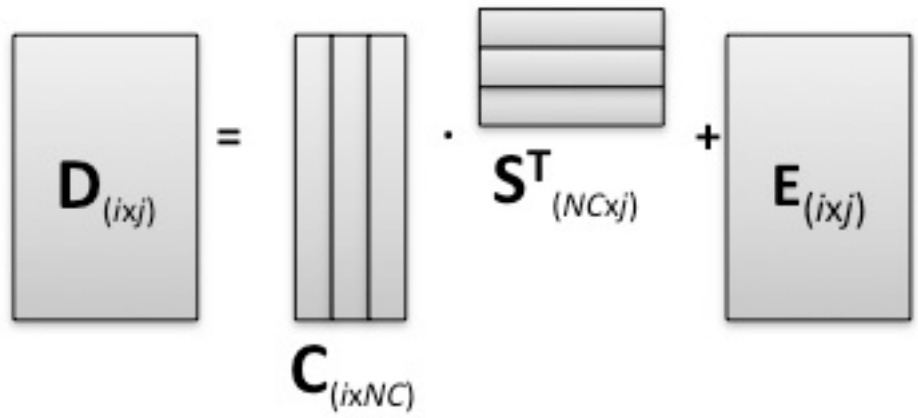

Figure 5: Representation of the bilinear model in terms of matrices.

MCR methods seek to obtain both sensitivities (i.e., pure spectra) (**S**) and concentration (**C**) given a measured data matrix **D**.

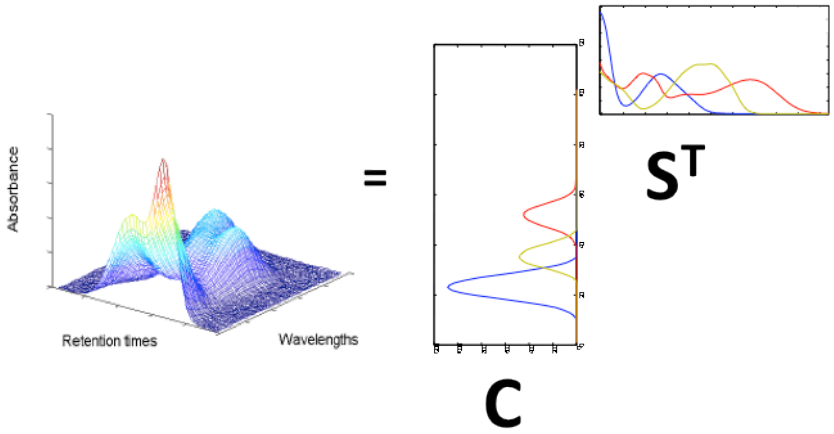

Figure 6: Representation of the bilinear structure of the data using the example of chromatographic peaks clouted and solved as three components.

Clearly there are some aspects in common with the principal component analysis, i.e. both are decomposition methods that describe the original data as the product of two matrices which summarize samples related information and variables contribution to the calculated set of components, however MCR components are not orthogonal and not sequential, hence MCR components are not ordered per variance explained, and the same variance can be explained by different components, it is commonly said that part of the variance is overlapped. Both PCA and MCR solution are not unique, while PCA uses orthogonality of the components as constraint to resolve rotational ambiguity (the fact that more than one solution with the same fit can be obtained up to rotation), MCR uses other constraints that will be described further in this section. Main advantage of this technique is the possibility of extending the resolution to several data matrices simultaneously (multiset structures described in Section 1.2) this feature is the key of the application of this methods to the data coming from hyphenated analytical instrumentation, but also allow application on data fusion and moreover reduce the rank deficiency problem and resolution ambiguities [40].

Using alternating least squares to determine MCR solution implies that an initial estimation of either the concentration or spectral matrix has to be provided, deciding the number of components to be fit. Thus, three issues are relevant in order to get optimal MCR solution: i) number of components; ii) initial estimates; iii) constraints. The following sub sections will illustrate each of these points.

2.2.1.3 Determination of number of components

The number of components, i.e. the underlying phenomena present in the data matrix D, need to be identified and set as one of the input in MCR-ALS, to this aim the knowledge of the chemical problem but also a visual analysis of the data set can help to speculate how many components are needed to solve the system. In the more general case, two methods are particularly useful and widely employed: singular value decomposition (SVD) and Evolving Factor Analysis (EFA) [41-42], which also can give an initial estimation of the data when the profile have a peak shape. The SVD method carry out a decomposition of the data matrix into the matrices of scores and loadings and the profile of percent variance explained (SCREE plot), as in principal component analysis, can be considered for the evaluation of the number of significant components. In case the trend is decreasing in smooth and monotonic way, hence furnishing a not clear indication on how many components to retain. The common procedure is to carry out the resolution with different number of components and then decide considering the goodness of the final results $\mathbf{C}$ and $\mathbf{S}^T$.

2.2.1.4 Initial estimation

The construction of an initial estimation is a central step in MCR. It is possible to estimate both of the two matrices in which the data set is decomposed ($\mathbf{C}$ or $\mathbf{S}^T$). During the iterative cycles these estimations are modified, recalculating in each step their values with the Equations 24 and 25. There are many possibilities to generate the estimation;

chemically meaningful estimates should be preferred with respect to random ones, this solution reduce computational time and convergence problems, moreover different initial estimations may lead to slightly different solutions because of the rotational ambiguity even if all solutions obtained present an optimal fit and are chemical feasible.

In the case in which the system under study is sufficiently know, initial estimate for the **S** matrix can be the spectra of pure components, but this is not necessary and initial estimations can be developed directly from the data or using auxiliary chemometric methods. The selection of a number of columns or rows equal to number of components can provide an initial estimation of respectively **C** and $\mathbf{S}^T$, for example in HPLC-DAD the spectra at the maxima of the observed chromatographic peaks would be a possible estimation of $\mathbf{S}^T$. Auxiliary chemometrics methods are usually preferred, these take advantage of the other direction of the data set; the most common procedure are the estimation calculated with evolving factor analysis (EFA) and the selection of variables with simple-to-use interactive self-modeling analysis (SIMPLISMA).

Evolving factor analysis performs PCA analyses expanding gradually the dimension of the data matrix in the rows direction, the dimension is enlarged by adding a new row at time. This procedure is performed from the first row to the last one of the data matrix (forward EFA) and in the other sense, from the bottom to the top (backward EFA). The eigenvalues obtained at each step are plotted and indicate the evolution of their magnitude along the rows direction, that e.g. in HPLC-DAD/GC-MS, represents the elution process: when the value is above the noise level (marked by the pool of non-significant eigenvalues) this indicate the appearance (in forward EFA) or decay (in backward EFA) of a new eluting compound. This provides a clear indication of the species present in the chromatogram as reported in Figure 7. In the more general case this indicate that a new phenomenon, or underlying factor takes place (forward EFA) or cease (backward EFA) [43,44].

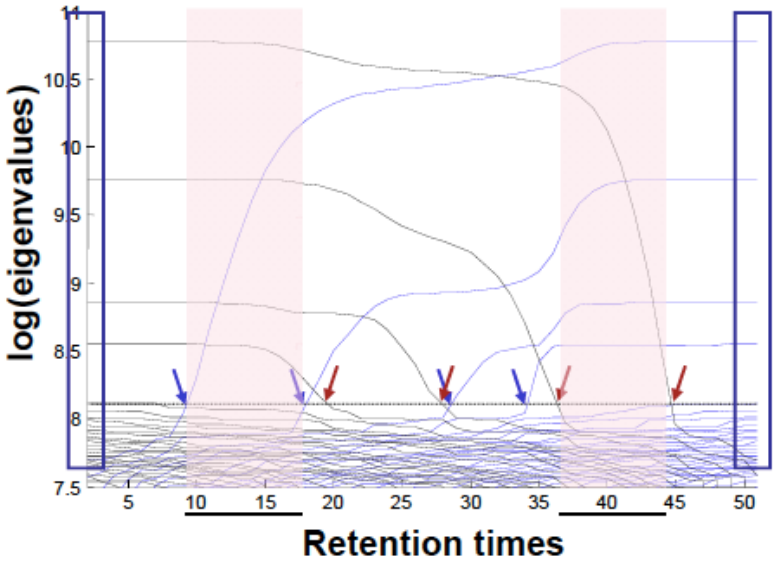

Figure 7: Intial estimation of C profile calculated with EFA. Blue line forward, red line backward, four components estimated are highlighted, blue arrows indicate the point in which the signal (chromatographic peak in this case) starts and red arrows where it decades.

Simplisma [45-46] belongs to a family of methods, which can find an initial estimation in any data set without requiring any special pattern in the variation along the concentration or the response of direction. It selects the most dissimilar, the so-called 'purest', rows or columns. It works selecting in a sequential way the variables in the rows or in the columns direction that have less information in common with the previous selected ones. This method allows generating the initial estimation of both **C** and $\mathbf{S}^T$ matrix: if the purest variables are selected in concentration direction, $\mathbf{S}^T$ estimates are obtained and if the selected variables are spectral responses a **C** estimation is generated. When pure variables selection is performed, best estimates are obtained when the search of purest variables is in the direction of the least overlapped information: in the case of chromatographic data set the selection of the purest elution times seems to be always the best option [47-48].

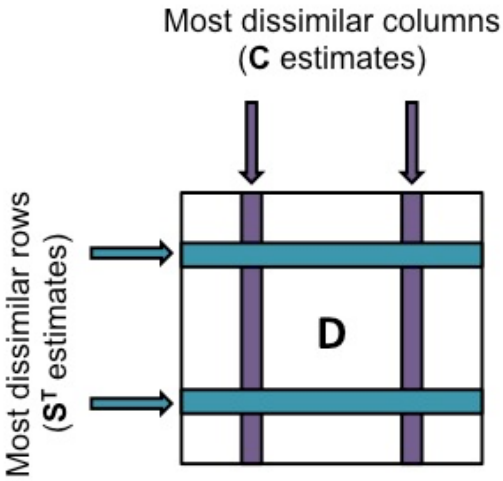

Figure 8: Selection of the purest variables with Simplisma, example of the possibility for creating the initial estimation.

2.2.1.5 Application of constraints

To resolve rotational ambiguity some constraints are applied. These are chemical or mathematical properties that the profiles **C** and/or $\mathbf{S}^T$ obey and that may help to obtain an optimal and chemically meaningful solution. The condition can be set in two ways: forcing the profile (or some elements in the profile) to be equal to some predefined value or to be higher or lower than it [49]. These two options define the equality and inequality condition respectively. Most common constraint applied in chromatography resolution problems are the following:

Non-negativity force concentrations and spectral profiles to be positives, clearly only in the case the chemical nature of the response is positive, which generally holds for spectra and concentrations;

Unimodality force profile to have a single maximum and it is usually applied to concentration profiles;

Selectivity is related to the information of presence/absence of specific components (species) in the experiments, if this selectivity is applied to single part of an experiment we talk about 'local rank' information. This allows taking advantage of experiments where only a component is

present, for example chromatographic injection of standard compounds.

The constraints of non-negativity and unimodality are linked to the chemical properties of the system while selectivity is a mathematical constraint, related to subspaces of the multiset, both strict selectivity and local rank can be separately applied and help to drastically decrease ambiguity in final resolution.

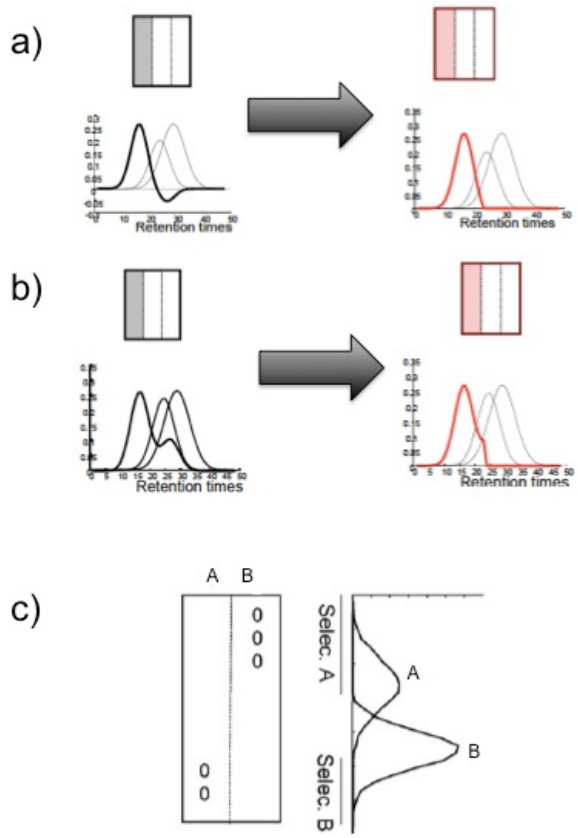

Figure 9: Graphical representation of a) non-negativity, b) unimodality and c) selectivity constraint.

The selectivity constraint practically is input by furnishing a matrix of dimensions equal to the number of samples (in the case of multiset with elution profiles for each sample, in the rows, it is equal to the number of $\mathbf{D}_i$ sub-matrices) per number of components: value 0 is assigned to cell for which the component is absent, and value 1 to all the other cells,

this constraint is usually related to experiments containing only a component prepared ad hoc as references experiments (standards). In the case of local rank, the constraint is imposed through a matrix with dimension equal to the resulting **C** matrix, (namely multiset dimensionality for rows and number of components for lines), if in some parts of the data is verified that only a component is present, all the other are set to be equal to zero and this value is reported in the correspondent cells, see Figure 9; while the other cells contain a Inf (Infinite) or NaN (not a number). The matrix of the local rank information is named *csel*, from the words 'concentration selectivity'. Local rank can be also applied with respect to $\mathbf{S}^T$ matrix constraining the spectra (spectra selectivity), in this case the input matrix is named *ssel*, the matrix is created in the same way as previously explained.
There are also other constraints like closure, for close system profiles, or hard modeling useful when kinetics or equilibriums are studied [50,51], here not described.

2.2.1.6 Reducing system complexity: working with windows

In case where the set of data is constituted by signals which are very complex from the point of view of comprising a huge number of components (e.g. NMR spectra) or when as in the case of elution signals the coelution of components is really strict, it is preferred to work separately on smaller clusters, to speed the calculation, but also to provide more accurate results and a careful application of the constraint. The selection of the segments of data that have to be treated separately can be decided by visual inspection; otherwise also the eventual presence of some target components we are interested in can be a parameter of decision. In the case treated in this thesis, a in house-made MATLAB™ routine is used, in order to select the same window in each chromatogram. This routine works comparing the spectra dimension of a reference compound, e.g. its UV spectrum corresponding to the retention time of the peak maximum in elution direction, with the

spectra of in peak clusters presents at the same retention time in the samples. This allows to identify the compound in real samples also if the elution pattern is not exactly reproduce and build multiset selecting the same elution windows.

Then a visual inspection of the resulting matrices is performed. In particular for each component, all the **C** sub-matrices are plotted and is verified that the solved species are always in a defined zone of the concentration profile matrix, this helps to understand if the resolution has correctly recognized the same components in all the experiments.
Finally one of the most important result to evaluate are the values of the peak areas, in fact MCR calculates the area of the solved **C** matrix, that for chromatographic data correspond to chromatographic peaks. This peak areas (but also concentration values of other kind of data) represent the real fingerprinting information provided from MCR resolution and can be analyzed with all the multivariate technique.

2.2.2 MCR-ALS as discriminant tool

Recently has been developed a new version of MCR called ALS-regression [52], this routine has been implemented to link directly the quantitative information of the standard compounds of a regression, used to optimize the peak resolution, to calibration and obtaining values of calibrated values of concentration for the **C** matrix (or peak area matrix in the case of chromatographic data). Concentration values used to build the calibration line are given in the form of equality constraint for the **C** matrix.
The algorithm calculated a calibration line in a regression step, during the ALS iterations, using as X values the 'csel' matrix holding the equality constraint, i.e. containing 1's and 0's related to class information of the samples; the Y values are the concentration values, **C** calculated at each iteration by the alternating, before the application of the correlation constraint. Then, after convergence, the parameters of

the calibration/regression line (slope and offset) are used to recalculate the final concentration values.

This algorithm was developed for the regression of continuous Y, however there is the possibility of using it for classification, since as in a PLS regression the concentration selectivity constraint can be translated in class information as described above. The salient features of MCR calculation obviously remain the same, so also in this case the number of component necessary to explain the variance can be determined with singular value decomposition. For the initial estimations classical methods can be used as Simplisma, EFA, or empirical/problems based ones (e.g. calculation of sample average for the class). The concentration selectivity matrix (csel) in this case is built as follow: components bearing the class membership: belong/not belong to the class is codified as 1/0 for the other components the values 'infinite' or 'not a number' are given.

1	Inf	Inf
1	Inf	Inf
1	Inf	Inf
-1	Inf	Inf
-1	Inf	Inf
-1	Inf	Inf
Inf	Inf	Inf
Inf	Inf	Inf
Inf	Inf	Inf
Inf	Inf	Inf

Figure 10: Example of a csel matrix for ALS-regression in classification, first column is the component which models the class characteristics (first three samples training set second part test set),.

The ALS-regression algorithm, is studied for data in which each sample corresponds to a row so in is actual form is not suitable for applications

on hyphenated data. This is the reason for which, in this thesis, it is applied as discriminant tool, not on the original data set but on the peak areas solved with a first step of MCR. Namely the data are resolved with a first step of MCR resolution, this is necessary to have a first reduction of variables and matrix rank, since the dimensionality of matrix passes from that of the multiset ([samples x elution time] x spectra dimension], to a fingerprinting matrix [samples x number of species eluted]. Then on the fingerprinting matrix of peak areas obtained is applied the ALS regression discriminant tool. In Figure 11 the relative scheme is illustrated.

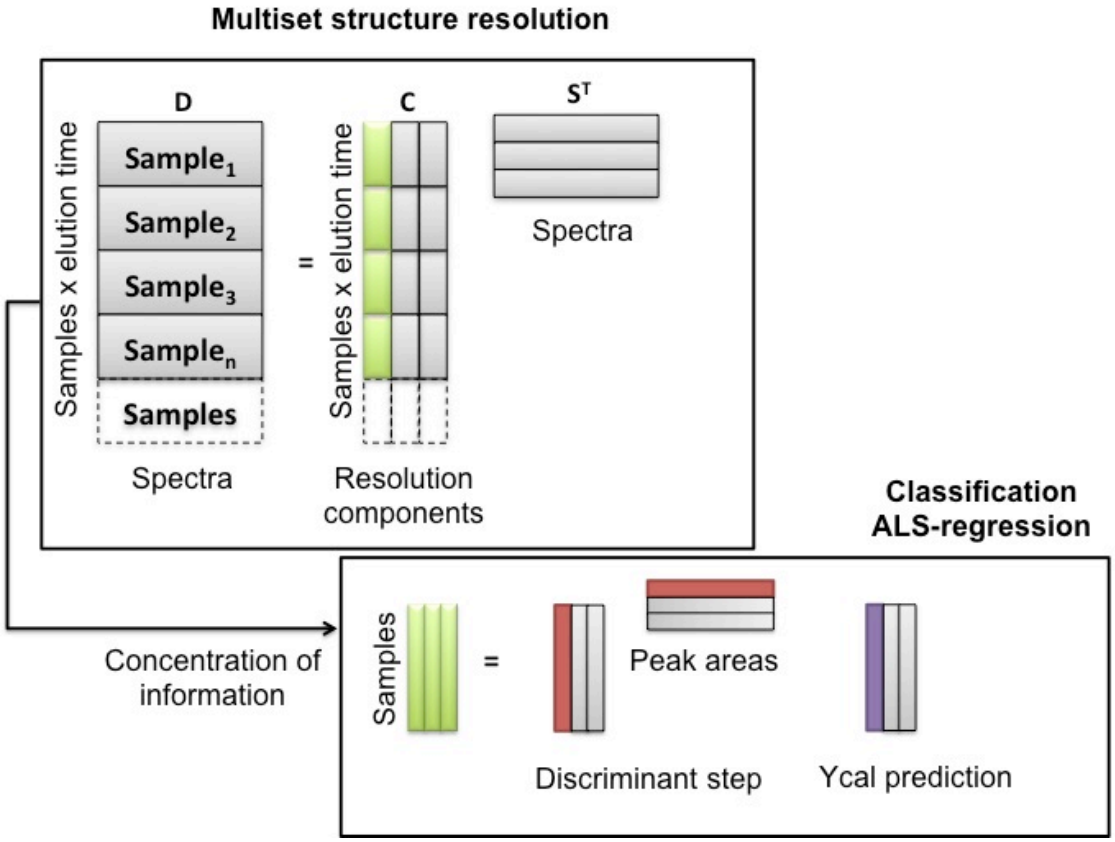

Figure 11: Scheme of the sequential application of MCR and reduction of dimensionality obtained on the final matrix.

2.3 N-way methods

Many cases in food analysis brings to a multi-way approach: for instance storage/ageing, in which modes are represented by samples x variables x time, or sensory analysis (samples x attributed x judges); but of course the most common case regards handling data from hyphenated analytical techniques (Chromatography-DAD or MS detector, fluorescence, etc.).
In this section the most common decomposition methods for multi-way arrays are described are PARAllel FACtor Analysis [53] (PARAFAC) and Tucker3 [54], these are used both for data compression, curve resolution and explorative analysis, but also as classification tools in NSIMCA [55] class modeling.

2.3.1 PARAFAC

The Parallel Factor Analysis (PARAFAC) is a decomposition method for multi-way array; in particular it presumes that the data array has a 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 (for the first mode also scores in analogy with PCA), and the components are named factors: on multi-way analysis the distinction between scores and loadings vectors is often not made and the term 'loadings' can be used for all the modes. For a three-way array $\underline{\mathbf{X}}$ of dimensions *I x J x K* with elements x_{ijk}, the PARAFAC model can be expressed as follow:

$$x_{ijk} = \sum_{f=1}^{F} a_{if}\, b_{jf} c_{kf} + e_{ijk} \text{ [33].}$$

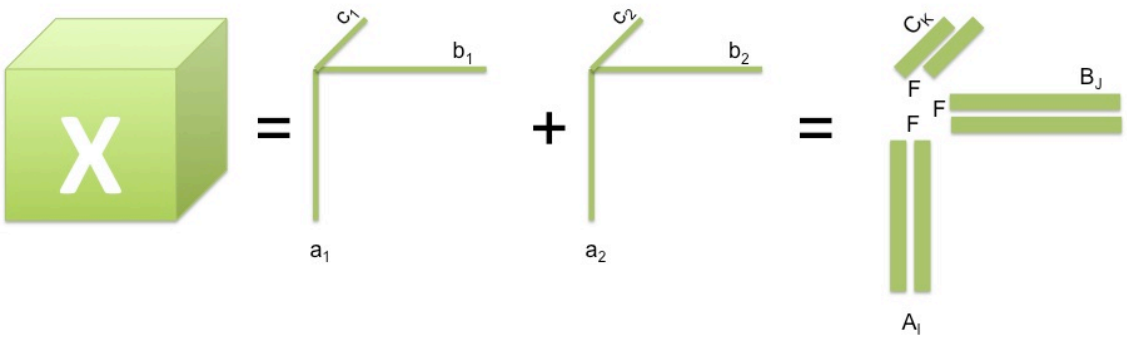

Figure 12: Schematic representation of PARAFAC decomposition with two factors.

A (*I*x*F*) with elements a_{if} is the first mode, score matrix, **B** (*J*x*F*) with elements b_{jf} and **C** (*K*x*F*) with elements c_{kf}, are the second and third modes loadings, respectively. F is the number of components used in the PARAFAC model and e_{ijk} is a residual term containing all the unexplained variation. The extracted factors (or components) are not forced to be orthogonal in a PARAFAC model; in fact, contrary to e.g. a PCA model, PARAFAC has a unique solution, i.e. there is only one set of **A, B** and **C** which can give a solution with that particular fit of the data array [56]. If the right number of components is chosen, if the data are approximately trilinear and the global minimum is found, then the solution is an estimate of the "true" solution, i.e. a chemical meaning solution, for example in the case of HPLC/DAD a solution where the PARAFAC factors for the spectral mode correspond to pure chemical compounds spectral profile, up to permutation and scaling.

It is not easy to select the correct number of components in a PARAFAC model, anyway there are several criteria to orient the choice. The most useful thing is the application of the knowledge of the underlying phenomena, namely the comparison of spectra or chromatographic profiles of some species can be very informative. The analysis of the residuals may as well indicate if the number of component selected is correct, since systematic variation in the residuals means that more components could be extracted. The Split half analysis is another useful

tool: the data are divided in two halves (taking care of representativeness of both halves) and the PARAFAC analysis is separately carried out in the two new data sets. Then it is checked the coherence of the results in terms of similar components profiles in all the array modes, analogous results indicate a robust model and that the correct number of components is chosen [57]. Finally core consistency close to 100%, low number of iterations and similarity of replicates PARAFAC runs, may also be considered as indication of a suitable number of chosen components [58].

2.3.2 Tucker3

Tucker3 decomposition can be seen as a generalization of PCA to higher order array. In fact in Tucker3 a three-dimensional array $\underline{\mathbf{X}}$, of dimension I x J x K, is decomposed into orthonormal triplets of loadings vectors (Figure 13); each triplet is called Component or Factor or Latent Variable (LV), and the advantage respect a PARAFAC decomposition is that the number of extracted components can be different for the three modes.

The extracted components are characterized by a set of scores **A** of dimensions $I \times P$, and two sets of loadings **B** and **C** of dimensions $J \times Q$ and $K \times R$, respectively. P, Q and R are the number of components extracted for each mode.

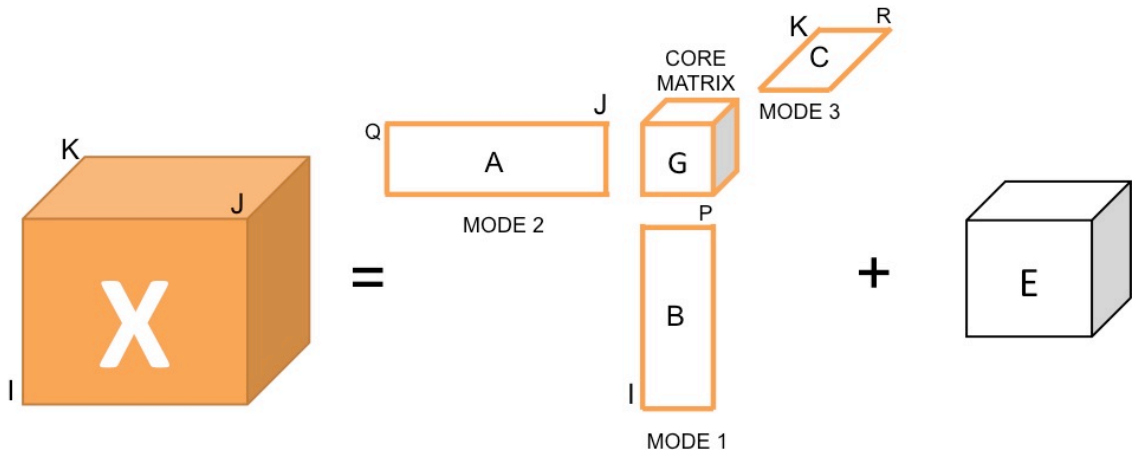

Figure 13: Schematic representation of a Tucker3 decomposition.

Mathematically the Tucker3 decomposition can be represented by the following Equation:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk} \quad [\text{Eq. } 34]$$

a_{ip} , b_{jq} and c_{kr} are elements of the loading matrices **A**, **B** and **C** respectively; g_{pqr} is the element of the core matrix $\underline{\mathbf{G}}$ of order P x Q x R; e_{ijk} denotes the error term or residuals. The model of the original data is the weighted sum of outer products between components in **A**, **B** and **C**. The array $\underline{\mathbf{G}}$, of dimension $P \times Q \times R$, with elements g_{pqr}, is called 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, i.e. 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 [59].

The Tucker3 model does not provide a unique solution. It is possible to estimate infinite of different solutions **A**, **B** and **C** and **G** that fit the same data array **X** equally well. This property is called 'rotational ambiguity' (as for MCR) and it has no impact on the interpretation as long as one is aware of it because the systematic behavior caught by one model is the same in all models. As PCA, the Tucker3 model, working on multi-way data, is able to compress data, extract features, explore data, generate parsimonious models etc. To evaluate the the number of latent variable, namely the correct model dimensionality, in analogy with PCA scree plot, the total number of factor, summed over each mode, against the variance explained by the model, is considered. Variance explained in cross-validation can also be used.

PARAFAC because of uniqueness is usually preferred in spectroscopic and calibration applications to solve the chromatographic peaks (as MCR), while Tucker3 is sometimes preferred in explorative data analysis because of factors orthogonality. However, since PARAFAC models are easier to interpret, if the Tucker3 core array may be made by rotation almost diagonal, PARAFAC has to be preferred. In many cases, the models can be quite similar and the choice may be based on the precision or the speed of the algorithm.

2.3.3 Multi-way classification

In this section the classification methods explained in Section 2.1.3 are extended to multi-way arrays; these methods are a valid alternative to the classical approaches to N-way data set which involves the initial unfolding of the structure converting it in a bidimensional matrix like illustrated in Chapter 1. Introduction.

2.3.3.1 NPLS-DA.

NPLS is defined as an extension of Partial Least Squares Regression (PLS)

to higher order arrays. It was first developed as a PARAFAC-like model [60] of $\underline{\mathbf{X}}$ and it was shown that the method could be easily extended to any desired order for both **X** and **Y** matrices. This method was further elaborated and lastly improved with respect to residual analyses by introducing a core array in the model of $\underline{\mathbf{X}}$ [61-62].
The model makes use of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ arrays of dimension $I \times J \times K$ and $I \times M \times N$ respectively, $\underline{\mathbf{X}}$ is modeled by a Tucker3 decomposition:

$$\mathbf{X} = \mathbf{T}\mathbf{G}_{\mathbf{X}}(\mathbf{W}^{\mathbf{K}} \otimes \mathbf{W}^{\mathbf{J}})^{\mathbf{T}} + \mathbf{E}_{\mathbf{X}} \quad [\text{Eq. } 35]$$

Where **X** is the array unfolded to an $I \times JK$ matrix, **T** is the first mode score matrix, $\mathbf{W}^{\mathbf{J}}$ and $\mathbf{W}^{\mathbf{K}}$ are the second mode weights and the third mode weights, respectively. The symbol ⊗ denotes the Kronecker product [63].
$\underline{\mathbf{G}}_{\mathbf{X}}$ is the core array of size $F \times F \times F$ where F is the number of components and it is defined as:

$$\underline{\mathbf{G}}_{\mathbf{X}} = \mathbf{T}^{+}\mathbf{X}((\mathbf{W}^{\mathbf{K}})^{+} \otimes (\mathbf{W}^{\mathbf{J}})^{+})^{\mathbf{T}} \quad [\text{Eq. } 36]$$

Analogously the model of $\underline{\mathbf{Y}}$ matrix is described by the following equation:

$$\underline{\mathbf{Y}} = \mathbf{U}\mathbf{G}_{\mathbf{Y}}(\mathbf{Q}^{\mathbf{N}} \otimes \mathbf{Q}^{\mathbf{M}})^{\mathbf{T}} + \mathbf{E}_{\mathbf{Y}} \quad [\text{Eq. } 37]$$

Where **U** is the first mode score matrix and $\mathbf{Q}^{\mathbf{M}}$ and $\mathbf{Q}^{\mathbf{N}}$ are the two loadings matrices. $\mathbf{E}_{\mathbf{X}}$ and $\mathbf{E}_{\mathbf{Y}}$ are **X** and **Y** residual matrices, respectively, see Figure 14.
In analogy with the traditional PLS algorithm, the weights are determined such that the scores obtained from the $\underline{\mathbf{X}}$ decomposition (**T**) have maximum covariance with the scores obtained from $\underline{\mathbf{Y}}$ decomposition (**U**).

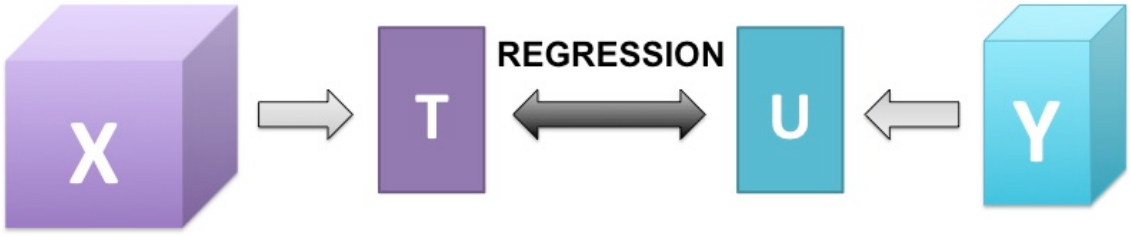

Figure 14: Schematic representation of NPLS decomposition of X (independent variables) and Y (dependent variables) arrays.

Moreover very recently [64] the Variable Influence on Projection (VIP) parameter, used in bilinear PLS-DA to assess the salient discriminant X-variables has been extended to NPLS.

2.3.3.2 NSIMCA

There are not many attempts in extending the class modeling methods to data set of higher order; even if discriminant approaches to multi-way classification exist such as the application of linear discriminant analysis on PARAFAC scores or the just cited multilinear discriminant partial least square (NPLS-DA). This have limited for year the use of SIMCA to bi-dimensional data set, and for the application on data of higher dimensionality an unfolding step was necessary and this procedure can bring to more complex models or with poor predictive ability.

Recently Cocchi et al., extended the SIMCA method to multi-way arrays [55]. This implementation takes advantage of the multi-way decomposition methods, namely PARAFAC or TUCKER3 to build a separate class model for each category, in analogy to the derivation of a separate PCA model for each category in bilinear SIMCA, and it is named NSIMCA (or multi-way SIMCA).

For each class array $\underline{\mathbf{X}}$ of dimension I x J x K, the following steps are performed:

i) Building an independent decomposition model for each class, choosing the appropriate class dimensionality (number of PARAFAC/Tucker3 factors).

Each training set array is separately decomposed through PARAFAC or Tucker3 methods. The proper number of factors is chosen in both cases according to the best classification performance in cross-validation (see Section 2.5 for illustration of cross-validation procedure).

i) To have a preliminary estimate of the range of dimensionality to be explored in the classification step, we calculate PARAFAC/Tucker3 models for a wide range of factors combination and looked at SCREE plot (Tucker3) or the above described model parameters for PARAFAC by using specific routines implemented in the n-way toolbox [78], narrowing on this basis the number of models to be tested. Finally for each class several models were evaluated by NSIMCA and the one, which give the higher Efficiency, i.e. the geometric average of Sensitivity and Specificity, was selected.

ii) Calculation of orthogonal (Q, sum of squared residuals) and scores distances (to give account of scores distance the Leverage values on mode 1 have been considered, see Equation below) for the training class objects in calibration and cross-validation.

$$\mathbf{H} = \mathbf{diag}[\mathbf{T}(\mathbf{T}\mathbf{T}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{T}^{\mathbf{T}}] \quad [\mathrm{Eq.}\,38]$$

iii) Calculation of the classification rules. Different classification rules are defined, hence evaluated, which differs for the listed aspects:

- Using original SIMCA or alternative SIMCA frameworks

(see Section 2.2.3.2 for classification formula);

- For the way the reference statistics limits are estimated for orthogonal distance and score distance.

In the case of scores distance, to obtain the leverage limit reference statistics for leverage both formulations by Forina [65], here referred to as Hlim_fit and Pomerentsev [66], here referred to as Hlim_fit(AP) have been considered. While the *c*-square distribution and the Pomerantsev [67] formulation are both considered to obtain Q-residual limits (named Qlim_fit and Qlim_fitAP). Moreover, for both original-SIMCA and alternative-SIMCA approaches, the same limits are also evaluated in leave one out cross-validation. In original-SIMCA this means that reference class variance has been estimated on residuals values for left-out samples in cross-validation loop (this will be referred to as orig-SIMCA (CV) classification criterion); in alternative-SIMCA this led to calculate the H and Q values for left-out samples in cross-validation loop (HCV, QCV) and their limits HlimCV and QlimCV by using the 95% percentile of the respective set of values.

iv) Estimation of orthogonal and scores distances for the objects belonging to the other classes, different from the one modeled, and for a test set if any, by considering their projection in the modeled class space.

v) On the basis of the classification rules assignation of each sample to one, more or none of the classes.

The NSIMCA code is explained in detail in Appendix II.

2.4 Data preprocessing

2.4.1 Alignment

The most diffused problem treating with chromatographic data is the misalignment in elution profiles of data. Usually these problems are due to different inconvenient in the experimental work like variation in the pressure values provided from chromatographic pump, or aging of the column, which, with time, tends to exert a greater counter pressure, or still, large differences in the concentration of the compounds eluted that is frequent when real samples are analyzed. Clearly peaks alignment is a fundamental pre-treatment step before the application of multivariate models, since for these models the same underlining process must correspond to the same variable and for chromatography this implies that the retention time of the compounds must be invariant across the samples. The alignment issue is treated by various authors [68-70], in this thesis alignment is achieved by using the Interval Correlation Optimized Shifting algorithm (icoShift) [71-72]. This optimizes the piece-wise cross-correlation using the Fast Fourier Transform and a greedy algorithm that allows for user-defined recursion. Among its characteristics rapidity and possibility of define accurately peak clusters on which the algorithm works, are the ones, which made us prefer this to other methods. Moreover even if the original routine was developed to align vectors (1D signals), recently the same authors developed the solution for extending the method to chromatogram with 2D detectors (ApplyIcoShift). The general procedure is that of converting the initial three-way data array to a matrix, considering the TIC signal or calculating the mean over wavelengths, here we reduced the array to the matrix samples vs. total ions count and aligned these signals. After the alignment of the two-way data, the displacement scheme applied for each sample is re-applied to each value in the other Mode (each point in

m/z or wavelength direction). In this way, alignment of the entire 2D-landascape is accomplished.

2.4.2 Baseline correction

When liquid chromatography is coupled with spectroscopic techniques, such as UV-Vis diode array detection, the eluents in the mobile phase very often present an interfering spectrum summed in the response. Usually the common practice is to select an eluent which has spectral bands outside the spectral range recorded, but this effect of the mobile phase is much more evident when a gradient elution is carried out. Normally the background is removed by subtraction of a blank chromatographic run, even if this procedure doesn't always give perfect results, because of variation in spectral intensity of the eluent spectrum during a chromatographic run. Moreover the blank run should be recorded at least once a week for the correction and it would be wasteful in terms of time and consumption of solvents. Baseline may also be corrected mathematically, we corrected for the base line according to the Elimination of Background Spectrum (EBS) method developed by Eilers [73]. This method is based on the Asymmetric Least Square algorithm AsLS (also known as weighted least squares) [74].

The EBS method is a two-steps procedure: first, all variation of the eluent spectra at baseline level is modeled in a background spectral subspace constructed by principal component analysis. Secondly, the spectra measured, during analyte elution, are corrected by performing an AsLS, advantage of the method are that it only needs the data of one single chromatographic run and that each spectrum during analyte elution can be analyzed separately, without using relation to other spectra during elution. The calculation can be performed simultaneously to all the samples. In asymmetric least squares positive residuals and negative residuals do not receive the same weights in this way it is avoided to have negative values in the baseline corrected signals.

2.4.3 Multi-way centering and scaling

In this section the common pretreatments applied to variables are discussed and in particular pretreatment of multi-way data will be explained [75].

Main aims in centering data are: i) to reduce the rank of the model; ii) to increase fit of the data, iii) specific removals of offsets if there are any, iv) to avoid numerical problems. In other words this pretreatment reduce the rank of the model avoiding using an extra component to describe the average behavior of the data. For two-way data centering consists in removing the column average.

In multi-way data arrays centering can apply to one of the mode or more than one modes and it has to be applied taking into account and preserving the multi-way data structure. Centering is done across one mode, e.g. centering across the first mode in a three-way array of dimension $\underline{\mathbf{X}}(I \times J \times K)$ means removing the mean from each columns of the matricized array $\mathbf{X}(I \times JK)$, as shown in Figure 15, the mathematical expression is:

$$y_{ijk} = x_{ijk} - \frac{\sum_{i=1}^{I} x_{ijk}}{I} \text{ [Eq. 39].}$$

where y_{ij} is an element of the centered matrix.

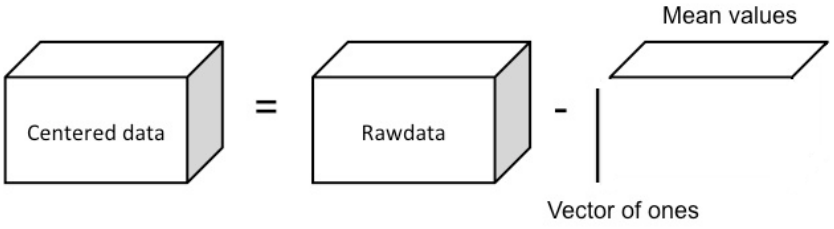

Figure 15: Graphical representation of a centering procedure on the first mode of a three-way array.

Unlike centering, scaling does not change the structure of the model, but gives different weights to parts of the data in fitting the model; it is used for several reasons, some important ones are: i) to adjust scale differenced, ii) to accommodate for heteroscedasticity, iii) to allow for different size of subset of data. Scaling is accomplished within a mode and it means that every entry in the same slab should get the same weight. For example, to scale in the second mode a three-way array $\underline{\mathbf{X}}(I \times J \times K)$ means matricizing the array to a $\mathbf{X}$ (IK, J) and for each j to calculate a scaling factor along IK and apply the same weight to the corresponding IK entry, as shown in the equations below:

$$y_{ijk} = \frac{x_{ijk}}{\sigma_j} \text{ [Eq. 40]}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{I} \sum_{k=1}^{K} x_{ijk}^2}{IK}} \text{ [Eq. 41].}$$

Equation 40 and 41 shows the scaling procedure within the second mode for a three-way array, in which y_{ijk} is the scaled array and σ_j is the standard deviation of the data, in this case the weight given to the data is the inverse of the standard deviation, so that each slab will get the same variance after the pretreatment.

In two-way matrices a most common scaling procedures is *autoscaling*, namely a combination of centering (removing columns average) and scaling to unit variance (weighting by its inverse standard deviation each column), this let the variance of each variable be identical and equal to 1. This type of scaling is especially useful when the variables are measured in different measurement units. In the multi-way case since centering is applied across one mode and scaling within a mode (in other words with different way of matricizing the array) the combination of centering and scaling in different modes affects each other operation. Thus, not only the pretreatments but as well the order in which they are

applied matters. This makes centering and scaling a little bit more tricky in the multi-way case. When the data are made up of several subsets of very different sizes, it may be advantageous to scale each block separately (*blockscaling*) in order to ensure that all the different blocks are allowed to influence the model. For multi-way arrays Blockscaling, e.g. within the second mode is accomplished by rearranging the three-way array to a bi-dimensional $IK \times J$ matrix and then weighting each variable belonging to the same block by:

$$w_j^{block} = \sqrt{\frac{SS_{TOT}}{SS_{block} \cdot n_{block}}} \text{ [Eq. 42] .}$$

In the Equation 42 SS_{TOT} is the total sum of squares over all J's variables, SS_{block} is the sum of squares over the J's variables belonging to the given block and n_{block} is the number of defined blocks. Also in the block-scaling case, the same general considerations drawn for scaling hold.

Pareto-scaling, is another type of scaling in which the applied weight is the inverse of the square root of the standard deviation, this is a milder scaling procedure with respect to weighting to the inverse standard deviation in the sense than small variance variables are less upweighted. Generally block-scaling or pareto-scaling are preferred to standard deviation scaling for chromatographic or spectroscopic data, in order to avoid to give too much importance to variables related to noise, e.g. baseline.

In pretreating multi-way data, each mode must be treated separately; in the case of deriving classification model on the basis of chromatographic data we applied center across the first mode, to remove constant sample contribution, and scaling in the other two modes, both pareto-scaling and block-scaling were tested as well as different order in which to apply the scaling.

2.5 Validation

After a model is calculated (every kind of model!) generally is a good norm to test if the information it gives are correct and if valid conclusions can be drawn from it, thus validation procedures have to be applied. Main aims of validation are: to investigate the quality of the solution, ensure that the obtained model is the simpler one and the more suitable and test its prediction capability on unknown samples. Furthermore, validation is useful to determine the correct number of components to use, to evaluate the estimated model parameters and above all for calibration models, to ensure that the estimated error of predictions is as close as possible to the experimental uncertainty. With the help of validation is also possible to verify the presence of outliers.

2.5.1 Internal validation

When the objects are few and it is not possible to sample a new set of objects, namely the so called test set, the objects are re-sampled from the same data set, this procedure is called *cross-validation*.

One (leave-one-out) or more (leave-more-out, venetian blind, random) samples are left out of the model at time and the model is built with the remaining ones and it is used to predict the left out samples, and the procedure is repeated since each sample is left out once. Root Mean Square Error in Cross-Validation (RMSECV) is computed with this equation:

$$RMSECV = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2} \quad [\text{Eq. } 43]$$

n is the number of objects, $\hat{x}_i$ are the *i*-predicted left out samples, x_i are the *i*-row values of the original data. Generally this value is the one used

for selecting the number of correct component, namely the minimum in the plot of the RMSECV in regression/calibration task.
It is necessary to highlight the difference between cross-validation in regression (PLS NPLS) and decomposition models (PCA PARAFAC Tucker3). In regression models the left-out data are used to calculate the prediction (Y values) on which the prediction error is calculated and used as parameters for the selection of the number of components. In a decomposition model the left-out data are projected into the loadings and find the scores of these left-out samples. As a consequence, the residuals from the recalculated model are not independent from the values that should be used for prediction, which will result in overfitting; i.e., the more components there are, the smaller the residuals. This is not appropriate because the whole idea of cross-validation is to avoid overfitting by estimating the model independently from the data to be modeled. That is the reason for which usually different cross-validation schemes are considered [76].

2.5.2 External validation

When is possible a second set of data are sampled for the validation, or also is a common procedure, if the number of sample allows it, to split the original data in two subsets, in proportion 2:1, the first subset is called *training-set* (or calibration set) and is used to built the model, the second is named *test-set* (or prediction set) and it is projected on the model. To select samples for the calibration and prediction sets one can recur to random selection, but to ensure a homogeneous composition of training and test set it is more suitable to use uniform spanning scheme such as in the Duplex algorithm [77]. This method gives an effective and balanced assignment of data objects in training and test sets, evaluating the mutual distances between pairs of points.

2.6 Software

All the calculation and graphical representation of the data in this thesis were performed with MATLAB™ (Mathworks, version R2010a). In-house-made MATLAB™ routines were written for:

i) The conversion of the ASCII files imported from software of instruments into a MATLAB™ readable file;

ii) Selection of chromatographic windows for the MCR application;

iii) Graphical representation of MCR components;

iv) Construction of calibration line with their confidence limits;

v) Selection of training set and test set with Duplex algorithm.

NSIMCA is written in MATLAB™ code, and it calls some routines contained in the N-way-Toolbox [78] available at the web site http://www.models.life.ku.dk/.

IsoShift is written in MATLAB™ code and available ad the web site http://www.models.life.ku.dk/.

Moreover, two MATLAB™ toolboxes were also used for the data analysis: the PLS-toolbox (Eigenvector Research Inc., version 6.5.2) [79-80] and the MCR GUI (multivariate curve resolution graphical user interface) [81] developed by the chemometrics group of Universitat de Barcelona and IDAEA-CSIC, which is available at the web site http://www.mcrals.info/.

References

[1] S.Wold; *Principal Component Analysis*. Chemometrics and Intelligent Laboratory Systems (1987) 2, 37-52.

[2] I. T. Joliffe, *Principal Component Analysis*, Second Edition, Springer (2002).

[3] K. Pearson; *On lines and planes of closest fit to systems of points in space*. Philosophical Magazine (1901), 2 (6), 559-572.

[4] R. Fisher, W. MacKenzie; *Studies in crop variation. II. The manurial respons of different potato varieties*. Journal of Agricultural Science (1923), 13, 311-320.

[5] F. Malinowski, D. Howery; *Factor Analysis in Chemistry*. Wiley, New York (1980).

[6] L.S. Ramos, K. R. Beebe, W. P. Carey, E. Sanchez, B. C. Erickson, B. E. Wilson, L. E. Wangen, B. R. Kowalski; *Chemometrics*. Analytical Chemistry (1986) 58, 294(R)-315(R).

[7] H. Wold; *Soft modeling by latent variables: the non-linear interative partial least sqyares (NIPALS) approach*. In: Gani, J. (ed.): Perspectives in probability and statistics. Applied Probability Trust, Sheffield, England, 1975.

[8] H. Wold; *Nonlinear estimations by iterative least squares procedure*. In: David, F. N. (ed.): Research papers in statistics, Festschrift for J. Neyman, Wiley, New York, 411-444, 1966.

[9] R.B. Cattel; *The scree test for the number of factors*. Multivariate Behavioral Research (1966), 1 245-276.

[10] G. H. Golub, Reinsch; *Singular value decomposition and least square solutions*. Numerische Mathematik (1970), 14 (2), 403-420.

[11] H. Hotelling; *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology (1933) 24, 417-441 498-520.

[12] L. S: Aiken, S. G. West, S. C. Pitts; *Multiple Linear Regression*. Handbook of Psychology (2003), 481-507.

[13] S. Wold, A. Ruhe, H. Wold, W. J. Dunn III; *The collinearity problem in linear regression, The partial least square approach to generalized inverses*. SIAM, Journal of Scientific and Statistical Computing (1984), 5 735-743.

[14] A. Hoskuldsson; *Prediction Methods in Science and Techonology,* Thor Publishing, Denmark, 1996.

[15] T. Næs, H. Martens; *Principal Component Regression in NIR Analysis: Viewpoints, Background Details and Selection of Component*. Journal of Chemometrics (1988) 2 156-167.

[16] S. de Jong, B. M. Wise, N. L. Ricker; *Canonical Partial Least Square and Continuum Power Regression*. Journal of Chemometrics (2001), 15 85-100.

[17] P. Geladi, B. R. Kowalski; *Partial Least Square Regression: a tutorial*. Analytica Chimica Acta (1986), 186 1-17.

[18] S. Wold, M. Sjöström, L. Eriksson; *PLS-regression: a basic tool of chemometrics*. Chemometrics and Intelligent Laboratory Systems (2001), 58 109-130.

[19] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke; *Supervised Pattern Recognition*. In: Handbook of Chemometrics and Qualimetrics: Part B. Elsevier, Amsterdam, 1998, 207-241.

[20] M. Barker, W. Rayens; *Partial Least Squares for Discrimination*. Journal of Chemometrics (2003), 17 166-173.

[21] S. Wold, E. Johansson, and M. Cocchi; *PLS—partial least squares projections to latent structures*. *3D QSAR in drug design* 1 (1993) 523-550.

[22] S. Wold, M. Sjöström; *SIMCA: A method for analyzing chemical data in terms of similarity and analogy*. Chemometrics: theory and applications.

[23] D. L. Massart; L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke; *Handbook of Chemometrics and Qualimetrics*, Elsevier, Amsterda, (1998).

[24] S. Wold; *Pattern recognition by means of disjoint principal component models*. Pattern Recognition (1997) 8 127-139.

[25] K. V. Branden, M. Hubert; *Robust classification in high dimension based on the SIMCA method*. Chemometrics and Intelligent Laboratory Systems (2005) 79 10-21.

[26] D. J. Louwerse; A. K. Smilde; *Multivariate statistical process control of batch processes based on three-way models*. Chemical engineering science (2000) 55 1225-1235.

[27] R. D. Maesschalck, A. Caldolfi, D. L. Massart, S. Heuerding; *Decision criteria for soft independent modelling of class analogy applied to NIR*. Chemometrics and Intelligent Laboratory Systems (1999) 47 65-77.

[28] S. Wold, M. Sjöström; Letter to the editor – *Comments on a recent evaluation of the SIMCA method*. Journal of Chemometrics (1987) 1 243-245.

[29] C. Albano, G. Blomqvist, D. Coomans, W. J. Dunn III, U. Edlund, B. Eliasson, S. Hellberg, E. Johansson, B. Nordin, D. Jonels, M. Sjöström, B. Soderstrom, H. Wold, S. Wold; *Patter recognition by means of disjoint principal components models: SIMCA Philosophy and Methods*. Proceedings of the Symposium on Applied Statistics, Copenhagen, Jan. 22 1981.

[30] O. M. Kvalheim, K. Oygard, O. Grahl-Nielsen; *SIMCA multivariate data analysis of blue mussels components in environ-mental pollution studies*. Analytica Chimica Acta (1983) 150 145-152.

[31] P.J. Gemperline, L. D. Webber, F. O. Cox; Analytical Chemistry (1989) 61 138-144.

[32] M. Daszykowski, K. Kaczmarek, I. Stanimirova, Y. Vander Heyden, B. Walczac; *Robust SIMCA-bounding influence of outliers*. Chemometrics and Intelligent Laboratory Systems (2006) 47 65-77.

[33] A. de Juan, S. C. Rutan, R. Tauler; *Introduction to multivariate curve resolution*. In: S. Brown, R. Tauler, R. Walzak (eds.) comprehensive Chemometrics, Oxford: Elsevier (2009), 2 259-259.

[34] R. Manne, B. V. Grande; *Resolution of Two-way data from hyphenated chromatography by means of elementary matrix transformations*. Chemometrics and Intelligent Laboratory Systems (2000) 59 35-46.

[35] C. Mason, M. Maeder, A. Whitson; *Resolving factor analysis*. Analytical Chemistry (2001) 73 1587-1594.

[36] R. Tauler; *Multivariate curve resolution applied to second order data*. Chemometrics and Intelligent Laboratory Systems (1995), 30 133-146.

[37] P. G. Gemperline; *A priori estimate of the elution profiles of the pure components in overlapped liquid chromatography peaks using Target Factor Analysis*. Journal of Chemical Information and Computer Sciences (1984) 24 206-212.

[38] B. G. M. Vandeginste, W. Derks, G. Kateman; *Multicomponent self-modelling curve resolution in high-performance liquid chromatography by Iterative Target Transformation Analysis*. Analytica Chimica Acta (1985) 173 253-264.

[39] S. C. Rutan, A. de Juan, R. Tauler; *Two-way data analysis: multivariate curve resolution – interative resolution methods*. In: S. Brown, R. Tauler, R. Walzak (eds.) comprehensive Chemometrics, Oxford: Elsevier (2009), 2 325-344.

[40] R. Tauler, M. Maider, A. de Juan; *Multiset data analysis: extended multivariate curve resolution*. In: S. Brown, R. Tauler, R. Walzak (eds.) comprehensive Chemometrics, Oxford: Elsevier (2009), 2 473-505.

[41] H. Gampp, M. Maeder, C. J. Meyer, A. D. Zuberbühler; *Calculation of Equilibrium constant from multiwavelength spectroscopic data. III. Model-free analysis of spectrophotometric and ESR titrations.* Talanta (1985) 32 1133-1139.

[42] M. Maeder, A. D. Zuberbühler; *The resolution of overlapping chromatographic peaks by evolving factor analysis*. Analytica Chimica Acta (1986) 181 287-291.

[43] A. de Juan, R. Tauler; *Factor Analysis of hyphenated chromatographic data Exploration, resolution and quantification of multicomponent systems*. Journal of Chromatography A (2007), 1158 184-195.

[44] A. R. Keller, D. L. Massart, G. O. De Beer; *Window evolving factor analysis for assessment of peak homogeneity in liquid chromatography*. Analytical Chemistry (1993), 65 471-475.

[45] W. Winding, J. Guilment; *Interactive Self-Modeling mixture analysis*. Analytical Chemistry (1991) 63 1425-1432.

[46] B. V. Grande, R. Manne; *Use of convexity for finding pure variables in in two-way data matrix from mixtures*. Chemometrics and Intelligent Laboratory Systems (2000) 50 19-33.

[47] W. Winding, N. B. Gallagher, J. M. Shaver, B. M. Wise; *A new approach for interactive self modeling mixture analysis*. Chemometrics and Intelligent Laboratory Systems (2005) 77 85-96.

[48] W. Winding; *Two-Way Data Analysis: Detection of Purest Variables*. Chemical and Biochemical Data Analysis (2009) 2 275-307.

[49] R. Tauler, A. de Juan; *Multivariate Curve Resolution*. In: Practical Guide to Chemometrics; Ed. P. Gemperline; CRC Press: Boca Raton, FL 2006 417-474.

[50] A. de Juan, M. Maeder, M. Martinez, R. Tauler; *Combining hard and soft modeling to solve kinetic problems*. Chemometrics and Intelligent Laboratory Systems (2000) 54 123-141.

[51] J. Diewok, A. de Juan, M. Maeder, R. Tauler, B. Lendl; *Application of a combination of hard and soft modelling for equilibrium systems to the quantitative analysis of pH-modulated mixture samples*. Analytical Chemistry (2003) 75 641-647.

[52] A. de Juan, R. Tauler; ALS-regression (In preparation).

[53] J. D: Carrol, J. Chang; *Analysis of individual differences in multidimensional scaling via N-way generalization and Eckart-Young decomposition*. Psychometrika (1970) 35 283-319.

[54] L. R. Tucker; *Some mathematical notes on three-mode factor analysis*. Psychometrika (1966) 31 279.

[55] C. Durante, R. Bro, M. Cocchi; *A classification tool for N-way array based on SIMCA methodology*. Chemometrics and Intelligent Laboratory Systems (2011) 106 (1), 73-85.
[56] R. Bro; PARAFAC. *Tutorial and application*. Chemometrics and Intelligent Laboratory Systems (1997) 38 (2), 149-171.

[57] R. A. Harshman, M. E. Lundy; *The PARAFAC model for three-way factor analysis and multidimensional scaling*. In: H. G. Law, C. W. Snyder, J. A. Hattie, R. P. McDonald (Eds.). Research methods for Multimode data analysis, Praeger, New York (1984).

[58] R. Bro, H. A. L. Kiers; *A new effcient method for determining the number of components in PARAFAC models*. Journal of Chemometrics (2003) 17 274-286.

[59] P. M. Kroonenberg, J. de Leeuw; *Principal component analysis of three-mode data by means of alternating least squares algorithms*. Psychometrika (1980) 45 69.

[60] A. K. Smilde; *Comments on multilinear PLS*. Journal of Chemometrics (1997) 11 367-377.

[61] S. de Jong; *Regression coefficients in multilinear PLS*. Journal of Chemometrics (1998) 12 77-81.

[62] R. Bro; Multiway calibration - *Multilinear PLS*. Journal of Chemometrics (1996) 10 47-61.

[63] J. R. Schott; *Matrix Analysis for statistics*. John Wiley and Sons, New York, 1997.

[64] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Chemom. Intell. Lab. Syst, submitted.

[65] M. Forina, M. Casale, P. Olivieri, S. Lanteri*; CAIMAN brothers: A family of powerful classification and class modeling techniques*. Chemometrics and Intelligent Laboratory Systems (2009) 96 239-245.

[66] A. L. Pomerantsev; *Acceptance areas for multivariate classification derived by projection methods*. Journal of Chemometrics (2008) 22, 601-609.

[67] R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan; *CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scaled functions*. Chemometrics and Intelligent Laboratory Systems (2009) 96 239–245.

[68] N. P. V. Nielsen, J. M. Carstensen, J. Smedsgaard; *Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping*. Journal of Chromatography A (1998) 805 17–35.

[69] G. Tomasi, F. van den Berg, C. Andersson; *Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data*. Journal of Chemometrics (2004) 18 231–241.

[70] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides; *Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data*. Journal of Chromatography A (2002) 961 237–244.

[71] F. Savorani, G. Tomasi, S. B. Engelsen; *icoshift: A versatile tool for the rapid alignment of 1D NMR spectra*. Journal of Magnetic Resonance (2010) 202 190–202.

[72] G. Tomasi, F. Savorani, S. B. Engelsen; icoshift: *An effective tool for the alignment of chromatographic data*. Journal of Chromatography A, (2011) 1218 7832– 7840.

[73] H. F. M. Boelens, R. J. Dijkstra, P. H. C. Eilers, F. Fitzpatrick, J. A. Westerhuis; *New Background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection*. Journal of Chromatography A (2004), 1057 21–30.

[74] P. H. C. Eilers; *Parametric Time Warping*. Analytical Chemistry (2004), 76 404–411.

[75] R. Bro, A. K. Smilde; *Centering and scaling in component analysis*. Journal of Chemometrics (2003), 17 16–33.

[76] R. Bro, K. Kjeldahl, A. K. Smilde, H. A. L. Kiers; *Cross-validation of component models: A critical look at current methods*. Analytical and Bioanalytical Chemistry (2008) 390 1241–1251.

[77] R. D. Snee; *Validation of Regression Models: Methods and Examples*. Technometrics (1977) 19 415-428.

[78] C. A. Andresson, R. Bro; *The N-way Toolbox for* MATLAB™. Chemometrics and Intelligent Laboratory Systems (2000) 52 1-4.

[79] B. M. Wise, N. B.Gallagher, R. Bro, J. M. Shaver, W. Winding, R. S. Koch: *PLS_Toolbox for use with* MATLAB™. Eigenvector Research, Inc., Wenatchee, USA, 2006.

[80] PLS-Toolbox Manual (ver. 4.0) for MATLAB™ (distributed by Eigenvector Research, WA, USA).

[81] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler; *A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in* MATLAB™. Chemometrcs and intelligent laboratory systems (2005) 76, 101-110.