A MID LEVEL DATA FUSION STRATEGY FOR THE VARIETAL CLASSIFICATION OF LAMBRUSCO P.D.O. WINES

M. Silvestri[1], A. Elia[1], D. Bertelli[2], E. Salvatore[1], C. Durante[1], M. Li vigni[1], A. Marchetti[1], M. Cocchi[1*]

[1]*Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 183, Modena (Italy)*

[2]*Department of Life Sciences, University of Modena and Reggio Emilia, Via Campi 183, Modena (Italy)*

*corresponding author

Marina Cocchi, PhD

Tel: 0039-059-2055029

Fax: 0039-059-373543

E-mail: marina.cocchi@unimore.it

Abstract

Nowadays the necessity to reveal the hidden information from complex data sets is increasing due to the development of high-throughput instrumentation. The possibility to jointly analyze data sets arising from different sources (e.g. different analytical determinations/platforms) allows capturing the latent information that would not be extracted by the individual analysis of each block of data. Several approaches are proposed in the literature and are generally referred to as data fusion approaches. In this work a mid level data fusion is proposed for the characterization of three varieties (*Salamino di Santa Croce*, *Grasparossa di Castelvetro*, *Sorbara*) of *Lambrusco* Wine, a typical P.D.O. wine of the District of Modena (Italy). Wine samples of the three different varieties were analyzed by means of ^{1}H-NMR spectroscopy, Emission-Excitation Fluorescence Spectroscopy and HPLC-DAD of the phenolic compounds.

Since the analytical outputs are characterized by different dimensionality (matrix and tensor), several multivariate analysis were applied (PCA, PARAFAC, MCR-ALS) in order to extract and merge, in a hierarchical way, the information present in each data set.

The results showed that this approach was able to well characterize *Lambrusco* samples giving also the possibility to understand the correlation between the sources of information arising from the three analytical techniques.

Keywords

Data-Fusion; HPLC-DAD; NMR; EEM; *Lambrusco* wine; Varietal-Classification; Multiset-MCR; PARAFAC; PLS-DA

1 Introduction

Data fusion is nowadays an emerging branch in chemometrics [1-8], in fact the possibility to jointly analyze by means of different analytical techniques a given set of samples enhances the quantity and the quality of the information which can be extracted. The nature and dimensionality of the data obtained by several instruments require the use of multivariate data analysis tools in data elaboration and fusion with the aim of capturing latent information that would not be extracted by the individual analysis of each block of data. When the analysis of a sample is performed at the same time by means of variables of different nature, not only the goodness of the sample characterization could improve, but also the correlation and the similar/different information content pertaining to the different variables can be assessed.

Several data fusion frameworks are reported in literature, and depending on the field of application, many terms are used interchangeably to indicate the same procedure, e.g. low-level data fusion is synonymous of concatenated or merged data fusion. In chemometrics data fusion techniques are commonly classified in three main groups [1,2]: (a) low-level data fusion, which consists of the simple concatenation of the data of different nature; (b) mid-level or hierarchical data fusion, which corresponds to the fusion of features extracted from the original data, i.e. prior to the fusion step some data reduction is operated, either by multivariate tools, e.g. Principal Component Analysis (PCA) retaining the scores for a given number of components, or through features selection strategies; (c) in high-level or decision level data fusion different models are built separately for each block of data and then the model responses are merged to produce a final "fused" response.

In particular, dealing with mid-level data fusion methodologies there are two levels of modeling: a low level, where data reduction is operated obtaining a latent variables model for each single analytical platform, and an higher multivariate data analysis level, where these latent variables are fused to build the final model. The main issues to face concern the features to retain from each model, the scaling to adopt and the most efficient method to be used for data reduction.

In the field of food analysis, which contemplates in almost all cases the presence of complex matrices, interferences, additional uncontrolled variability due to biological, chemical and physical transformations, the necessity to unveil the hidden information related to the investigated product is a crucial task to face and data fusion methodologies may represent a very efficient tool.

This work focuses on the development of a model able to distinguish between three *Lambrusco* wines with Protected Denomination of Origin (PDO), typical of the District of Modena (Italy).

Lambrusco wines are the most exported wines all over the world (more than 450.000 hL/year produced) and are the oldest wines in the Region of Emilia Romagna.

Since the three varieties, considered in this study, belong to the same family of grapes and the production procedure are quite similar, all *Lambrusco* are red sparkling wines produced with a second fermentation in pressurized tanks, investigating the differences between the three typologies of wines from a chemical point of view is a difficult task to face.

The aim of this work is twofold: on one hand to develop a classification model for the identification of the three typologies of *Lambrusco* wines, on the other hand, to better understand the correlation and analogy between the considered sources of information. In order to characterize *Lambrusco* wines a mid-level data fusion approach is proposed. *Lambrusco* wine samples were analyzed by means of Nuclear Magnetic Resonance (^{1}H-NMR), Fluorescence Excitation-Emission Matrix Spectroscopy (EEM), and liquid chromatography (HPLC-DAD). The first two are non-destructive techniques useful to acquire a fingerprint of the food product, while HPLC-DAD was used to extract and characterize the phenolic fraction.

The data arising from the three analytical techniques, considered in this study, are arranged as two and three way arrays. Thus, appropriate different data analysis tools such as PCA, Parallel Factor analysis (PARAFAC) and Multivariate Curve Resolution (MCR), were used as explorative analysis and data reduction tools, in particular PCA for ^{1}H-NMR data, parallel factor analysis (PARAFAC) for EEM data and multivariate curve resolution (MCR) was used to resolve the phenolic fraction from HPLC-DAD data. The features extracted in the data reduction step (PCA scores, mode one PARAFAC loadings and the peaks areas of MCR resolved components) were then merged together and a classification model based on PLS-DA was finally computed for the merged data set. In addition, to compare the results of data fusion with the performance of the single set of data, three classification models, based on PLS-DA and NPLS-DA were built for the three separate data sets, respectively.

2 Materials and methods

2.1 Experimental

According with their regulation (DM July 27, 2009; GU no. 184-187-188, 13/08/2009) four typologies of *Lambrusco* wines can be produced as PDO product, namely: *Lambrusco Grasparossa di Castelvetro* PDO, *Lambrusco di Sorbara* PDO, *Lambrusco Salamino di Santa Croce* PDO and *Lambrusco* di Modena PDO. The ampelographic composition, set by the production codes, is defined as follow: at least 60% of *Lambrusco di Sorbara* grapes, up to 40% of *Lambrusco Salamino*

grapes, up to a maximum 15% of other *Lambrusco* grapes either of one variety or in combination for *Lambrusco di Sorbara*; for *Lambrusco Salamino di Santa Croce* and *Lambrusco Grasparossa di Castelvetro* at least 85% of grapes from vines of the same name, respectively, and the remaining 15% of other *Lambrusco* (*Ancellotta, Fortana,* and *Malbo Gentile*) grapes; at least 85% of *Lambrusco* grapes of the three PDO different varieties, up to a maximum 15% of grapes from *Ancellotta, Malbo Gentile*, and *Fontana* vines for *Lambrusco di Modena*.

Fifty-eight *Lambrusco* wine samples, produced in the 2009 harvest, were provided by several local wineries through the producers consortium. A total of nineteen *Lambrusco Grasparossa di Casltelvetro* PDO, twenty *Lambrusco Salamino di Santa Croce* PDO and nineteen *Lambrusco di Sorbara* PDO were analyzed with the three analytical techniques described below.

2.1.1 ^{1}H-NMR data

The acquisition of the ^{1}H-NMR spectra is described in detail in [10]. Wine samples were buffered at pH 2.00 in order to avoid the chemical-shift variation of pH dependent signals and then lyophilized to reduce the ethanol signals. ^{1}H-NMR spectra were acquired with a Bruker FT-NMR Avance 400 spectrometer (Bruker Biospin GmbH Rheinstetten, Karlsruhe, Germany) operating at 400.13 MHz. The experiments were performed at a temperature of 300 K and non-spinning. The ^{1}H NMR data were acquired using the Bruker spin−echo sequence "cpmgpr.fb" (Carr−Purcell−Meiboom−Gill, Bruker Library) with water presaturation, applied to suppress broad resonance signals. [11].

2.1.2 EEM data

The acquisition of EEM signals was performed on degased wine samples using a FLS920 fluorescence spectrometer (Edinburgh Photonics) equipped with a variable-angle front-face accessory, to ensure that reflected light, scattered radiation, and depolarization phenomena were minimized. The angle of incidence, defined as the angle between the excitation beam and a perpendicular to the cell surface, was 30°. Wine samples were placed in a 3 mL quartz cell, and spectra were recorded at 20 °C. The excitation range spanned the region from 340 nm to 240 nm (5 nm steps), the emission range considered was from 500 nm to 300 nm (1nm steps). The landscapes were registered as multiple emission spectra at decreasing excitation wavelengths (from 340 nm to 240 nm), total scanning time per sample was about ten minutes.

2.1.3 HPLC-DAD data

The chromatographic runs conditions are treated in detail in our previous works [9] here only main analysis steps will be recalled. Separation of the phenolic compounds by means of solid phase extraction, Supelco DSC-18 cartridges with 6 mL tubes, was performed prior to the acquisition of the chromatograms. Samples were analyzed by reverse phase liquid chromatography by using a

Beckman System Gold coupled with a diode array detector. The column used was a reverse-phase Atlantis dC18 Waters-Milford-MA. The mobile phase used was formed by two solvents: solvent A: water (0.1% TFA); solvent B: 80 % acetonitrile and 20% water (0.1% TFA). An elution linear gradient was used. The wavelength range in the diode array detector was from 220 to 430 nm with a resolution of $\Delta\lambda$=2 nm.

2.2 Data analysis

2.2.1 Decomposition Methods

Since the analytical techniques considered in this study produce outputs that are characterized by different data structure, i.e. data matrix and three way arrays, several chemometric algorithms were applied in order to extract and merge, at mid-level fusion, the information presents in each data set. The NMR data set was compressed by principal component analysis (PCA); the EEM data array was analyzed by parallel factor analysis (PARAFAC) [12], in fact a PARAFAC model ideally decomposes the fluorescence landscapes into trilinear components according to the number of fluorophores present in the samples. Finally, for HPLC-DAD data, considering the presence of shift in the retention time and also the interest in quantifying some of the phenolic compounds, multivariate curve resolution (MCR) was used to resolve the chemical components present and, the peaks areas of the resolved components were then used in building the fused data set.

PARAFAC [12] is a trilinear decomposition method for multi-way arrays. It assumes that the data are organized in a trilinear structure and can be decomposed as a sum of triple outer product of vectors:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad \text{Eq. (1)}$$

The three sets of vector, ***a***, ***b*** and ***c***, are called loadings (the term scores is also used for the loadings of the first or samples mode). Constraints such as non-negativity, unimodality etc. can be eventually applied in order to obtain a chemical meaningful solution.

Multivariate Curve Resolution (MCR) is a bilinear decomposition method [13] based on the multi-wavelength extension of Beer's absorption law: $X = C \times S^T + E$, where ***C*** holds the concentration profiles and ***S*** the pure components spectra. In this work the alternating least squares (ALS) algorithm MCR-ALS [14] was used. In MCR the resolved components are not forced to be orthogonal. Moreover MCR suffers from, rotational as well as intensity ambiguity and in order to minimize these issues several constraints can be applied. The constraints, or rather, the translation in a mathematical way of a chemical/physical behavior, help the resolution step and at the same

time assist in giving chemical meaningful sense to the resolved profiles. Two types of constraints can be applied: soft constraints, such as non-negativity, unimodality, selectivity and closure, allow reducing rotational ambiguity; hard constraints, based on physicochemical models such as kinetic or equilibrium model, reduce at the same time both rotational and intensity ambiguities.

2.2.2 Classification Methods

The discriminant partial least squares regression methods (PLS-DA) [15] and its multi-way extension, multilinear PLS-DA (NPLS-DA) [16], were used in this work. These two classification techniques derive from the well-known regression algorithms PLS and NPLS where the dependent variable block, **Y**, holds the class information: as many y-variables as number of classes are defined as dummies variables assuming values one/minus one to indicate class membership (one/zero notation can be used as well). We adopted as classification rule that samples are assigned to the class for which the predicted y-value is higher, e.g. if the predicted vector of responses for an unknown sample, is: [-0.7 0.7 0.2] (in the case of a three classes problem), it will be assigned to class two.

In order to assess PLS-DA/NPLS-DA dimensionality, i.e. number of components, the minimum classification error rate in cross-validation (venetian blind, ten splits) was considered.

2.2.2.1 NPLS-DA

Briefly, considering an $\underline{\mathbf{X}}$ array of dimension $I \times J \times K$, the NPLS model is obtained by modeling $\underline{\mathbf{X}}$ as in Tucker3 decomposition:

$$\mathbf{X} = \mathbf{T}\mathbf{G}_X(\mathbf{W}^K \otimes \mathbf{W}^J)^T + \mathbf{E}_X \qquad \text{Eq. (2)}$$

where $\mathbf{X}$ is the $\underline{\mathbf{X}}$ array unfolded to an $I \times JK$ matrix**,** $\mathbf{T}$ holds the first mode scores (samples mode), $\mathbf{W}^J$ and $\mathbf{W}^K$ are the second and the third mode weights, respectively. The symbol $\otimes$ denotes the Kronecker product. $\mathbf{G}_X$ is the matricized core array of size $F \times F \times F$ where F is the number of NPLS components (factors) and it is defined by:

$$\mathbf{G}_X = \mathbf{T}^+\mathbf{X}((\mathbf{W}^K)^+ \otimes (\mathbf{W}^J)^+)^T \qquad \text{Eq. (3)}$$

In the case of a two-ways data matrix, $\mathbf{Y}_{I,M}$ is defined by:

$$\mathbf{Y} = \mathbf{U}\mathbf{Q} + \mathbf{E}_Y \qquad \text{Eq. (4)}$$

where $\mathbf{U}$ holds the $\mathbf{Y}$ scores and $\mathbf{Q}$ is the $\mathbf{Y}$-loadings matrix. $\mathbf{E}_X$ and $\mathbf{E}_Y$ hold $\underline{\mathbf{X}}$ and $\mathbf{Y}$ residuals, respectively ($\mathbf{E}_X$ is in matricized form $I \times JK$).

In analogy with the two-ways PLS algorithm, the weights are determined such that the scores obtained from the $\underline{\mathbf{X}}$ decomposition ($\mathbf{T}$) have maximum covariance with the scores obtained from $\mathbf{Y}$

decomposition (inner relation: $\mathbf{U} = \mathbf{TB} + \mathbf{E}_U$). Regression coefficients that apply directly to $\mathbf{X}(I \times JK)$ may also be derived [17] to re-express the NPLS model as:

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B}_{PLS} \qquad \text{Eq. (5)}$$

2.3 Data split

The dataset, which contains 58 samples, was split in training (forty-five samples, fourteen *Lambrusco Grasparossa*, sixteen *Lambrusco Salamino* and fifteen *Lambrusco Sorbara*) and test set (thirteen samples, five *Lambrusco Grasparossa*, four *Lambrusco Salamino* and four *Lambrusco Sorbara*) using the Duplex algorithm [18] on the NMR data set. The same split was maintained for the other data sets and the fused data set, after checking, by exploratory data analysis, that both sets spanned the whole variability domain.

2.4 Preprocessing and analysis of separate data set

2.4.1 NMR data set

Prior to data analysis, several preprocessing were tried and applied on NMR signals. The region below 1.24 ppm was cut, since containing ethanol signals not completely removed in the lyophilization step. To remove the inhomogeneous pH dependent chemical shifts, all spectra were aligned by means of Icoshift [19]. Weighted least squares WLS [20] was used for baseline correction. The ^{1}H-NMR spectra of wines samples are characterized by region in which a lot of intense signals are present (e.g. 3ppm–5ppm) and regions with low intensity signals (e.g. aromatic region at 7ppm). For these reasons, after normalization (1-norm), block scaling was applied in order to give to each region the same variance. In practice block scaling applies the same weight, equal to the inverse of the standard deviation of the block, to each variable belonging to the same block. As a result, each block will have variance equal to one after block scaling, but the ratio of the variance between any two variables inside a block is preserved. This is different compared to block autoscaling where each variable is first scaled to unit variance (autoscaling), and then each block is scaled to equal variance. As a result each block will have unit variance and each variable inside a block will have the same variance equal to $1/n_{block}$, where n_{block} is the number of variables in a given block.

Three regions were defined for block scaling, the first one from 1.23 ppm to 3.14 ppm, the second from 3.15 ppm to 5.32ppm and the last one from 5.33 ppm to 7.93 ppm. Results of preprocessing for ^{1}H-NMR signals are reported in Figure 1.

Then, a PCA model, based on four principal components, chosen according to minimum RMSECV and explaining 86% of total variance, was considered and the extracted scores vectors were used to build the fused dataset.

2.4.2 EEM data set

Regarding the EEM data, some regions were cut because affected by noise and Rayleigh scattering. In particular, the excitation regions below 250 nm and the emission region over 450nm. A PARAFAC model, with no pretreatment applied, based on four factors constrained for non-negativity in the second and third mode, was built considering the best compromise between explained variance, core consistency and split half analysis [21] results.

2.4.3 HPLC-DAD data set

The data analysis related to the HPLC-DAD dataset is described in detail in our previous work [9]. To summarize, a multiset MCR-ALS approach was used to resolve the chromatographic runs. The whole elution domain was divided in eight elution windows, reported in Table 1, in order to better achieve a complete resolution of the system under investigation. The elution windows will be mentioned in the text with the name of a compound, for example Gallic Acid or p-Coumaric Acid, presents in that windows. Two windows, namely Windows1 and Windows2 were mathematically resolved by the MCR-ALS approach but not labeled as the others since no one of the compounds present was confirmed using reference standards. The MCR-ALS setting were the same for each elution window, namely the number of components was determined by PCA, and initial estimation of the pure spectral profile by Simplisma. Some constraints were applied to resolve rotational ambiguity: non-negativity in both concentration and spectra, and unimodality was applied to concentration profiles; finally selectivity or local rank were applied where we have further information about the system, for example for chromatographic runs of standards, or in the case in which the chromatographic profile of an experiment is known and some components are absent in some windows profiles.

From the resolved concentration profiles of each elution window the peaks areas of each chemical component were calculated by integration. Thus the thirty-nine peak areas obtained were analyzed by PCA, applying autoscaling, to have an overview of *Lambrusco* varieties separation by this technique.

2.5 Mid- Level Data fusion

A mid-level data fusion approach was adopted for the joint analysis of wine samples using the HPLC-DAD, ^{1}H-NMR and EEM information. In Figure 2 a schematization of the data fusion

framework is proposed. In this case, since the data sets are of different order (two and three way arrays), the low-level approach based on simple concatenation and scaling of the variables of different nature, was not considered.

As shown, the fused data set was assembled by using the four PCA scores from NMR data, the four PARAFAC samples loadings from EEM data and the areas of the thirty-nine resolved components by MCR of HPLC-DAD data. Both autoscaling and block-scaling (each data set corresponding to an analytical techniques was considered as a block) were tested as data preprocessing for the fused data set. Based on explorative PCA of the training set, autoscaling was finally selected as giving better categories separation. Then a PLS discriminant model was calculated on the fused data set.

2.6 Software

Multivariate Curve resolution was carried out by means of MCR-ALS GUI (http://www.ub.edu/mcr/web_mcr/mcrals.html). Preprocessings, fusion of data, classification rate (as well in CV) for PLS-DA and NPLS-DA were assessed by using homemade Matlab routines (Mathworks, MA, USA); PCA, PARAFAC, PLS-DA and NPLS-DA models were calculated by using the PLS Toolbox 6.5 (distributed by Eigenvector Research Inc., WA, USA).

3. Results and discussion

This section is articulated in four parts. In the first three, the description of exploratory analysis results, classification models and features extraction, are presented for the separate datasets, namely, "*^{1}H-NMR (NMR) dataset*", "*EEM dataset*" and "*HPLC-DAD dataset*".

In the last part, the fused dataset is considered and the application of the mid-level approach is described in detail.

3.1 ^{1}H-NMR dataset

^{1}H-NMR signals acquired on complex matrices are characterized by having a great number of peaks, high overlapping and a huge number of variables, which can difficultly be reduced by simple down sampling of the data because peaks width are very narrow. An analytical resolution of these signals by means of curve resolution techniques could be a difficult task. Signals present in NMR spectra of complex matrices, are referred to hundreds or even thousands of molecules and span in some cases the whole spectral region. In order to extract the majority of the information presents and to reduce the number of variables to be used for the data fusion step we decided to use NMR

spectra as a fingerprinting technique, hence, a PCA model was built on the training set. A four components PCA model, chosen on the basis of minimum cross validation error and able to describe the 86% of the total variance was calculated using the data preprocessing described in previous section 2.4.1.

As shown in Figure 3a, there are several part of the spectra having high loadings values for all components. As an example, the doublet at 1.4 ppm, attributable to CH_3 group of lactic acid, is positively associated to the first, second and fourth component and negatively to the third. On the other hand, several regions of the spectra cannot be associated to a single signal since many overlapped multiplets are present, e.g. anomeric region from 3 ppm to 5 ppm or aromatic region over 7 ppm. In these cases, an overall contribution, related to a class of compound, could be attributed to a given component, as in the case of aromatic region, in which polyphenol signals present high loadings for the third component, or the case of sugars signals present in the anomeric region and positively correlated with the fourth component. Considering the scores plot in Figure 3b it appears that the separation of the samples belonging to the three classes is not clear as they overlap. The third component, highly related to the aromatic region of the signal, seems to be able to distinguish, but not completely, at positive values *Salamino* and *Grasparossa* samples from *Sorbara* samples present at negative values.

The classification results obtained by PLS-DA on this data set are reported in Table 1 together with the classification models obtained for each separate data set. For all classification models presented in this work, the dataset was split in a training set of 45 samples and a test set of 13 samples, as described in the previous section 2.3.

The PLS-DA results show that few *Salamino* samples are misclassified for the training and one *Salamino* and one *Sorbara* samples for the test. The same *Sorbara* sample is misclassified for the test set in all models.

3.2 EEM dataset

The EEM data array (45 training samples, 161 emission wavelengths and 21 excitation wavelengths) was analyzed by means of PARAFAC. For the development of the most suitable model, second and third modes were constrained using non-negativity. A four factors model was chosen representing the best compromise between explained variance (99%), core consistency and split half analysis results.

The model obtained presents results that are in good agreement with works presented in literature [22]. In figure 4a, loadings for second (emission) and third (excitation) modes are reported. The first

resolved factor presents a broad maximum in excitation at 260 nm – 280 nm and a corresponding maximum in emission at 360 nm, very similar to the profile of vanillic acid. The second factor has fluorescent properties close to p-Coumaric acid, trans-resveratrol, trans-piceid and gentisic acid. The third factor well matches the wavelengths of catechin and epicatechin, having maximum in excitation at 280 nm and 320 nm in emission. The fourth factor cannot be directly attributed to a molecule or class of molecules and deepest investigations must be performed in order to characterize it. From the first mode loadings plot in Figure 4b, it is difficult to highlight a separation of samples belonging to the different classes. The first factor score values of *Sorbara* samples, on average, present higher values with respect to *Grasparossa* and *Salamino* samples, hence, higher concentration for vanilic acid or molecules having the same fluorescent properties.

A classification model based on NPLS-DA was built using nine latent variables, (lowest cross validation classification error rate), results are reported in Table 1. In this case, only one *Grasparossa* sample is misclassified for the training set. Concerning the test set, as for the PLS-DA model based on ^{1}H-NMR signals, the same *Sorbara* sample is misclassified; also one *Grasparossa* sample is misclassified.

3.3 HPLC-DAD dataset

Outputs from HPLC-DAD analysis are arranged in a cube in which the first mode is related to samples, the second to the UV spectra at different elution times and the third to chromatograms at different wavelengths. The multiset MCR approach can be applied to this kind of data unfolding the original three-way array in a multiset structure via column-wise augmentation. Having the dataset a very high chemical rank, hence, high rank for the resolution using MCR, we decided to split in eight windows the elution path, and then, to resolve separately each window applying constraints of non-negativity, unimodality and local rank as described in detail in reference [9]. A chromatogram, in which the elution windows are reported, is shown in Figure 5. Since each window presents different number of compounds, each MCR model was built considering different number of components. In Table 2 the number of components used to resolve each windows and the explained variances of all models are reported.

Thirty-nine areas, extracted from the eight MCR-ALS models, were merged together in a new block of variables. A PCA analysis was performed on this dataset on autoscaled variables (four components explain 60% of total variance). From the scores plot of the first two components, reported in Figure 6a, can be highlighted a rough separation of the three typologies of *Lambrusco* wines. In particular, all *Grasparossa* samples have negative scores values on PC2, *Salamino*

samples have positive values on PC2 and almost all *Sorbara* samples have negative scores values on PC1 and positive scores values on PC2; Salamino category overlaps with the other two. This result confirms the ability of phenolic compounds used as "fingerprinting technique" for the varietal discrimination of wines [22-25]. From the loadings plot, Figure 6b, is possible to point out that, having almost all the loadings a positive contribution for the first component, the separation between *Sorbara* samples and the other two varieties is manly due to a global lower concentration of phenolic compounds on this typology of *Lambrusco* wine. Regarding the second component, mainly responsible of the separation between *Grasparossa* and the other two categories, a complex loadings profile is present; hence, many compounds are present at different concentration in the two wines. For example, (+)-catechin and unassigned compounds SC3 (in the window of syringic and caffeic acids) and p-C5 (in the window of p-coumaric acid), having high positive loadings, are influent in giving positive scores values to *Sorbara* and most of *Salamino* samples, on the other hand, variables such as SC4 and Quer5 are important, having high negative loading, to characterize *Grasparossa* samples.

The peaks areas of resolved components by MCR models were used to build a classification model based on PLS-DA obtaining good discrimination (Table 1).

Summarizing, PLS-DA results are best for the model based on resolved areas from HPLC-DAD. In general the models performance is similar, as well as the models dimensionality (except for the EEM data).

3.4 Mid level data fusion

Once obtained, from the single data sets, the new blocks of variables (scores, peaks areas, etc.), the data-fusion step has to be completed merging them in a unique new block of variables. This step is the most crucial when dealing with mid-level data fusion. As in the case of low-level methodologies, the way in which the different variables are concatenated, normalized and scaled, the number and typology of extracted variables and their magnitude can bring to models that could be very different. In many cases, a "trials and errors" approach can be followed and the comparison of the obtained models can help the choice of the best one. Depending on the data analysis used in the first step of reduction, some consideration can be done in order to provide the best way to merge the new variables together without other "*a priori*" information. If the reduction step is obtained for example *via* variable selection, hence, the more representative of original variables are maintained, the quantity of information that each variable carries is far less than the information which is carried out by a latent variable extracted by means of a decomposition method. The intrinsic variance that

each variable bring, with respect to the original data, can be a suitable way to scale the variables in the new fused block. In this work, the extracted variables arise from decomposition methods; in the HPLC-DAD case each peak area is referred to a single compound, in the cases of EEM and ^{1}H-NMR to a group or class of compounds. The thirty-nine peaks areas from HPLC-DAD, the four scores from NMR dataset and the four PARAFAC loadings from EEM are able to describe most of the variance of the original data. For these reasons it was decided to circumscribe the scaling options for the merging procedure to few possibilities, namely, block scaling to equal variance for each block, block-autoscaling (the difference between these types of scaling is discussed in section 2.4.1) and autoscaling. Even if few minor differences were noticed in the PCA based explorative analysis, with respect to other scaling procedures, the classification model obtained on fused dataset using autoscaling gave better results.

Indirectly, using autoscaling as merging strategy, a higher importance is given to the block of variables more numerous, hence, the HPLC-DAD part. Being the HPLC-DAD analysis the only one that is referred to analytically resolved concentration, attributable to individual compounds, was decided to accept the fact that the global weight of these variables was higher if compared to the others extracted from NMR and EEM datasets.

In Figure 7a the scores plot of the first two components of the PCA model (three principal component chosen by minimum CV error, 51% variance explained) built using the autoscaled fused dataset are reported. A similar clustering of the three categories with respect to the one observed in the scores plot reported in Figure 6a, related to HPLC-DAD dataset, is observed. In particular, *Grasparossa* samples have negative scores values on the second component, *Salamino* samples have positive scores values on PC2 and *Sorbara* samples have negative scores values on PC1 and positive scores values on PC2. Some improvements can be noticed in the separation of samples belonging to different classes, as matter of fact, no one of the *Sorbara* samples is confused within the Salamino cluster in the first quarter. In Figure 7b the first two components loadings plot is reported and highlights that all variables contribute to the position in the scores space of the samples. In particular, the first and second factors from the EEM dataset seem to be most of all responsible of the improvement in the separation of *Salamino* and *Sorbara* samples, having high negative weight for second component, while the fourth EEM factor seems characteristic for *Grasparossa* class. Almost all the variables arising from the HPLC-DAD dataset are located at positive values for the first component. The third component extracted from the NMR dataset show positive loadings in both PCs contributing to separation of Salamino samples from Sorbara ones while the second component seems characteristic of *Grasparossa* samples.

A PLS-DA model, based on five latent variables chosen accordingly to minimum cross validation classification errors, was built using the fused dataset and the same scaling procedure adopted for the explorative PCA analysis. Results of the model, reported in Table 3, confirm the improvement with respect to the models obtained from the separate datasets. Figure 8a shows the scores plot of the first three latent variables and highlights the good separation among the three varieties of *Lambrusco* wines. In particular, all *Salamino* samples are placed at negative values for the first and second latent variable, *Sorbara* samples are situated at positive values for the third latent variable and close to zero for the others, whilst *Grasparossa* samples present high positive values for the second latent variable.

The close examination of the PLS weights plot, reported in Figure 8b, allows identifying which are the variables that most of all are responsible of the separation of the samples belonging to different classes and on the other hand to evaluate the correlation among them.

Considering *Sorbara* samples, many variables seem to be relevant, in particular the first and second factors from EEM dataset, p-Coumaric acid, Caffeyc acid and two unassigned compounds belonging to the window2 cluster and myrecitin elution window. As described above, the second EEM component extracted by the PARAFAC model presents fluorescent properties compatible with several compounds such as, trans-resveratrol, trans-piceid, gentisic acid and p-Coumaric acid, which was one of the resolved compound by MCR analysis of HPLC-DAD data and stays very close to the second EEM factor in the scores space obtained by the PLS-DA model. The first factor from EEM is placed close to a group of unassigned compounds resolved starting from HPLC-DAD part, hence, future investigations will be oriented to the assessment of the presence of vanillic acid in that cluster of compounds. Also the first component extracted from the ^{1}H-NMR dataset, which brings mainly the information of the anomeric region of the NMR signal, seems to give a relevant contribution for the discrimination of *Sorbara* samples.

The third component from the ^{1}H-NMR signals elaboration, mainly associated with the phenolic signal region, and a wide group of variables obtained via HPLC-DAD analysis, such as Gallic acid, myrecitin, quercytin etc., are relevant for the classification of Salamino samples.

Grasparossa samples, on the other hand, are described mostly by the fourth component arising from the EEM data and syringic acid and other molecules obtained from the analysis of HPLC-DAD data.

Finally, on the basis of the PLS weights plot (Figure 8b) and the VIP scores (plot not shown for sake of brevity) in order to evaluate if using a reduced number of features from the HPLC-DAD data set, which requires longer acquisition time and higher computational complexity in the

resolution step, could be sufficient to achieve a satisfactory classification model, a data fused model was considered including only the concentration of the compounds resolved in the window of syringic and caffeic acids together with the four PCA scores from NMR data and the four PARAFAC samples loadings from EEM. The results were quite satisfactory, showing the same classification performance on the test set and two classification errors in the training set, one *Grasparossa* and one *Sorbara* sample, respectively.

4 Conclusions

In food analysis the presence of interferences, uncontrolled variability due to biological and chemical-physical transformation combined with the development of new high-throughput analytical platforms make mandatory the use of chemometrics in order to unveil the hidden information able to characterize, discriminate, evaluate the products under investigation. The overall approach proposed, based on a mid-level data fusion application, allowed to deal with data of different nature and dimensionality (matrices and tensors) in a global way. In this respect, it suggests a suitable strategy generally applicable, whose potentiality may be also better appreciated in more complex cases with respect to the actual one. The possibility to treat jointly several analytical data was useful to develop a classification model for the three P.D.O. *Lambrusco* wines and to understand the correlations between the features extracted from the original separate datasets. In particular, we have seen that part of information extracted by complete resolution of several phenolic compounds through MCR analysis of HPLC-DAD data set can be recovered by faster approaches, such as EEM or NMR.

When the *a priori* knowledge about the investigated samples is not complete, the data fusion approach has proven to be a powerful tool for features/information discovery, since complementary knowledge can be point out when the problem is tackled in a comprehensive way.

Moreover, better results in classification were obtained when data fusion approach was used with respect to the analysis of separate datasets.

However, it may be that especially in the case of three way data structure in the compression step not all the information can be properly recovered. Thus, a further step will be the use of techniques that allow to extract common latent structure from the shared mode of the different source of data (in this case of study the sample mode) such as Coupled Matrix Tensor Factorization (CMTF) [26].

Acknowledgements

This work was supported by the AGER, *Agroalimentare e Ricerca*, cooperative project between grant-making foundations under the section "wine growing and producing" project. New analytical methodologies for varietal and geographical traceability of enological products; contract n. 2011-0285. We are also grateful to *Consorzio Marchio Storico Lambruschi Modenesi* for use of their facilities during sampling procedures.

A special thank to Prof. Rasmus Bro for availability of EEM instrumentation and assistance in elaboration of EEM data.

References

[1] J. Forshed, H. Idborg, S. P. Jacobsson, Evaluation of different techniques for data fusion of LC/MS and 1H-NMR, Chemometrics and Intelligent Laboratory Systems, (2007) 85, 102–109

[2] L. Vera, L. Aceña, J. Guasch, R. Boqué, M. Mestres, O. Busto, Discrimination and sensory description of beers through data fusion, Talanta, (2011) 87,136-142

[3] I.V. Mechelen, A.K. Smilde, A generic linked-mode decomposition model for data fusion, Chemometrics and Intelligent Laboratory Systems (2010) 104, 83–94.

[4] Fernández, C., Pilar Callao, M., Soledad Larrechi, M., UV-visible-DAD and 1H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS, Talanta (2013) 117, 75-80.

[5] Silvestri M., Bertacchini L., Durante C., Marchetti A., Salvatore E., Cocchi M, Application of data fusion techniques to direct geographical traceability indicators, Analytica Chimica Acta, (2013) 769, 1-9.

[6] Smolinska, A., Blanchet, L., Coulier, L., Ampt, K.A.M., Luider, T., Hintzen, R.Q., Wijmenga, S.S., Buydens, L.M.C., Interpretation and visualization of non-linear data fusion in kernel space: Study on metabolomic characterization of progression of multiple sclerosis, PLoS ONE 7 (2012) 6, art. no. e38163.

[7] I. Stanimirova, C. Boucon, B. Walczak Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction, Talanta (2011) 83, 1239–1246.

[8] Y. Lin, S.D. Brown, "Wavelet multiscale regression from the perspective of data fusion: new conceptual approaches". Analytical Bioanalytical Chemistry, 380 (2004) 445-452.

[9] E. Salvatore, M. Cocchi, A. Marchetti, F. Marini, A. de Juan, "Determination of phenolic compounds and authentication of PDO Lambrusco wines by HPLC-DAD and chemometric techniques". Analytica Chimica Acta (2013) 761, 34-45

[10] G. Papotti, D. Bertelli, R. Graziosi, M. Silvestri, L. Bertacchini, C.Durante, M. Plessi. Application of One- and Two-Dimensional NMR Spectroscopy for the Characterization of Protected Designation of Origin Lambrusco Wines of Modena, Journal of Agricultural and Food Chemistry, (2013) 61(8), 1741-1746

[11] S. Meiboom, D. Gill, Modified Spin- Echo Method for Measuring Nuclear Relaxation Times, Review of Scientific Instruments, (1958) 29, 688−691

[12] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems, (1997) 38(2), 149-171

[13] J. Jaumot , R. Gargallo , de Juan A, R. Tauler, A graphical user-friendly interface for MCR - ALS: a new tool for multivariate curve resolution in MATLAB, Chemometrics and Intelligent Laboratory Systems, (2005) 76(1), 101-110

[14] MCR-ALS Toolbox (http://www.ub.edu/mcr/web_mcr/down_mcrt.html)

[15] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Analytical Methods, (2013) 5 , 3790-3798

[16] R. Bro, Multiway calibration. Multi-linear PLS, Journal of Chemometrics, (1996) 10, 47–61.

[17] S. De Jong, Regression coefficients in multilinear PLS, Journal of Chemometrics (1998) 12, 77–81.

[18] R. D. Snee, Validation of Regression Models: Methods and Examples, Technometrics, (1977), 19(4), 415-428

[19] G. Tomasi, F. Savorani, S.B. Engelsen, icoshift: An effective tool for the alignment of chromatographic data, Journal of Chromatography A, (2011) 1218(43), 7832-7840

[20] 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(1), 21-30.

[21] C. M. Andersen, R. Bro, Practical aspects of PARAFAC modeling of fluorescence excitation-emission data, Journal of Chemometrics, (2003) 17, 200–215.

[22] D. Airado-Rodríguez, T. Galeano-Díaz, I. Durán-Merás, J. P. Wold, Usefulness of Fluorescence Excitation−Emission Matrices in Combination with PARAFAC, as Fingerprints of Red Wines, Journal of Agricultural and Food Chemistry, (2009) 57(5), 1711–1720.

[23] M. García-Marino, J.M. Hernández-Hierro, C. Santos-Buelga, J.C. Rivas-Gonzalo, M.T. Escribano-Bailón, Multivariate analysis of the polyphenol composition of Tempranillo and Graciano red wines, Talanta, (2011) 85 , 2060–2066.

[24] A. Andreu-Navarro, P. Russo, M.P. Aguilar-Caballos, J.M. Fernández-Romero, A. Gómez-Hens, Usefulness of terbium-sensitised luminescence detection for the chemometric classification of wines by their content in phenolic compounds, Food Chemistry, (2011) 124(4), 1753–1759

[25] R. D. Di Paola-Naranjo, M.V. Baroni, N.S. Podio, H.R. Rubinstein, M.P. Fabiani, R.G. Badini, M. Inga, H.A. Ostera, M. Cagnoni, E. Gallegos, E. Gautier, P. Peral-García, J. Hoogewerff, D.A. Wunderlin, Fingerprints for Main Varieties of Argentinean Wines: Terroir Differentiation by Inorganic, Organic, and Stable Isotopic Analyses Coupled to Chemometrics, Journal of Agricultural and Food Chemistry, (2011) 59(14), 7854–7865

[26] E. Acar, M. A. Rasmussen, F. Savorani, T. Næs, R. Bro, Understanding data fusion within the framework of coupled matrix and tensor factorizations, Chemometrics and Intelligent Laboratory Systems, (2013) 129, 53-63.

Captions of Figures

Figure 1: Effects of preprocessing on ^{1}H-NMR signals

Figure 2: Data analysis flow and schematization of the mid-level data fusion process

Figure 3: a) Loadings plot of the PCA model built using *" ^{1}H-NMR dataset "*, b) Scores plot of the PCA model built using *" ^{1}H-NMR dataset "*

Figure 4: a) Loadings of second and third modes of the PARAFAC model built using *"EEM dataset"*, b) Loadings of the first (samples) mode of the PARAFAC model built using *"EEM dataset"*

Figure 5: HPLC-DAD TIC chromatogram of a sample: the sub division in elution windows is shown

Figure 6: Scores plot (a) and loadings plot (b) of the PCA model built using *"HPLC-DAD dataset"*. Variables labels in loadings plot as reported in Table 1.

Figure 7: a) Scores plot of the PCA model built using *"Fused dataset"*, b) Loadings plot of the PCA model built using *"Fused dataset"*

Figure 8: a) Scores plot of the PLS-DA model built using *"Fused dataset"*, b) Weights plot of the PLS-DA model built using *"Fused dataset"*

Table(s)

Table 1: Results of PLS discriminant analysis for the single data sets

Dataset	Model	Latent Variables	N° of missclassified samples in CV			% Correct classification rate in CV			N° of missclassified Test set samples			% Correct classification rate Test set			N° of missclassified Training set samples			% Correct classification rate Training set		
			Gra	*Sal*	*Sor*	*Gra*	*Sal*	*Sor*	*Gra*	*Sal*	*Sor*	*Gra*	*Sal*	*Sor*	*Gra*	*Sal*	*Sor*	*Gra*	*Sal*	*Sor*
HPLC-DAD	*PLS-DA*	3	0	2	0	100	88	100	0	0	1	100	100	75	0	1	0	100	94	100
^{1}H-NMR	*PLS-DA*	4	2	5	0	86	69	100	0	1	1	100	75	75	0	3	0	100	81	100
EEM	*NPLS-DA*	9	3	3	1	79	81	93	1	0	1	80	100	75	1	0	0	93	100	100

Table(s)

Table 2: Results from MCR analysis of HPLC-DAD data set

Eluition Windows	Window name in figures	Resolved MCR components	Explained Variance %
Gallic Acid	GA	3	99.2
(+)-Catechin	C	3	98.0
Window 1	Window1	3	97.1
Caffeic and Syringic Acids	CS	4	98.8
p-Coumaric Acid	p-C	7	99.3
Window 2	Window2	9	98.6
Myricetin	Myr	4	99.5
Quercetin	Quer	6	99.7

Table(s)

Table 3: Results of PLS discirminant analysis for the fused data set

Model	Latent Variables	N° of missclassified samples in CV			% Correct classification rate in CV			N° of missclassified Test set samples			% Correct classification rate Test set			N° of missclassified Training set samples			% Correct classification rate Training set		
		Gra	***Sal***	***Sor***	***Gra***	***Sal***	***Sor***	***Gra***	***Sal***	***Sor***	***Gra***	***Sal***	***Sor***	***Gra***	***Sal***	***Sor***	***Gra***	***Sal***	***Sor***
PLS-DA	**5**	0	2	0	100	88	100	0	0	1	100	100	75	0	0	0	100	100	100

Figure1
Click here to download high resolution image

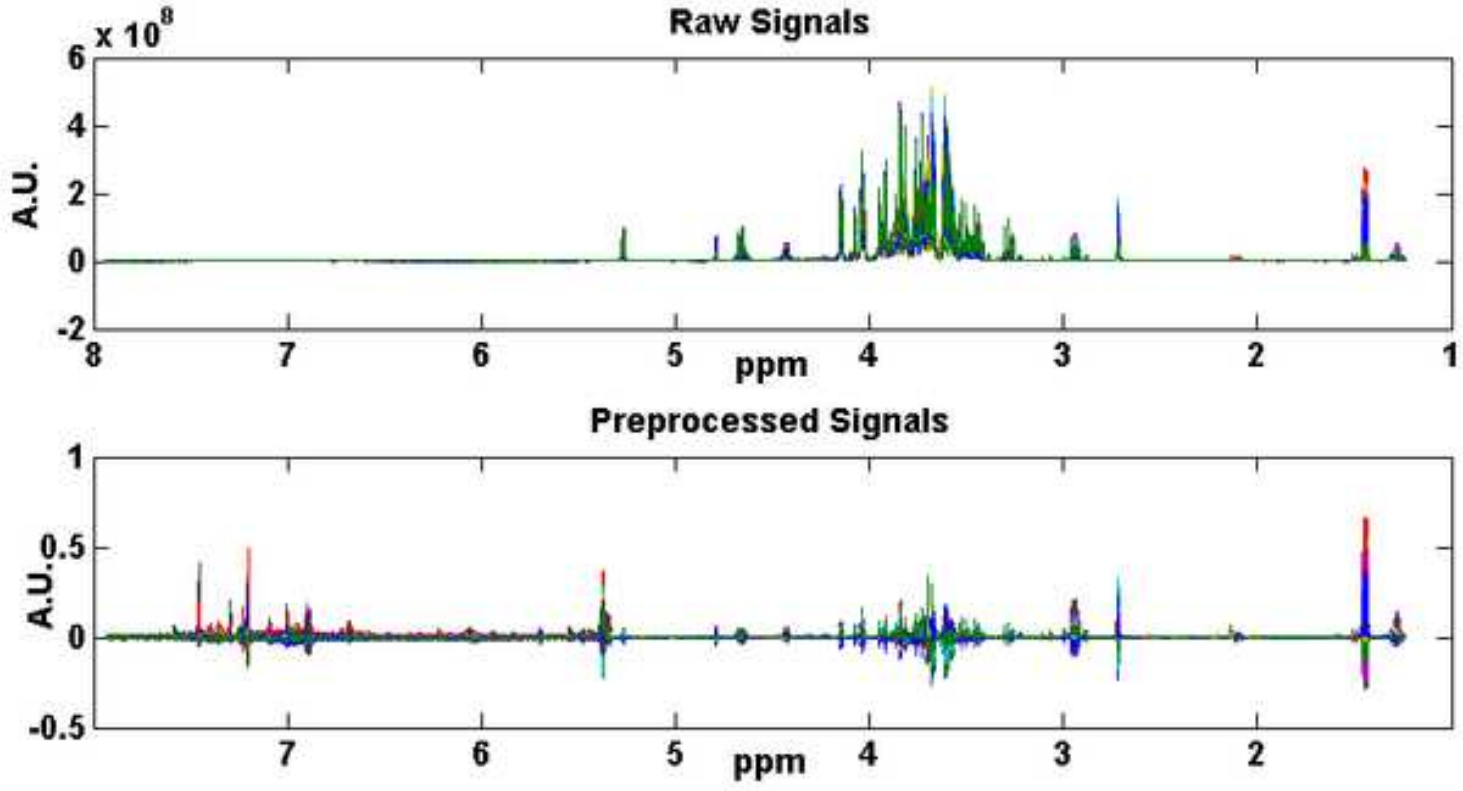


Figure2 coloured for web

Click here to download high resolution image

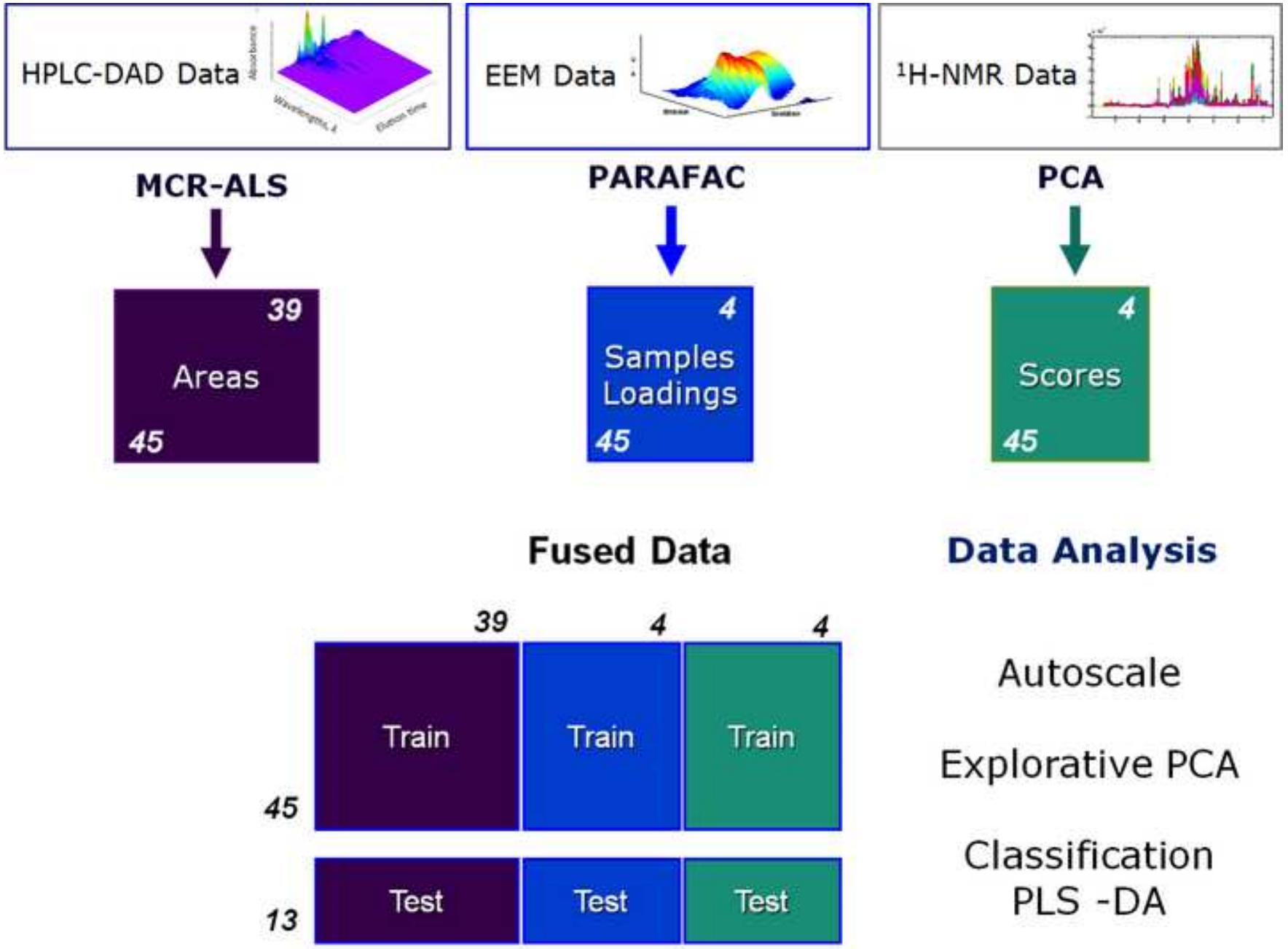


Figure3 coloured for web

Click here to download high resolution image

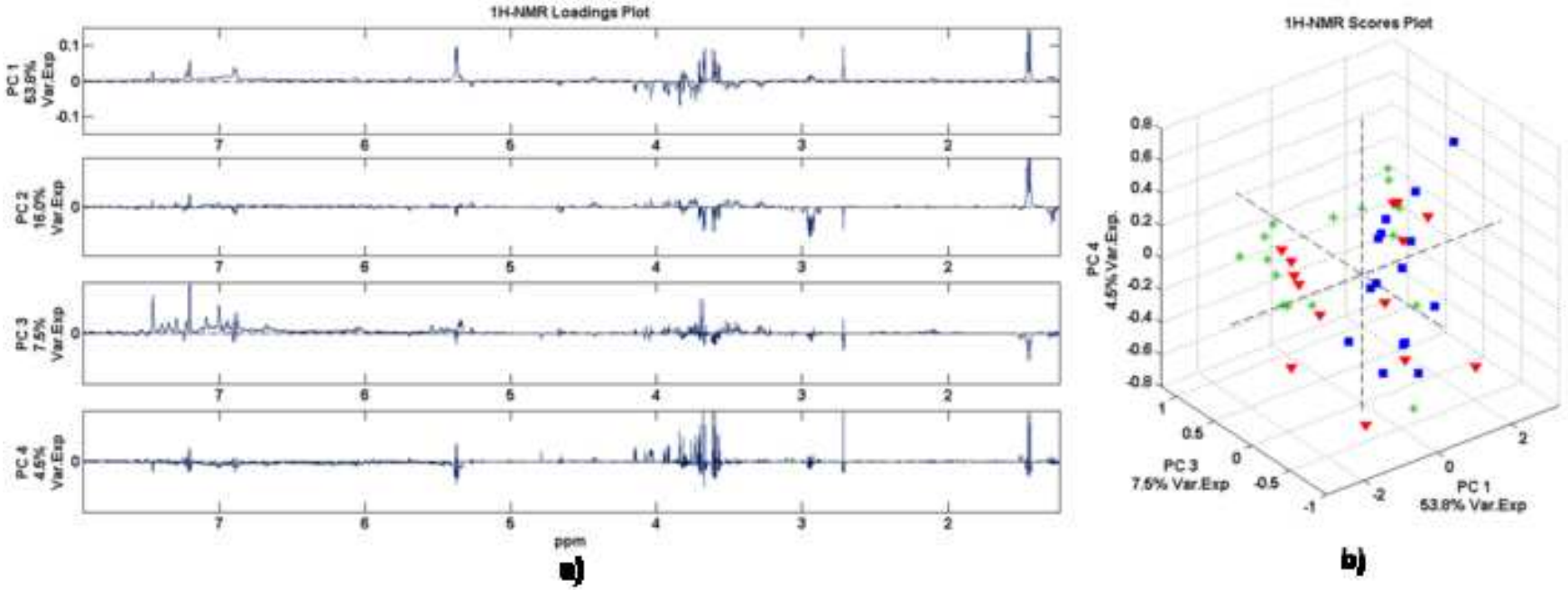
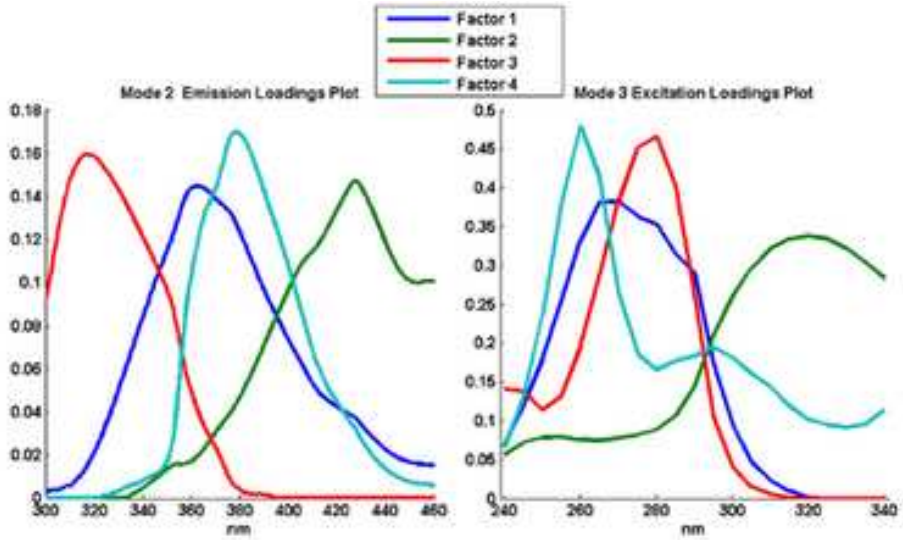
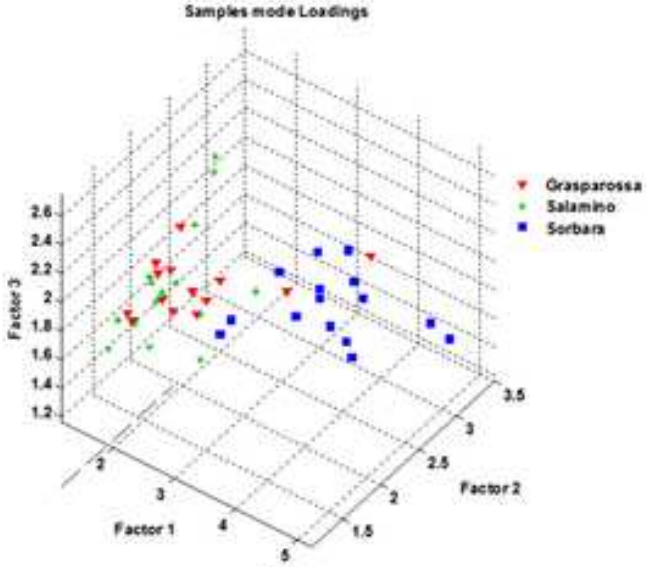


Figure4 coloured for web
Click here to download high resolution image



a)



b)

Figure5 coloured for web
Click here to download high resolution image

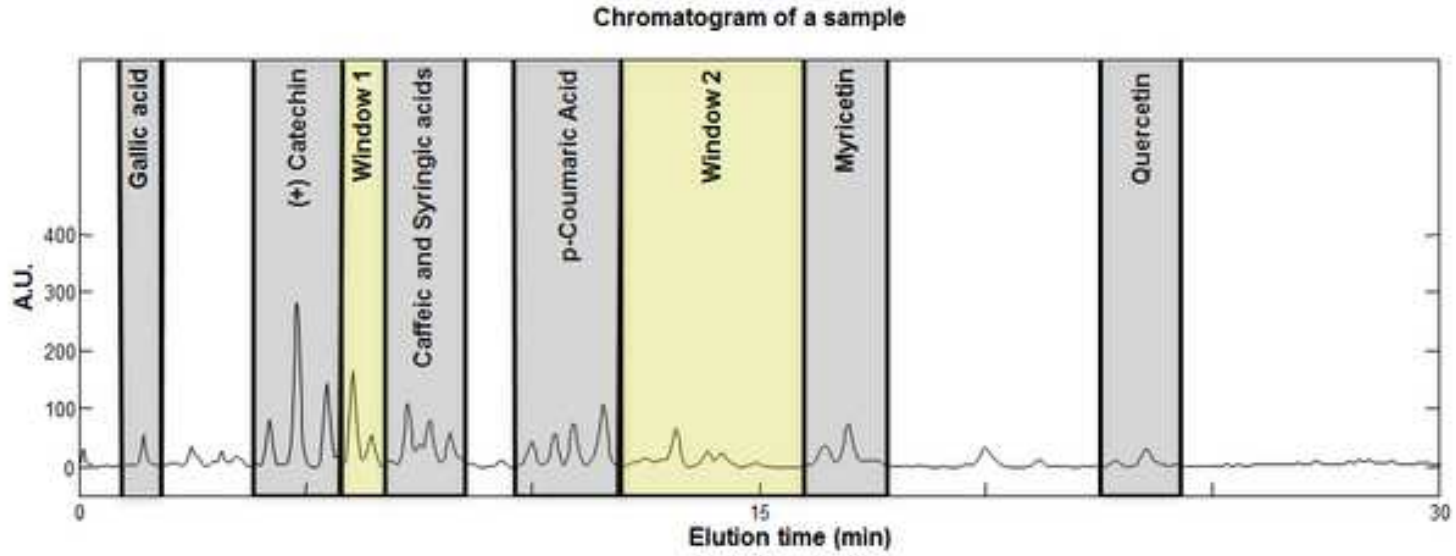


Figure6 coloured for web

Click here to download high resolution image

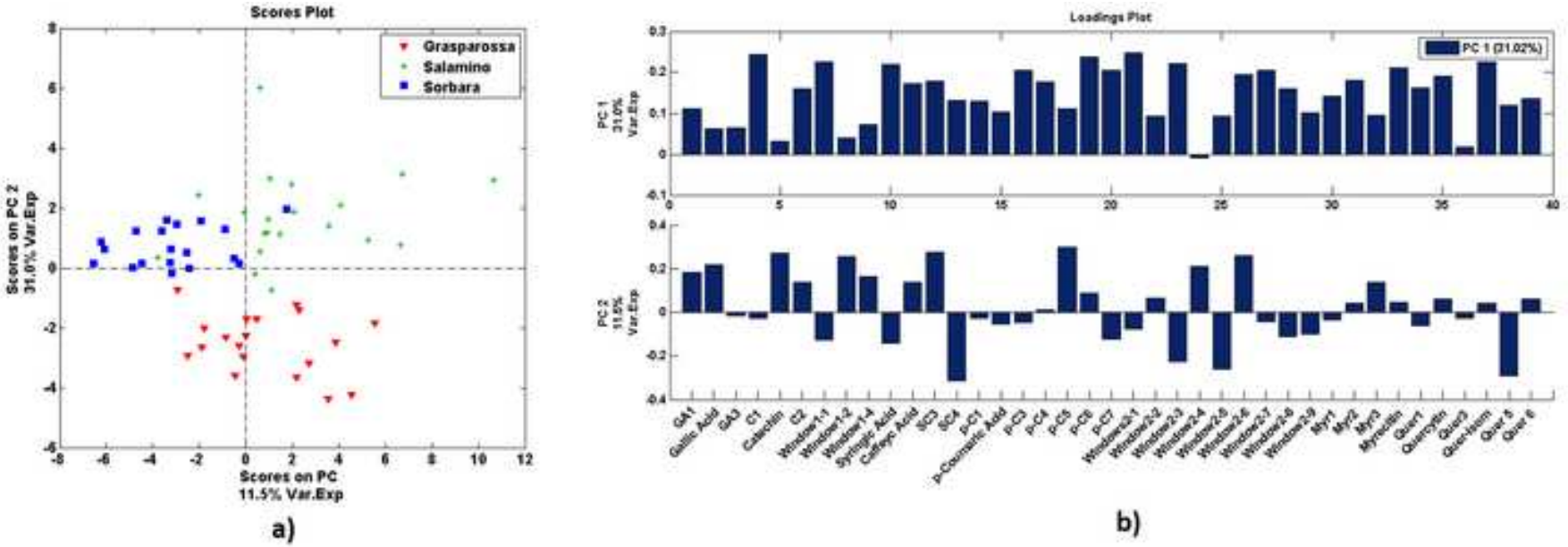
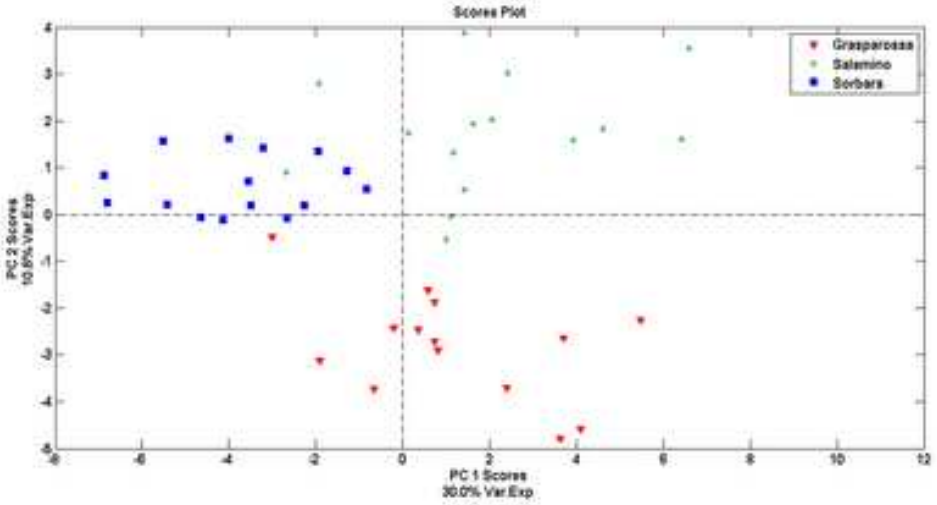
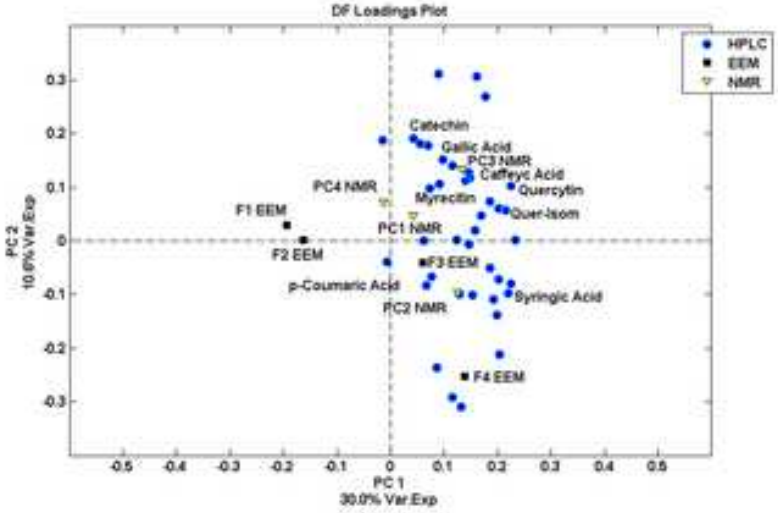


Figure7 coloured for web
Click here to download high resolution image

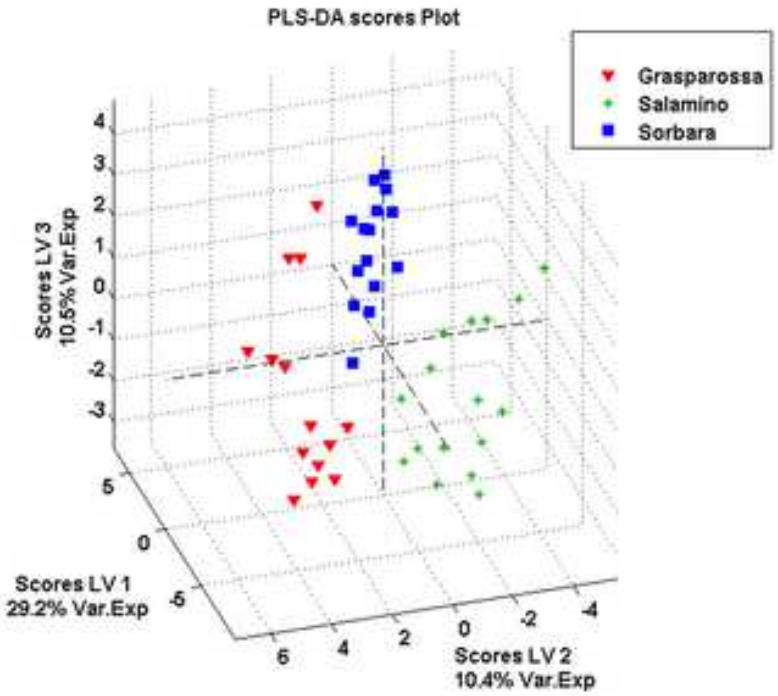


a)

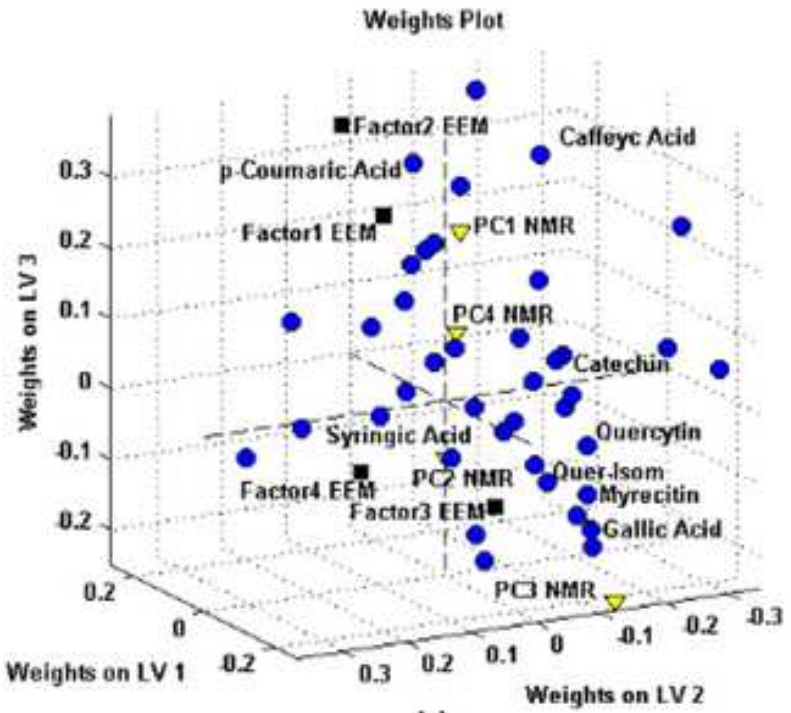


b)

Figure8 coloured for web
Click here to download high resolution image



a)



b)