4. RESULTS

4.1 Mathematical resolution of chromatographic problems

4.1.1 Peaks coelution

The problem of coelution regards the data set Lambrusco, the data were collected with the specific purpose of reducing the time analysis, this at the expense of resolution. As a consequence many coeluted peaks are obtained and these are solved with chemometrics tool a posteriori. MCR is applied for the resolution, and the comparison with the same data solved with PARAFAC models is also reported.

Chromatographic analysis of the Lambrusco wines mostly focused on the extraction of phenolic compounds, in particular 7 of these were quantified, by using mixtures reference standard compounds, and these are Gallic acid, (+)-Catechin, Syringic Acid, Caffeic Acid, p-Coumaric acid, Vanillin, Myrecetin and Quercetin. These are common compounds which belongs to the categories of phenolic acids, flavonoids, in Figure 1 are reported the structure of the target compounds. Recently many authors in enology have studied this class of molecules [1-4]. At the beginning the interest was towards their antioxidant properties which makes the red wine to assume protection role for cardiovascular disease, phenols are present in wine due to their natural content in grapes or as a result of contact with wood during the aging process. Most recently these have been investigated since they seem to be the species of major interest in the chemotaxonomic differentiation of vegetables and they have also the advantages of being widespread among a wide number of plants. Some phenolic groups are powerful tools for pattern recognition in different fruits, for example flavonoids, which are secondary metabolites ubiquitously in plants where they function as antioxidant. Indeed, evidence revealing that phenols may constitute a fingerprint to differentiate wines from pure varieties has been found. Moreover many studies regarding the evolution of phenol during common process in enology have been carried out [5]. Among the most important phenol in wine characterization must be also named other classes such as stilbenoids [6]

(e.g. resveratrol [7]), anthocyanins [8] (pigments which gives the red purple to plants and fruits: malvidin, delphinidin), tannins, cynnamic acids, flavonoids, dihydroflavonols and flavanols.

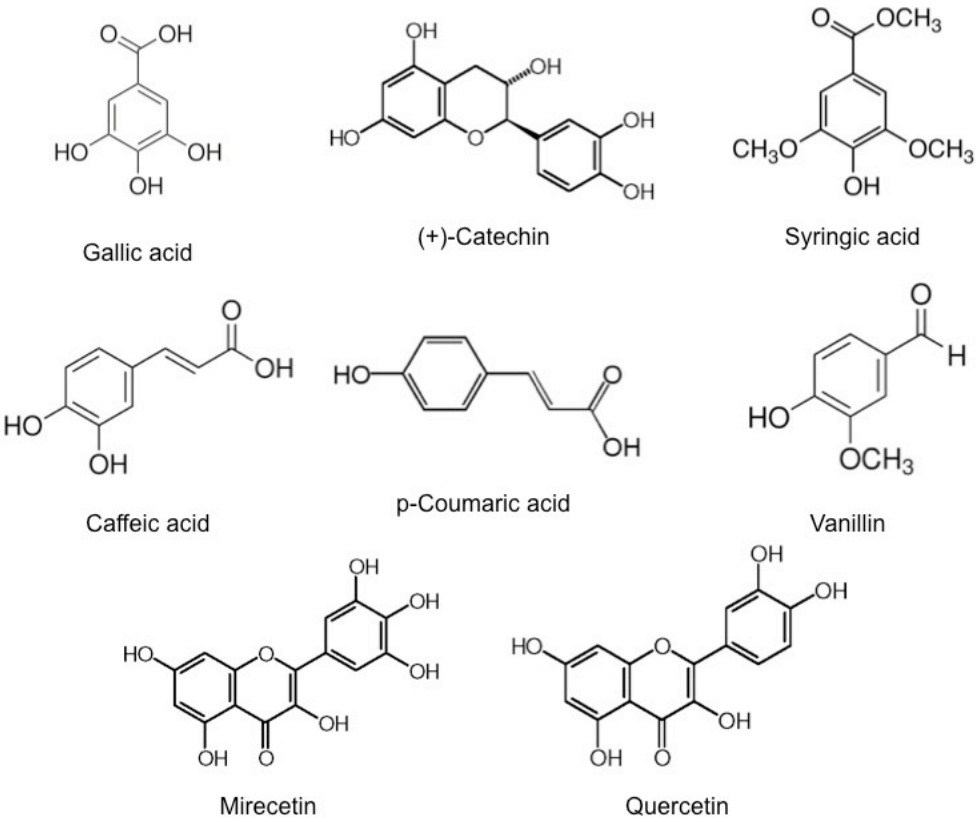


Figure 1: Structures of the phenolic compounds of interest.

A part of the total content of phenolic fraction regards also glycosylated phenols, which are a consistent percent. With respect to the total wine composition, the phenols represent a minor fraction with a concentration range of 200-500 mgL^{-1} for red wine, which presents higher concentrations respect to the white one, due to the different vinification process [9].

To achieve an effective resolution by MCR approach, the chromatograms were divided in six different retention time regions, namely six elution windows, and from each of this a data set or a multiset is obtained and these blocks are analyzed separately. The six windows were defined considering

the retention time (the location) of each of the quantified compounds in the chromatographic run. To set the six elution windows in each run, an homemade MATLAB™ routine has been written, the routine performs the following steps: i) it finds the peak maximum of each target compound in the sample chromatogram, to check the correspondence with the standard compound, namely the UV spectrum corresponding to a peak maximum is compared with the spectrum of the reference standard compound; ii) it builds the relative window a defined number of time points on both sides of the maximum found (for 50 points it is equivalent to 3-4 minutes of elution time).

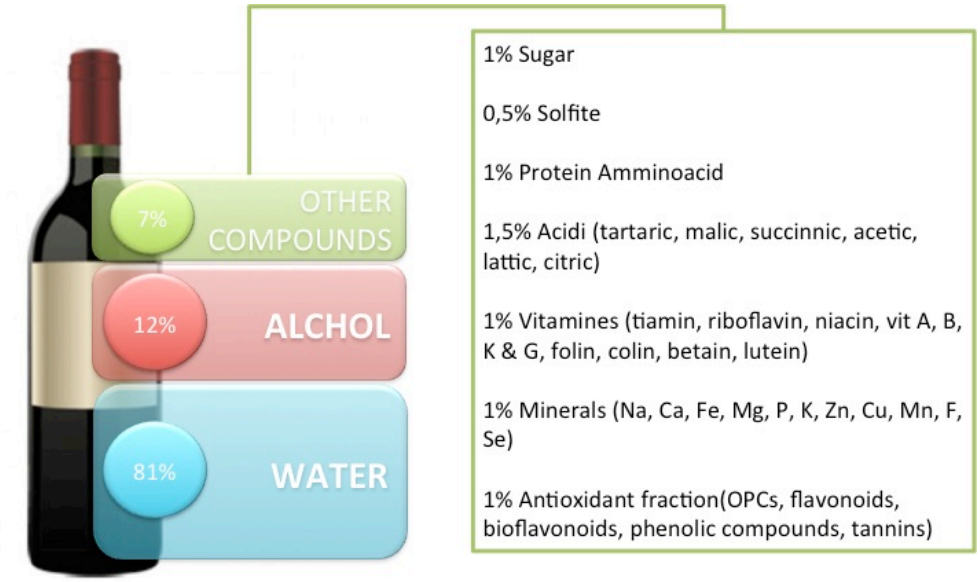


Figure 2: Percent composition of wine constituents, with a focus on the minor ones.

This way to build the elution windows, i.e., looking at the spectral similarity of the sample spectra with the standard spectrum, is more reliable than just using the retention time of the target compound, since shifts in elution time can occur among standard and sample chromatographic runs. The retention time of the standard has been taken as the reference to build elution windows only when the spectral information was not conclusive, i.e., when the sought compound in the sample was present at very low content or

found to be absent after the resolution analysis. Each elution window in a sample run contains a peaks cluster with, at least, one analyte and some co-eluted compounds. The six multiset structures formed by standards and sample runs are named after the phenolic compounds eluted in them, as follows: 1. Gallic acid, 2. Catechin, 3. Caffeic and Syringic acids, 4. Coumaric and Vanillin, 5. Myrecetin and 6. Quercetin. Figure 3 reports a mean chromatogram of sample 0012GRA_X compared with a chromatographic run, containing a mixture of standards of the seven compounds analyzed, the elution windows considered with the relative names are highlighted by vertical lines.

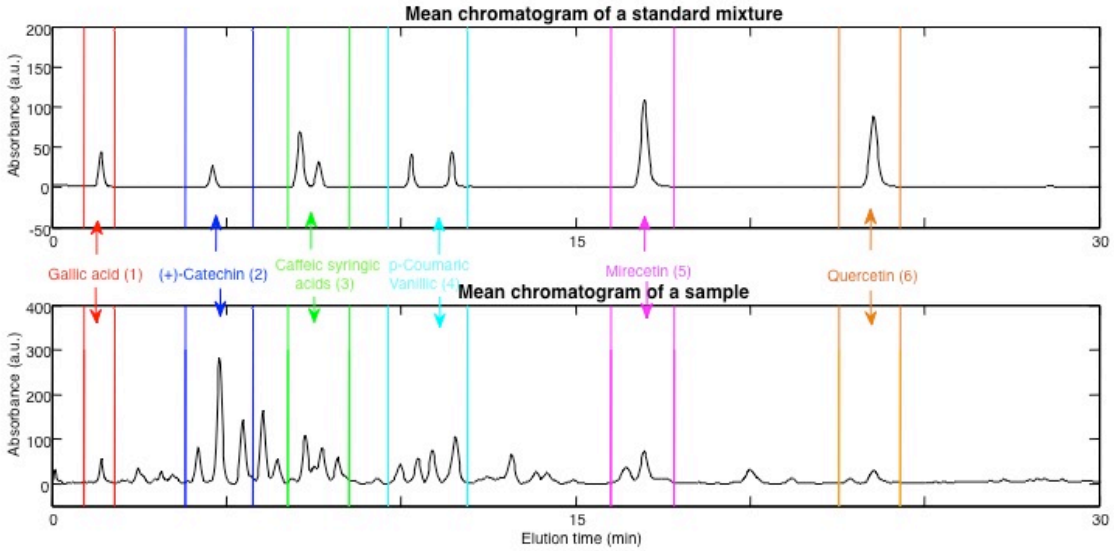


Figure 3: The six elution windows are shown: in the mean chromatogram of the standard mixtures and in the mean chromatogram of sample 0012_GRA; [1]Gallic acid, [2](+)-Catechin, [3]Syringic Caffeic acid, [4]Vanillin p-Coumaric acid, [5]Myrecetin [6]Quercetin.

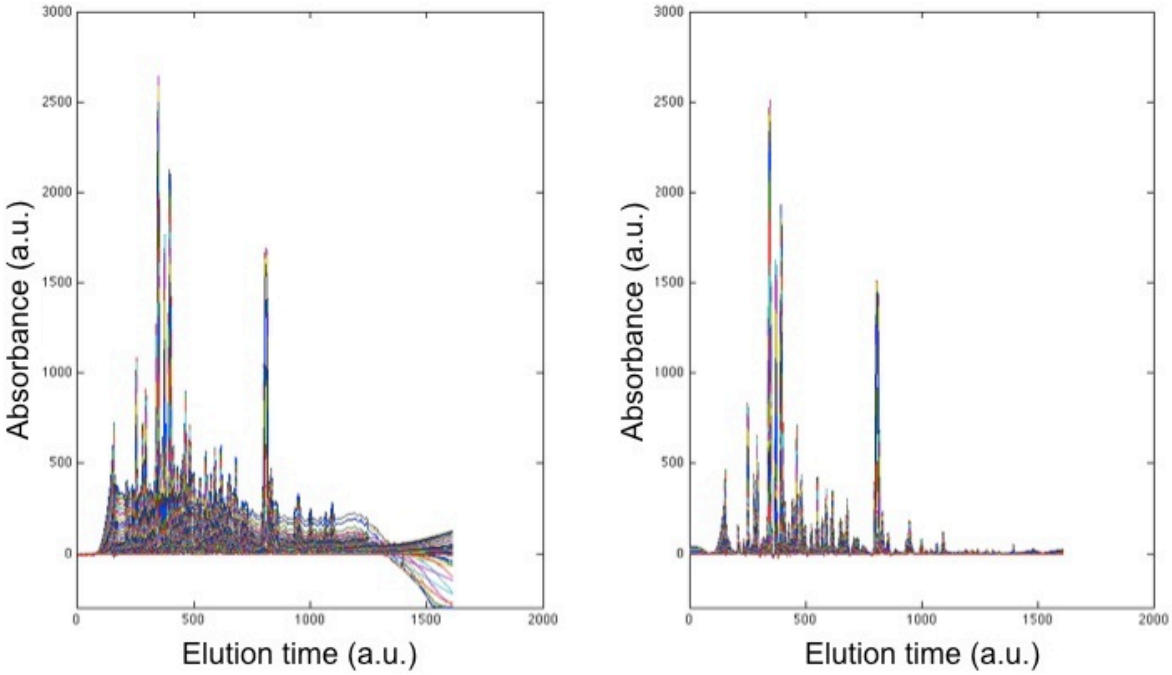


Figure 4: Elimination of Background Signal (EBS), application of the algorithm on a real sample of Lambrusco data set.

These data are analyzed by liquid chromatography with mobile phase acetonitrile-water, as described in Chapter 3. As a consequence the data obtained needs to be corrected with EBS (described in Section 2.4.2) to suppress the baseline contribution due to variations of mobile phase absorbance. In Figure 4, it is reported an example of how this kind of data pretreatment change the signal.

4.1.1.1 MCR-ALS results

Each multiset is organized in the same way, namely appending the matrices one on the top of the other, in the first part are positioned the experiments regarding the standard compounds, and then all the samples analyzed respecting in each multiset the same alphabetic order of the sample and the separation among the three categories as reported in Figure 5. The multisets were separately analyzed by MCR-ALS and the results obtained consist of the resolved elution profiles and related pure spectra. The MCR model quality parameters and the number of components, corresponding to the

compounds resolved in each multiset are reported in Table 1. For all MCR models, the explained variance is close to 99 % or higher, which indicates the validity of the results. In all cases, the resolved elution profiles and related pure spectra were consistent with the information in the raw data. Increasing the number of components in the different multisets did not lead to better fit or more interpretable results. Below, the resolution of two of the multisets analyzed is described in more detail since the optimization involved the use of some constraints and allows discussion of common problems encountered in the analysis of this kind of chromatographic data.

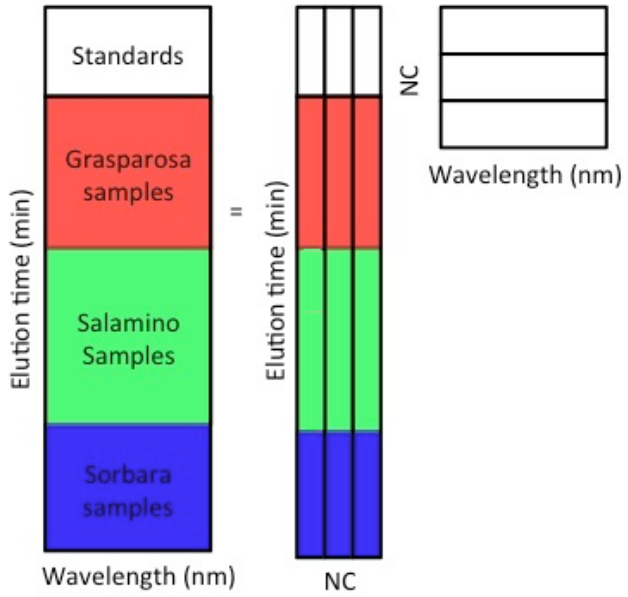


Figure 5: Scheme of the MCR bilinear model representing the multiset structure built using all the chromatographic experiments of the Lambrusco wine samples.

Multiset	Number of component	% Explained variance	% Lack of fit exp
Gallic Acid	3	99.2	9.1
(+)-Catechin	3	98.9	10.2
Syringic, Caffeic acids	4	98.8	11.0
p-Coumaric, Vanillin	7	99.3	8.3
Myrecetin	4	99.5	7.0
Quercetin	6	99.7	5.1

Table 1: Resolution parameters of MCR models for the six multisets.

These multisets resolution, regards the elution windows number 4 and 6 of the phenolic compounds p-Coumaric acid-Vanillin and Quercetin, and are used as example to explain how to optimize the results obtained and interpret them.

p-Coumaric acid and Vanillin are examples of compounds that were very difficult to identify from the sole inspection of the spectral information in the raw data linked to sample chromatographic runs. This can be due to a high coelution of these two phenols with other compounds or to the fact that these compounds are very minor or absent in the samples analyzed. Knowing which the real situation is would be difficult if these chromatograms were analyzed without additional information. The use of chromatographic standard runs of both analytes in the multiset and the application of the constraint of correspondence among species in these standard runs with very well known composition helps to define the identity of these species and to model unequivocally the presence or absence of these compounds in the sample runs. A total of seven compounds were modeled in the multiset. The $\mathbf{S}^T$ matrix with the seven spectra is reported in Figure 6, including those of the two analytes. The **C** matrix of the whole multiset is shown in Figure 7, zooming on the sample 0019SAL_X, where it is clear that the p-Coumaric acid peak is present and perfectly resolved from the coeluting compounds, whereas Vanillin is detected in a very minor proportion and may likely be absent in the samples analyzed. Similar conclusions about this pair of compounds are obtained in the rest of sample runs analyzed. Once quantitative information is obtained, it will be seen whether Vanillin is a minor compound or whether it can be considered absent (if its concentration is below the detection limit).

In Quercetin multiset, the use of local rank information helps in solving the Quercetin peak from a strongly coeluted interference, despite the large spectra similarity between them and the minor presence of the second compound. The local rank constraint is applied only in some specific zones of the chromatogram, where the Quercetin peak is clearly separated from the neighbor interference. In this elution region, the interference is set to be

absent. It is important to say that, in a multiset structure, this local rank constraint has only been applied to some sample runs where this differentiation is clear. This strategy makes possible a perfect resolution of the two species, which otherwise would be seen as a unique component because of the high elution and spectral overlap in many runs. The $\mathbf{S}^T$ matrix in Figure 8 shows the compounds' spectra, where bold line and dashed line belong, respectively, to Quercetin and the interference. In Figure 9 the zoom on the sample 0006GRA in the **C** matrix shows the elution region in the chromatogram where the absence of the interference has been set as a local rank constraint.

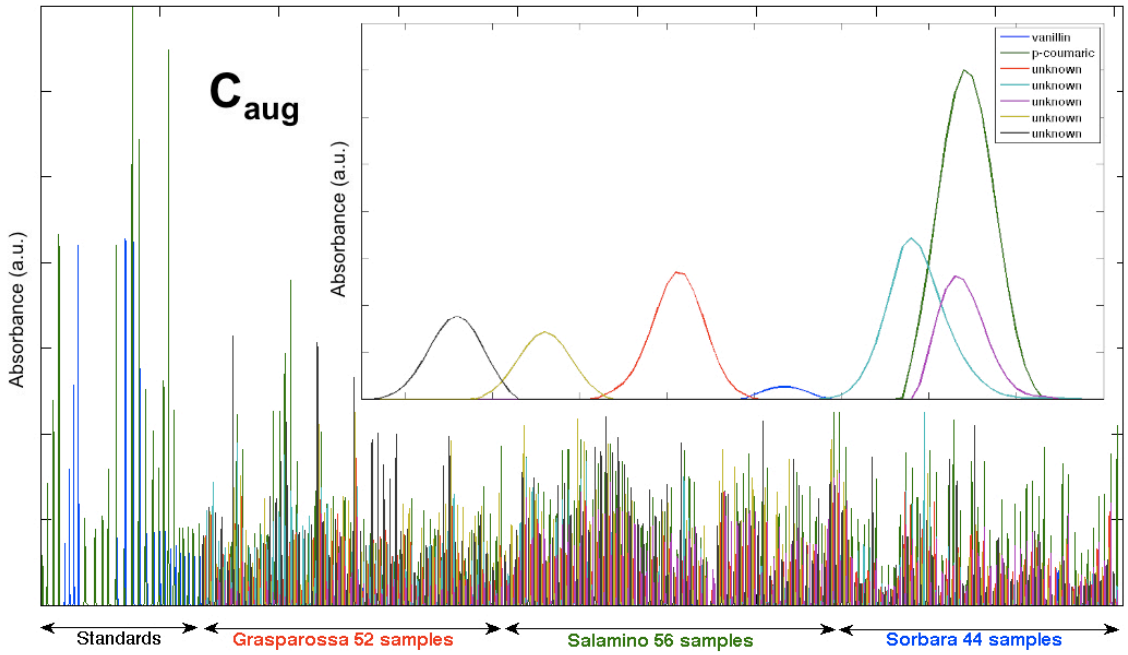


Figure 6: Matrix of the solved elution profiles for the p-Coumaric acid and Vanillin multiset; **C** profiles and the zoom of sample 0019_SAL, p-Coumaric (green line) and the Vanillin (blue line).

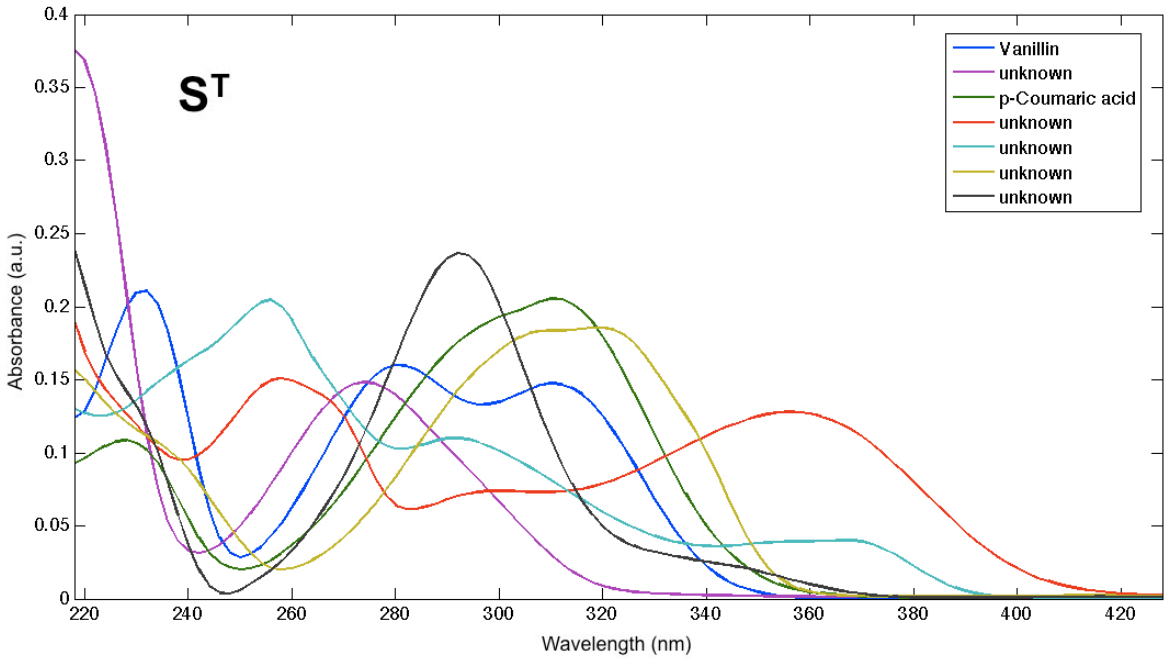


Figure 7: Matrix of the pure spectra $\mathbf{S}^T$ for the p-Coumaric and Vanillin multiset, p-Coumaric acid (green line) and Vanillin (blue line).

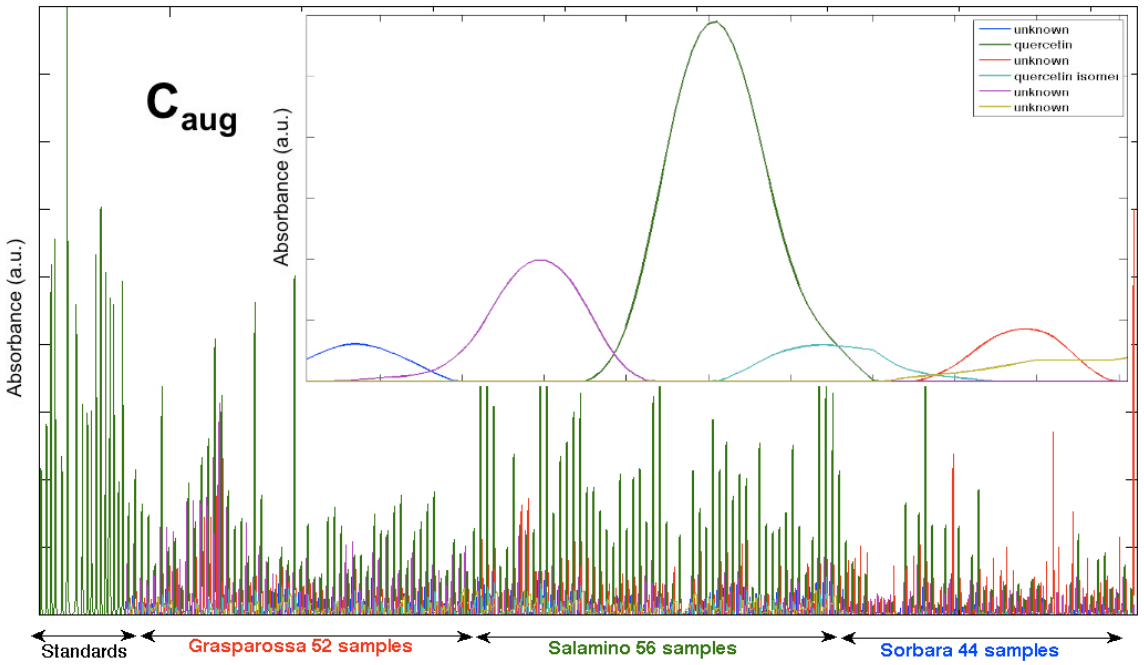


Figure 8: Matrix of the solved elution profiles for the Quercetin multiset: **C** profiles and the zoom of sample 0011_GRA, Quercetin (green line) and the interferents (cyano line).

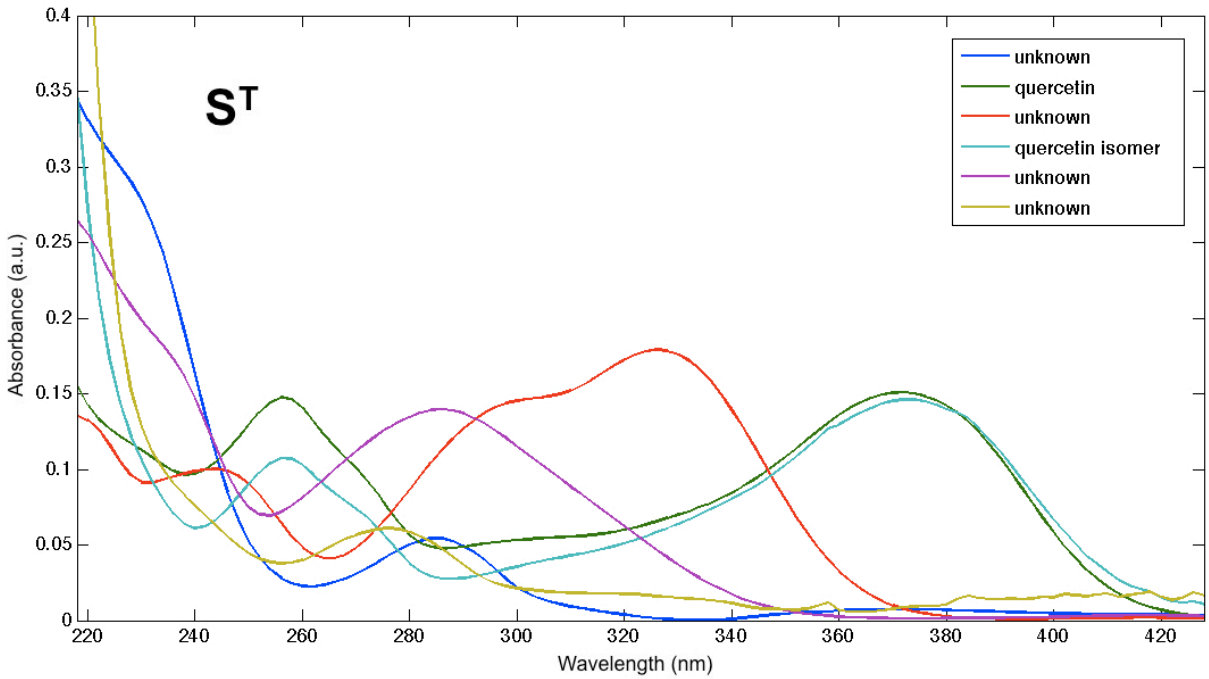


Figure 9: Matrix of the pure spectra $\mathbf{S}^T$ for the Quercetin multiset, Quercetin (green line) and the interferents (cyano line).

4.1.2 Quantification of target compounds

4.1.2.1 Calibration lines and validation

Multivariate curve resolution provides elution profiles and spectra for each analyzed compound, but can also offer quantitative information. This information can be derived from the area of the resolved chromatographic peaks, obtained from the elution profiles of each $\mathbf{C}_k$ submatrix in $\mathbf{C}_{aug}$ (Figure 10). When runs of mixtures of standards compounds at different concentration levels are analyzed with the wine sample runs, the peak areas obtained from the MCR-ALS results for the standards can be correlated to the real concentrations, like in univariate calibration. Since the information of each compound is separated in a particular elution profile, a calibration line for each of them is built, which allows the estimation of concentration in the analyzed samples. As described in the introduction of this section, calibration curves were derived for the quantification of the eight analytes: Gallic, Caffeic, Syringic, p-Coumaric and Vanillin, (+)-Catechin, Myrecetin and Quercetin.

For this purpose, standard solutions of single phenolic compounds and their mixtures with different concentration were prepared. Table 2 reports the quality parameters of the individual calibration models of the eight phenols analyzed. The values of limit of detection (LOD) and of quantification (LOQ) obtained by this technique are comparable with values obtained by other authors by using HPLC-UV in operational conditions ensuring good peak separation [10]. The advantage of the methodology presented in this work is that comparable results are obtained in a shorter analysis time, even in the presence of strong coeluting interferences. The concentration levels of standard solutions prepared for each analyte were the following expressed in mg L^{-1}:

Gallic acid: 0.5-1-5-10;

(+)-Catechin: 0.3-1-4-10;

Syringic acid: 0.5-1-2-4;

Caffeic acid: 0.5-1-2-4;

p-Coumaric acid: 0.5-1-2-5;

Vanillin: 0.5-2-5-10;

Myrecetin: 1-2-5-10;

Quercetin: 0.5-2-5-10.

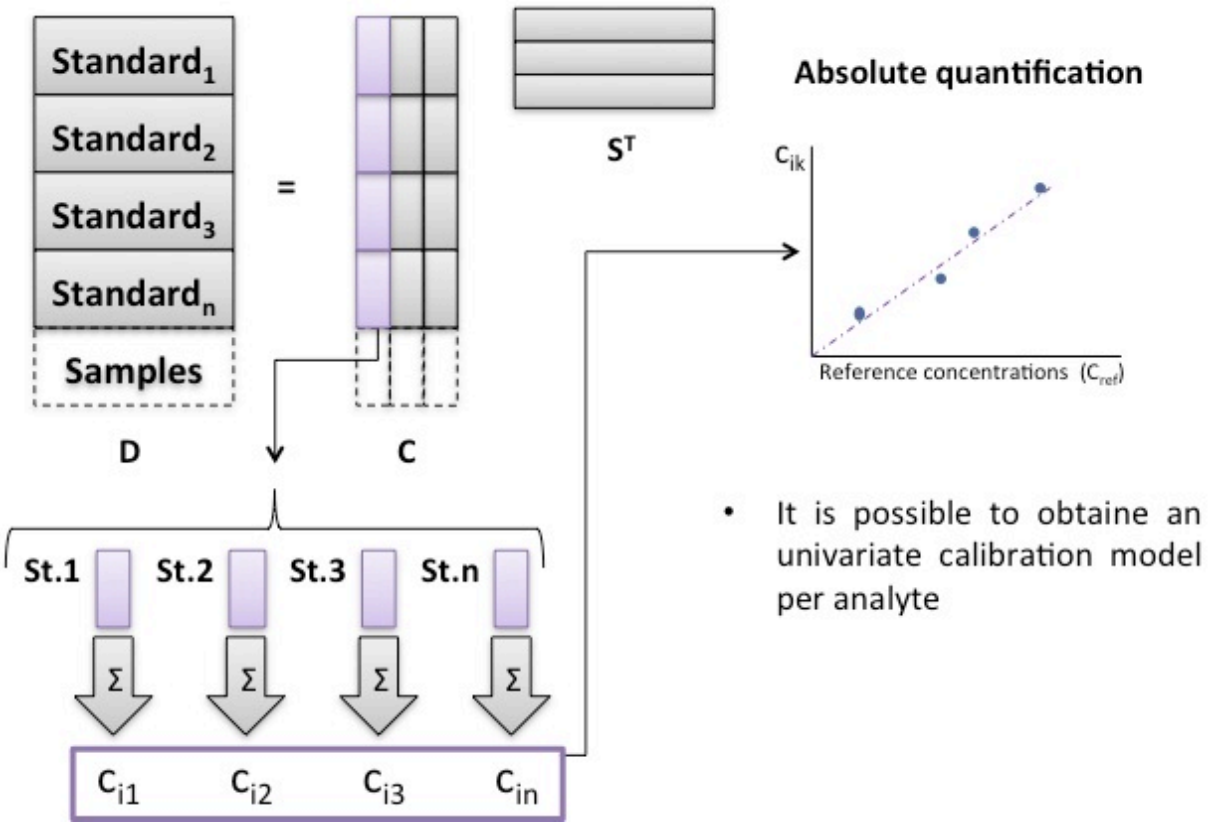


Figure 10: Representation of the quantitative information (peaks area matrix) obtained from the C matrix in the MCR resolution and absolute quantification performed with standard solutions.

The calibration models were used to determine the concentrations of the phenolic compounds in the 110 wine samples. These values are reported in Appendix III. Summarizing the quantitative information about the content of phenolic compounds in the wine samples analyzed, the minimum, maximum and mean value, of concentrations, of each Lambrusco variety, are listed in Table 2. As general conclusions, Vanillin is found to be in concentrations lower than LOD in all samples and, hence, can be assumed to be absent in these Lambrusco varieties. Quercetin is the most abundant phenolic

compound in all varieties, Gallic and p-Coumaric acids seem to dominate in some varieties and Syringic and Caffeic acid seem to be minor constituents in the three kinds of wines studied. Among varieties, Sorbara, in general, contains lower concentrations of almost all the compounds compared with the other two varieties that have more similar concentrations. This difference can be easily seen also by visual inspection since the Grasparossa and Salamino varieties present a dark purple color. Other trends can be seen, such as the high amount of (+)-Catechin in Salamino or the high amount of Syringic acid in Grasparossa. However, the possibility to differentiate wine varieties from the information related to the abundance of the phenolic compounds will be much better exploited when the whole fingerprint information is taken into account by multivariate analysis, as it will be shown in the next sections.

To verify the reliability and robustness of the model, an additional set of samples has been analyzed after several months applying the same experimental conditions and taking advantage of the previous MCR models for the resolution of the chromatographic peaks. The samples of this validation data set are in total 12, six wines are of the same Lambrusco wine varieties as the original data set, but collected at a second time, namely: 2 Grasparossa, 2 Salamino and 2 Sorbara; and the other six samples, here named 'extraterritorial' wines, belong to the same grape variety, namely Lambrusco, but the cultivation of the grape is located outside the district of Modena. Hence, since these latter wines do not respect the suitable geographical characteristics, they are not recognized within the 'Protected Designation of Origin' label. The multiset obtained for these samples has been analyzed in the same way as for the previously described 110 samples. Each pure spectral profile obtained from the resolution of the first multiset is used as initial estimation for the resolution of these validation sets, obtaining a very satisfactory matching among the first and second set of resolved spectra. In this case the peak areas allow the quantification of the phenolic compounds using the same calibration strategy as in Table 2. The concentration values obtained for these samples are reported in Table 3.

Table 2: Calibration parameters for analytes, LOD LOQ and concentration values are expressed in mg L^{-1}, the values in grey are below LOD.

Compound	Calibration lines	S_m[1]	S_b[2]	RMSEC[3]	R^2	LOD[4]	LOQ[5]	Grasparossa[6] max-min mean conc	Salamino max-min mean conc	Sorbara max-min mean conc
Gallic Ac.	9872x – 887	189	23	0.06	0.995	0.09	0.11	6.78 – 2.99 3.66	8.87 – 3.02 4.38	8.06 – 3.00 3.72
(+)-Catechin	10385x – 765	356	78	0.05	0.997	0.04	0.26	6.30 – 2.99 3.43	6.91 – 3.02 3.90	7.52 – 3.13 3.95
Syringic Ac.	8705x – 451	157	231	0.09	0.994	0.13	0.32	2.07 – 0.83 1.30	2.01 – 0.18(<LOQ) 0.99	1.30 – 0.32(<LOQ) 0.67
Caffeic Ac.	12905x – 3125	141	418	0.29	0.997	0.34	0.57	2.78 – 0.60 1.39	3.12 – 0.40(<LOQ) 1.70	3.27 – 0.79 1.68
p-Coumaric	13289x – 1023	291	120	0.32	0.993	0.04	0.13	6.10 – 0.72 2.26	4.71 – 0.28 2.47	5.96 – 0.90 2.80
Vanillin	7235x + 1251	248	37	0.12	0.996	0.09	0.14	Max 0.02(<LOD)	Max 0.07(<LOD)	Max 0.03(<LOD)
Myrecetin	2855x + 2158	81	348	0.13	0.992	1.12	1.98	7.23 – 0.76(<LOD) 3.01	7.25 – 0.84(<LOD) 3.40	5.48 – 0.89(<LOD) 1.72(<LOQ)
Quercetin	11541x + 2489	76	199	0.43	0.996	0.26	0.38	13.64 – 2.35 4.82	13.75 – 1.60 5.70	9.33 –1.80 3.05

[1] Standard deviation of slope of the calibration line
[2] Standard deviation of offset of calibration line
[3] Root mean square error in calibration $\sqrt{[\sum(c_{pred}-c_{true})^2/n-1]}$
[4] Limit of detection (LOD) $3S_b/b$
[5] Limit of quantitation (LOQ) $10S_b/b$
[6] Concentration values for each analyte in each wine variety, first line: maximum - minimum and second line: average

Table 3: Concentration values of the phenolic compounds for the test samples, all the concentration are expressed in mg L^{-1}, values of Vanillin are below the LOD.

Test samples		Gallic Ac.	(+)-Catechin	Syringic Ac.	Caffeic Ac.	p-Coumaric Ac.	Vanillin	Myrecetin	Quercetin
Grasparossa	s0041	4.12	3.20	1.78	1.45	3.01	-	0.94(<LOD)	1.79
	s0044	5.38	3.06	3.28	2.69	2.65	-	7.51	18.96
Salamino	s0042	10.45	3.09	2.37	3.69	1.92	-	7.33	10.10
	s0044	9.28	3.07	2.08	2.88	1.88	-	4.97	6,27
Sorbara	s0035	12.18	3.09	2.21	3.09	0.93	-	4.37	4.99
	s0037	19.79	3.10	3.26	3.09	1.22	-	4.24	3.60
Extraterritorial	x0003	24.29	3.18	1.68	2.69	5.11	-	6.06	6.00
	x0007	13.26	3.13	0.72	1.39	3.12	-	1.07(<LOD)	1.94
	x0011	11.42	3.03	0.77	1.71	4.67	-	1.42(<LOQ)	3.26
	x0012	15.18	3.78	2.43	4.77	1.71	-	6.45	13.43
	x0016	19.79	3.75	0.18(<LOQ)	1.57	2.13	-	1.62(<LOQ)	2.87
	x0019	22.94	3.15	0.75	1.94	2.30	-	1.15(<LOQ)	1.97

The concentration values of the external samples of PDO wines are comprised in the range previously determined for the related wine class, which confirms the reliability of the quantitative information obtained. Some expected differences in concentration levels of polyphenols are observed for the samples belonging to the 'extraterritorial' samples, which seems promising in the perspective of deriving authentication models. Moreover, standards solution at the maximum and minimum concentration values used to build the calibration lines were also analyzed and values of RMSEP between 0.13 and 0.72 were obtained, fairly similar to those related to the calibration models in Table 2.

4.1.3 Screening analysis of the data

The peak areas obtained from the resolution results, belonging to both phenolic targets and additional resolved compounds represent the relative quantitative information of all the eluted species and are the fingerprint information of the samples. These data are collected in a matrix in which the columns (variables) correspond to the peak areas of the resolved compounds (phenolic or non-identified) and the holds the different wine sample analyzed. This matrix of peak areas is studied with principal component analysis, which decomposes the data in a scores matrix (wine samples coordinates in PC space) and a loadings matrix (expressing the contribution of each peak area in the PCA model, hence related to wine composition). The interpretation of both plots allows us to explore the possibility to distinguish among wine varietals and to identify the compounds responsible for their characterization. Two pretreatments are applied to the matrix of peak areas prior to PCA: normalization to unit area per rows and auto-scaling per columns. Normalization along the rows is primarily applied to reduce the effect of the total intensity due only to variation in samples total concentration that may be due to differences of volume in the sample extraction. This is done because the interest of the analysis is on the relative amount of compounds within a sample and not on the differences of global intensity among fingerprint profiles. Auto-scaling along the columns is done

to give the same potential importance to all compounds analyzed.

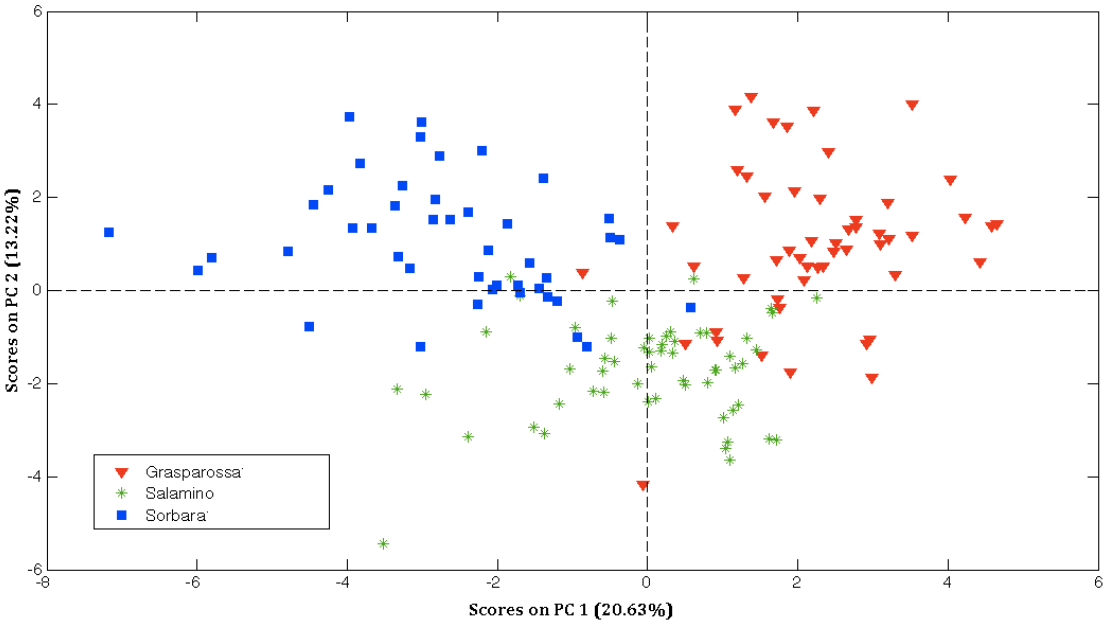


Figure 11: Scores plot, PC1 vs. PC2, of PCA of the fingerprinting information data (110 samples x 27 peaks area).

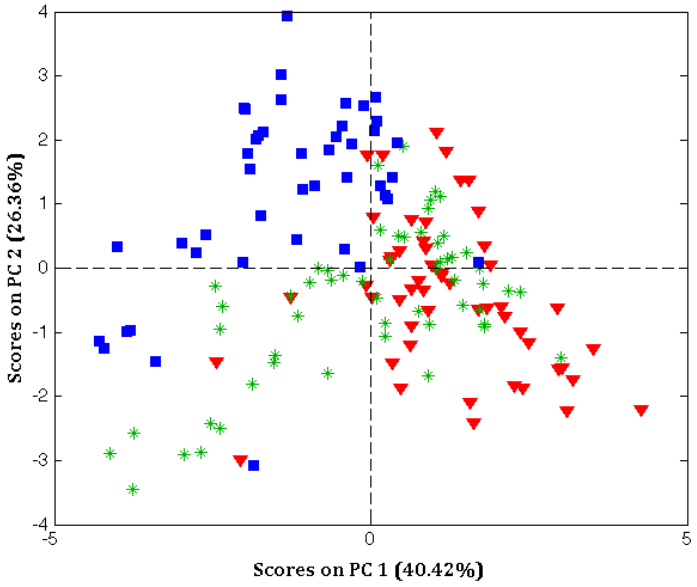


Figure 12: Scores plot, PC1 vs. PC2 of PCA of quantified phenolic compounds data (110 samples x 8 peaks area).

PCA results from the analysis of the complete fingerprinting information containing all resolved compounds (target phenol compounds and unidentified species) are shown in Figure 11. The first principal component helps to differentiate between the two varieties Sorbara and Grasparossa while the second principal component separates these classes from the third one of Salamino. The loadings plot (Figure 13) shows that Quercetin, Syringic and Myrecetin characterize the Grasparossa samples (same position in the loadings plot as Grasparossa samples, red triangles, in the corresponding scores plot) and, indeed, these compounds have clearly the lowest concentrations in the Sorbara variety, which is oriented in the opposite direction on scores plot (blue squares). p-Coumaric and Caffeic acids differentiate the Sorbara variety from the rest and Gallic acid seems to be clearly more present in the Salamino variety when compared with the other two. Besides, the correlation among phenol concentrations can also be clearly seen. For instance, Quercetin and Myrecetin concentration values shows a direct correlation and an inverse correlation with Catechin, which suggests that maybe not all these compounds need to be determined for authentication purposes. Conversely, p-Coumaric acid seems completely uncorrelated to the previous compounds and slightly correlated with Caffeic acid.

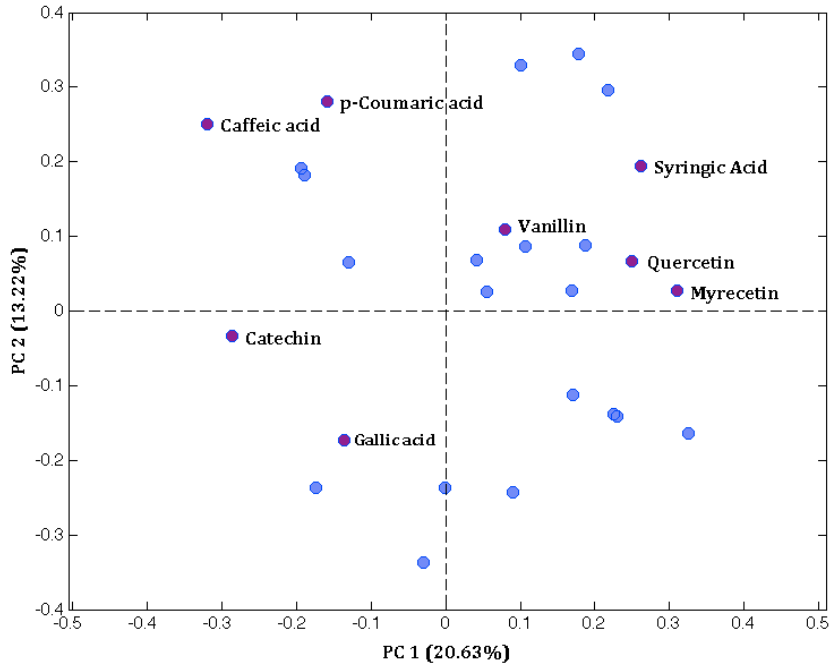


Figure 13: Loadings plot of PCA model including all 27 peaks area, highlighting the variables useful for the differentiation of the classes.

The scores plot resulting from PCA performed on a data matrix containing only the peak areas of the eight quantified phenolic compounds (Figure 12) is also provided. This plot highlights that the sole use of the phenolic compounds allows only the differentiation of the Sorbara variety from the other two varieties. Without the introduction of the information related to the additional compounds obtained from the multivariate resolution analysis, it would not be possible to distinguish the two varieties with the most similar total concentrations, i.e., Grasparossa and Salamino. This point clarifies the importance of the quantitative data provided by MCR-ALS, which gives the peak areas of all the resolved species in each of the six elution windows in addition to the target compounds. This additional information allows a better separation of the wine varieties.

4.1.4 PARAFAC results

The same elution windows solved with MCR were also considered for testing the second order calibration capability of PRAFAC algorithm. The multiset shown in Figure 3 was refolded to a 3-way array. Since the elution profile still present the problem of misalignment, before applying PARAFAC, which would suffer from lack of trilinearity due to shift, these were corrected with the Icoshift algorithm. An alternative could be to use PARAFAC2 instead of PARAFAC, which handles also with slab-wise deviation. The solution of applying Icoshift and then PARAFAC was chosen here because more rapid, since PARAFAC2 algorithm sometimes is time consuming and the aim was only to confirm the resolution results obtained from MCR. Elution profiles were constrained with non-negativity and unimodality and spectra profile were constrained with unimodality. The number of components was selected considering the values of core consistency and the resemblance of the components spectra with standards compounds. The number of resolved components and the value of core consistency obtained for each model are reported in Table 4, for Quercetin and p-Coumaric elution windows the number of component of PARAFAC model is lower with the respect to MCR, this because chromatographic peaks of interferents and minor compounds are of difficult resolution and this explain how important is the possibility of applying different constraints in the algorithm depending on the situation. Anyway, spectra obtained from the PARAFAC resolution, are consistent with those obtained from MCR-ALS for all the components resolved, as shown in Figure 14 and 15.

Data set	Number of component	Core consistency %
Gallic Acid	3	83
(+)-Catechin	3	92
Syringic, Caffeic acids	4	80
p-Coumaric, Vanillin	3	61
Myrecetin	4	86
Quercetin	4	99

Table 4: Resolution parameters of PARAFAC models for the six data sets.

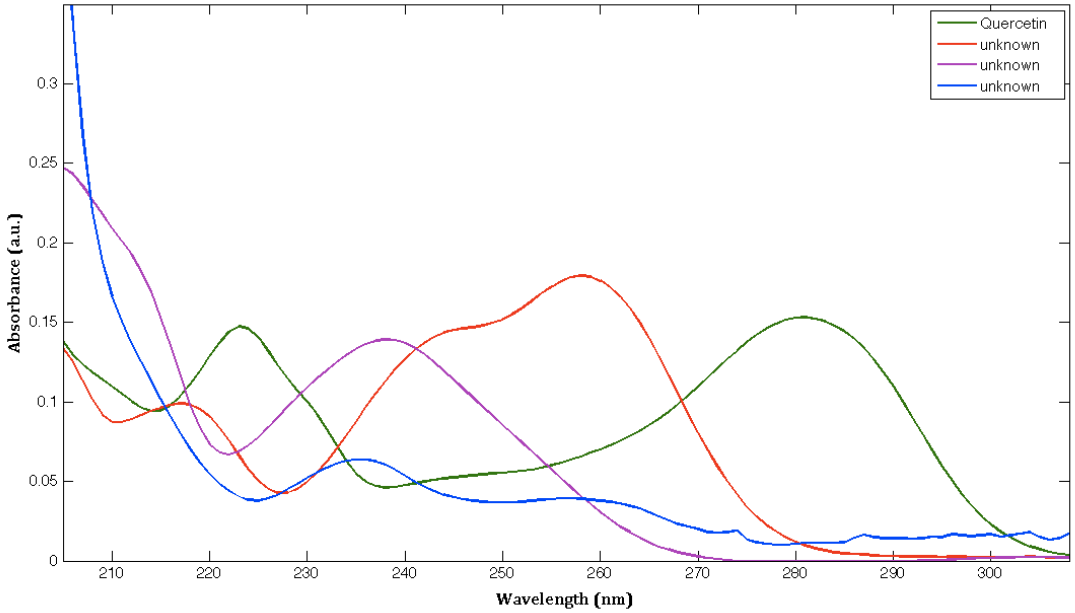


Figure 14: Lambrusco data set – PARAFAC resolution. Spectra obtained fro the resolution of Quercetin elution window, to compare with spectra in Figure 9.

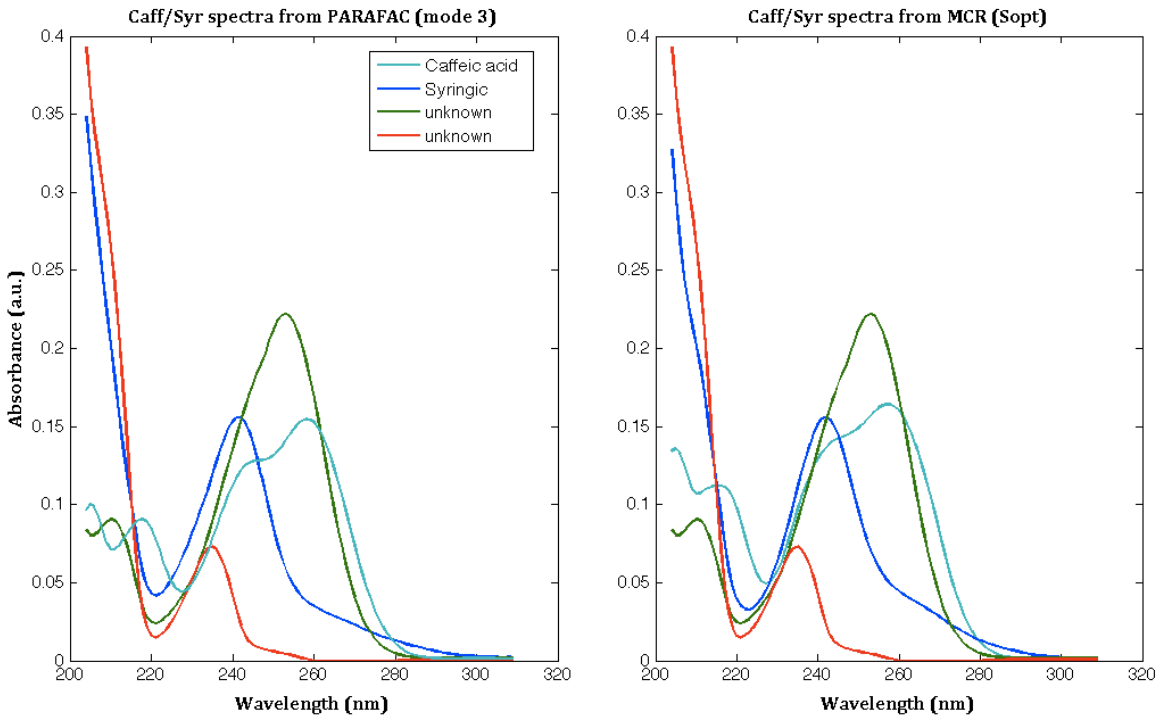


Figure 15: Lambrusco data set – PARAFAC resolution. Spectra of the Caffeic and Syringic acid elution window, with PARAFAC and MCR resolution.

The PARAFAC scores obtained by each of the six elution window were merged together in a data matrix and analyzed with PCA, the obtained scores plot reported in Figure 16, shows a distribution similar to that obtained in Figure 11 from MCR results.

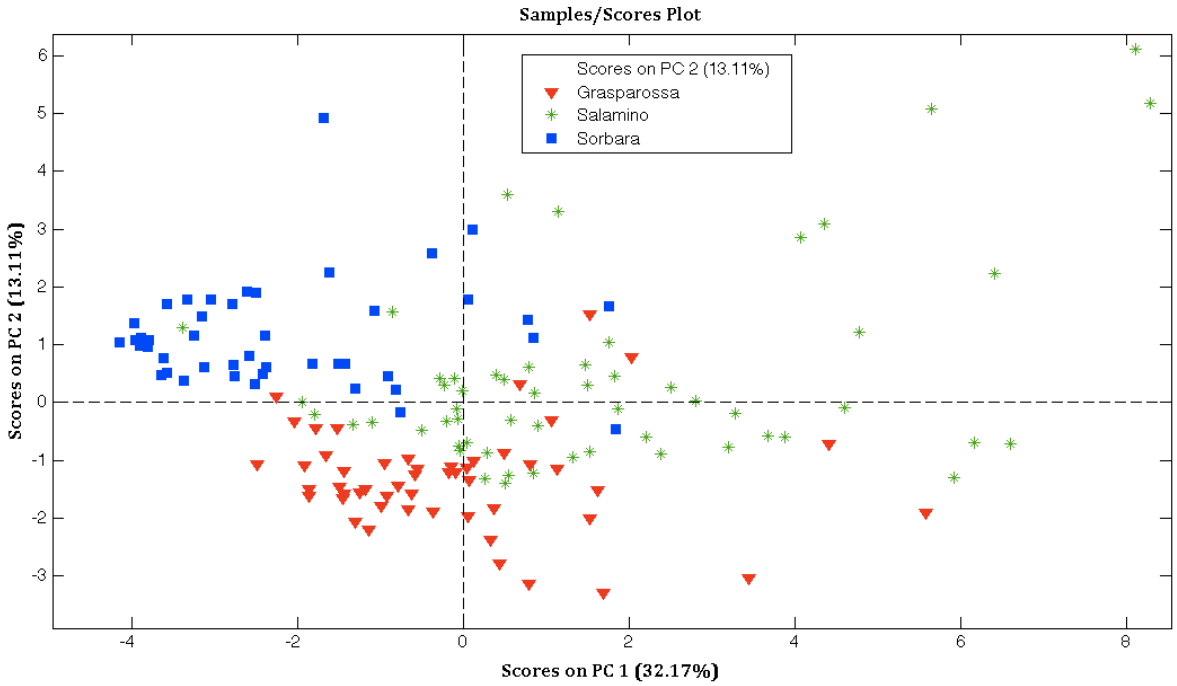


Figure 16: Lambrusco data set – PCA on the score of PARAFAC resolution.

4.2 Multiset & multi-way classification

4.2.1 Lambrusco data set classification

This section deals with the differentiation among the three Lambrusco wine varieties. The results obtained with the MCR resolution methodology are elaborated with different classification approaches, classical and MCR-ALS discriminant. Moreover, these results are compared with multiway classification approaches on the raw three-way HPLC-DAD data (samples x retention times x wavelengths).

4.2.1.1 PLS-DA

Since the results obtained from the principal component analysis of the areas values coming from the MCR resolution of the six windows previously discussed were promising (Section 4.1.4 Figure 11) the resolution was extended also to other two elution windows containing peaks with a relevant signal as shown in Figure 17. The resolved compounds in this windows were not identified, since we did not have reference standard and an HPLC-MS identification is under progress but not accomplished yet, so clearly in this case there is not information on the spectrum of standard compounds, but anyway they belong to the total fingerprinting information of the Lambrusco varieties. Resolution parameters for the two multisets are reported in Table 5.

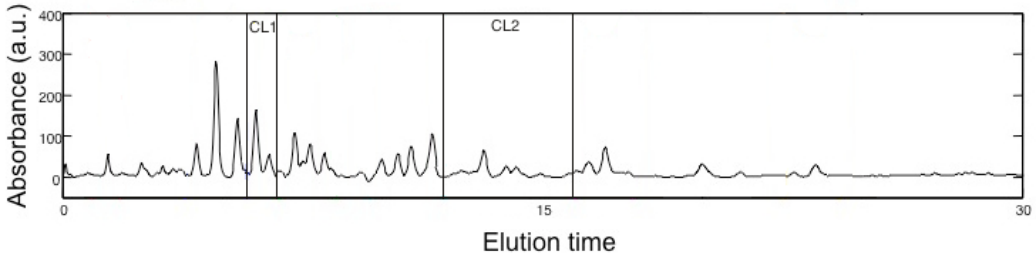


Figure 17: Elution windows containing cluster 1 and cluster 2 unknown compounds resolved for the discrimination of classes.

Multiset	Number of component	% Explained variance	% Lack of fit exp
Cluster 1	3	97.1	9.3
Cluster 2	9	98.6	6.8

Table 5: Resolution parameters for the two data sets resolved.

Final fingerprinting information extracted through MCR resolution consist in the peak area of 39 compounds, namely 39 variables that are then used for the characterization of the three wine categories. Even if the number of sample analyzed is 110 the final dimensions of the matrix is 152x39, since some samples were replicated during the analysis to verify the reproducibility of the results. Then the data are split in two subsets in proportion 1:3 and the first is used as training set and the second as test set, these subset are selected with Duplex algorithm maintaining the replicates in the same set. Considering that for each class some replicates are inserted in the calculation the total samples for Grasparossa are 52 (34 train, 18 test), Salamino 56 samples (37 train, 19 test) and Sorbara 44 samples (29 train, 15 test). The data were first normalized on rows direction (per sample); with this operation the values of area are made comparable independently of eventual error due to different dilution committed in the extraction phase or natural differences in the water content of the samples. Then, in order to let each variable (resolved compounds) to contribute the same to the model data were columns autoscaled. The number of PLS components was selected according to minimum classification error in leave one out cross validation and seven components were selected. Samples are assigned to the class if the value of Y predicted is higher than a threshold calculated considering the minimum error of classification for the training set (assignation of PLS-toolbox).

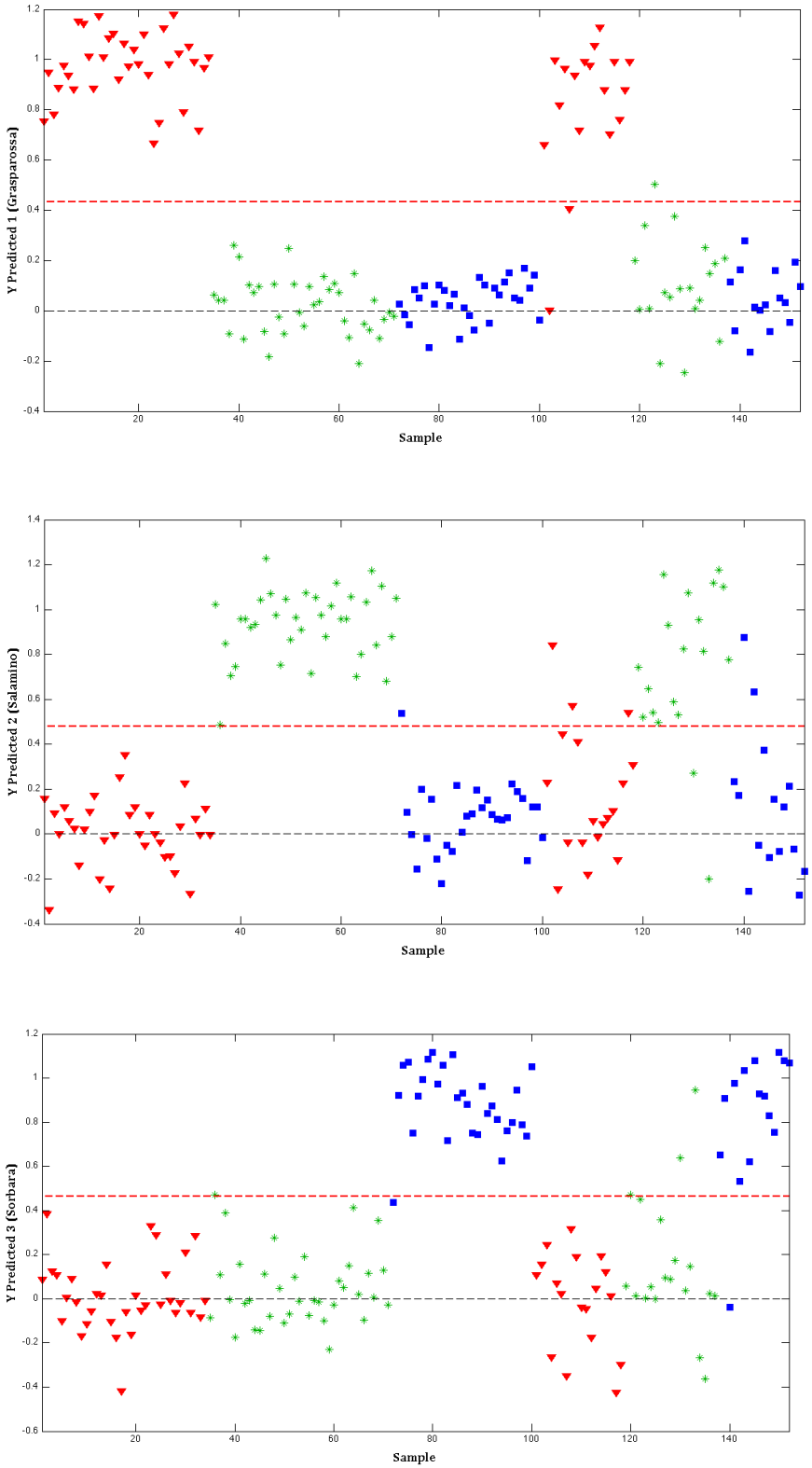


Figure 18: Plots of the Y predicted values for the PLS-DA model of Lambrusco data set, 7 latent variables were selected.

Figure 18 reports the values of Y predicted for each samples and the values of sensitivity and specificity for each class can be found in Table 6.

	CL1 Grasparossa	CL 2 Salamino	CL3 Sorbara
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/100	100/98	97/99
CV	100/98	95/95	97/94
TEST	89/97	89/85	93/92

Table 6: Lambrusco data set – PLS-DA on peak areas. Sensitivity and specificity in calibration, prediction and cross-validation, 7 latent variables were selected.

In Figure 19 are reported the VIP scores for the three classes, the variables with value higher than 1 are considered most relevant for class discrimination. It is interesting to notice that the varieties Grasparossa and Sorbara seems to be characterize by the same phenolic compounds, even if from the regression vector plot (Figure 21) it can be evinced that in many case Grasparossa variety is characterized by a higher content of those compounds and Sorbara by lower contents. Salamino wine is discriminated mainly for three compounds, for which this variety presents higher concentration with respect to the other two. The compounds name is reported in Table 7, for the previously quantified compounds. Some other constituents were tentatively identified on the basis of mass spectrometry detection, comparing retention times and mass spectrum and UV spectrum solved with MCR with literature [11].

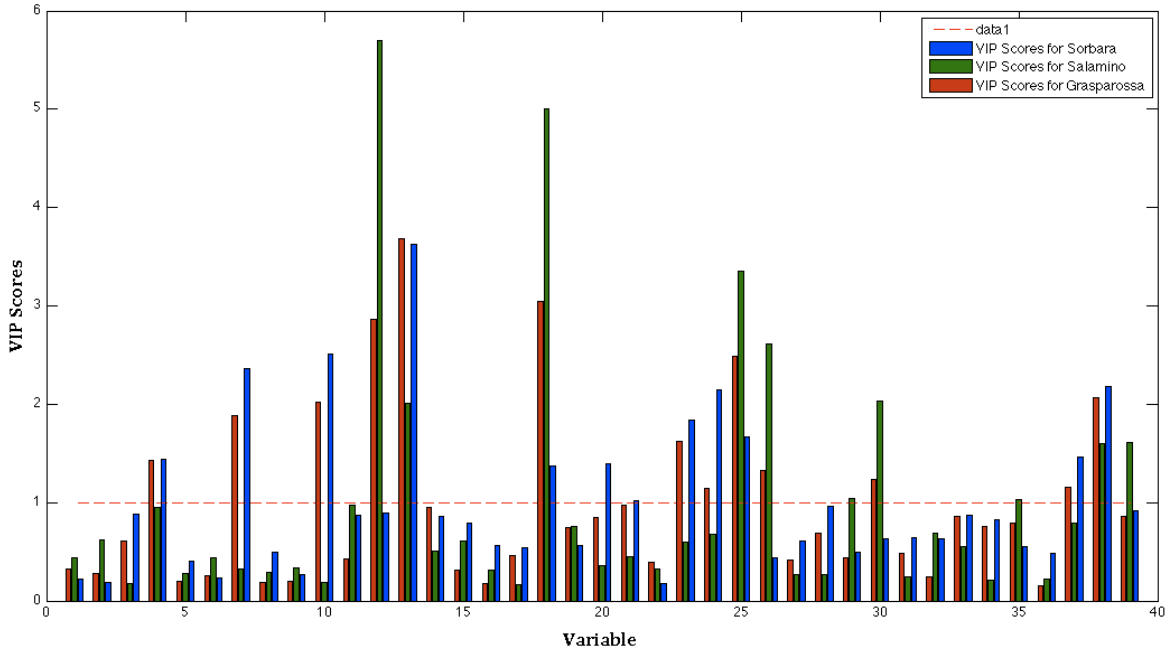


Figure 19: Score VIP for the PLS-DA model of Lambrusco data set.

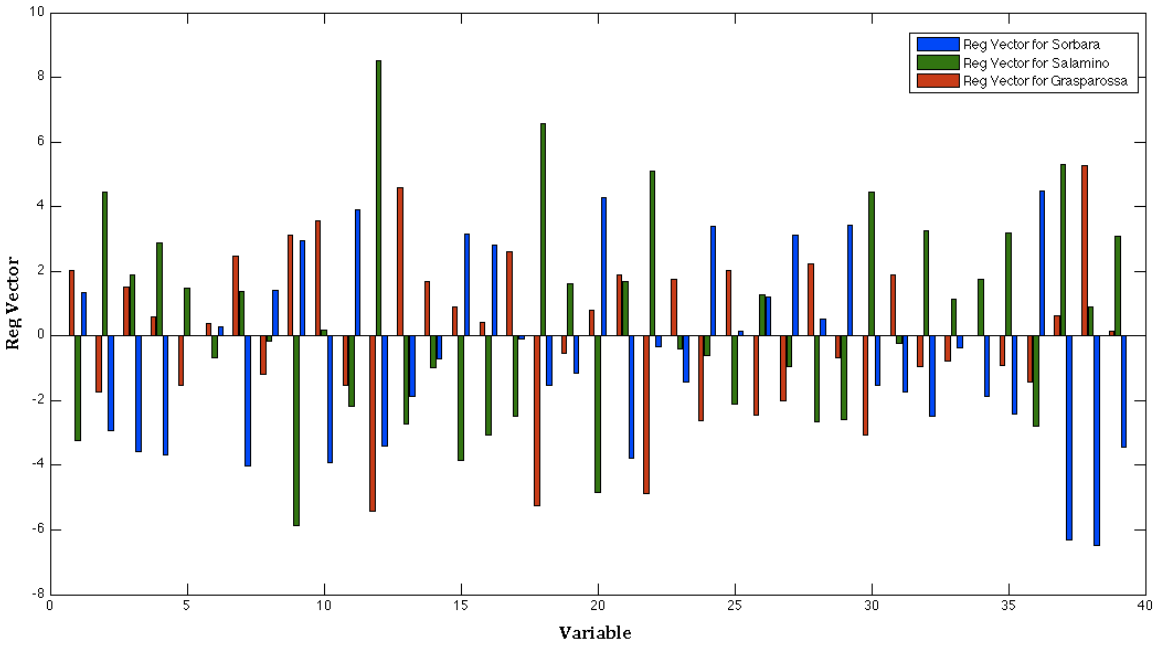


Figure 20: Regression vectors for the PLS-DA model of Lambrusco data set.

Peak Number	Compound	λmax (nm)	[M-H], Fragments (m/z)
1	unknown		
2	**Protocatechuic**	294	153, 109
3	**Gallic acid**	272	
4	unknown		
5	**Catechin**	280	
6	**trans-Caftaric acid**	330, 298	311, 179
7	unknown		
8	**trans-Fertaric acid**	330	325, 193
9	unknown		
10	**Syring acid**	276	
11	**Caffeic acid**	324	
12	**Procyanidin B2**	280	577, 423, 290
13	**(-)-Epicatechin**	274	289
14	unknown		
15	**p-Coumaric acid**	310	163, 119
16	unknown		
17	unknown		
18	**(-)-Epicatechin**	274	290
19	unknown		
20	**Ethyl Gallate**	274	197, 169
21	unknown		
22	unknown		
23	unknown		
24	unknown		
25	unknown		
26	**Astilbin**	288	449, 303
27	unknown		
28	**trans-Resveratrol (Glu)**	307, 319	
29	unknown		
30	unknown		
31	**cis-Resveratrol (Glu)**	285	389, 227
32	unknown		
33	**Mirecetin**	370, 300	
34	**trans-Resveratrol**	306, 319	
35	**Quercetin**	369, 300	
36	unknown		
37	**Quercetin isom**	370, 300	
38	**cis-Resveratrol**	284	
39	unknown		

Table 7: Lambrusco data set – Identification of resolved compounds, in yellow the quantified ones, in bold the tentatively identified.

4.2.1.2 MCR-ALS discrimination

The same data set described in Section 4.2.1.1 and the same training and test sets were considered. The discriminant method applied, called MCR regression, has been described in detail in Chapter 2. Before the analysis some pretreatments were tested selecting rows normalization and columns scaling as the ones giving best results. The MCR-ALS regression is separately applied to each class (on the training set).

Five components were selected as suitable for all classes considering the minimum value of explained variance and lack of fit (see Section 2.2.1); as initial estimation of the $\mathbf{S}^T$ matrix three components were given as the mean values, for each of the 39 areas values, calculated over the samples of the training set for each class. The initial **S** estimations, for the other two components, were assembled, by taking randomly two samples among the whole training set.

The only constraints applied in the calculation are the non-negativity for the spectra profiles and the selectivity constraint expressing the class membership. Non-negativity allows obtaining a different interpretation of the loading in the discrimination, since the values maintain their original meaning of chromatographic areas and describe a characteristic profile for the class. The selectivity constraint implements class membership, i.e. for three components in each class model the corresponding columns in the selectivity matrix takes value of 1 for samples belonging to the given class and zero for the others. So for instance in the model of class Grasparossa, the component whose spectral initial estimate is given as the average of Grasparossa training samples, is constrained and 1/0 are assigned to the training samples to the corresponding belonging/not belonging to this class, while the other four components are not constrained, i.e. assigning Inf or NaN to all the samples, both training and test, see Section 2.2.2.

In Figure 22 are reported the values of Y predicted by the MCR-ALS regression model for each sample, the threshold line is calculated considering the minimum of classification error in the training set (best compromise between sensitivity and specificity see Figure 21). The values of

the threshold indicates if a sample belong or not to the modeled class, each figure is referred to one class and the distinction among training set and test set is evidenced by a vertical line. The final results obtained, expressed in terms of sensitivity and specificity, are reported in Table 8.

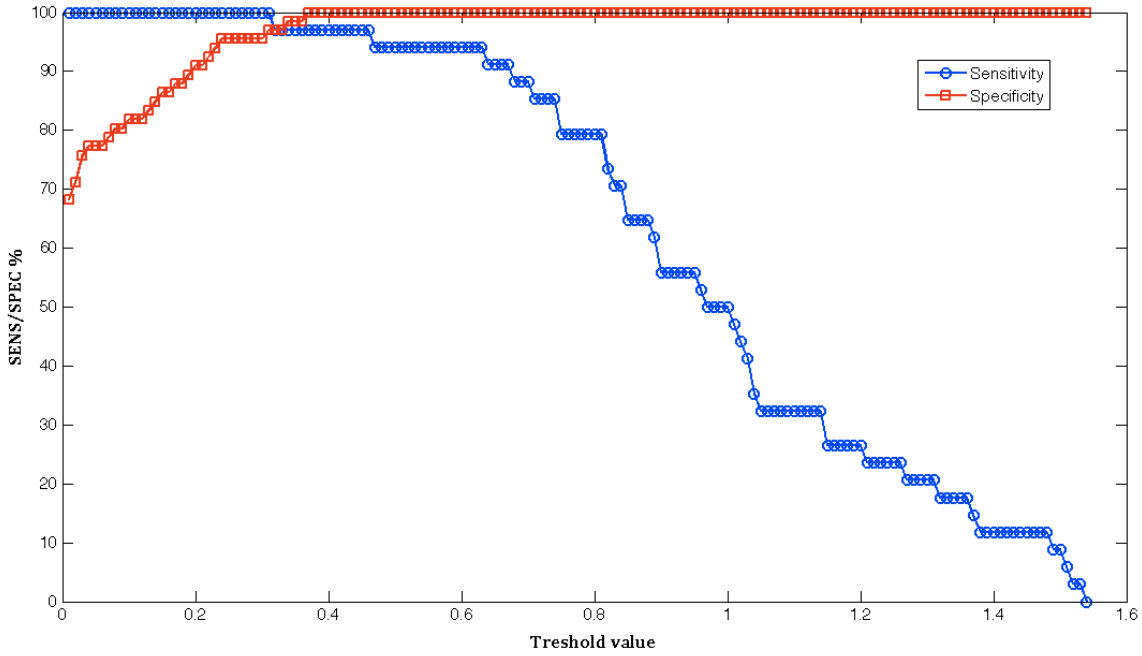


Figure 21: Lambrusco data set – ALS-regression. Selection of the threshold value for Grasparossa class.

Values obtained are of the same order of that calculated for the PLS-DA discriminant model, even if for class 2 and 3, PLS-DA still result the best discriminant approach.

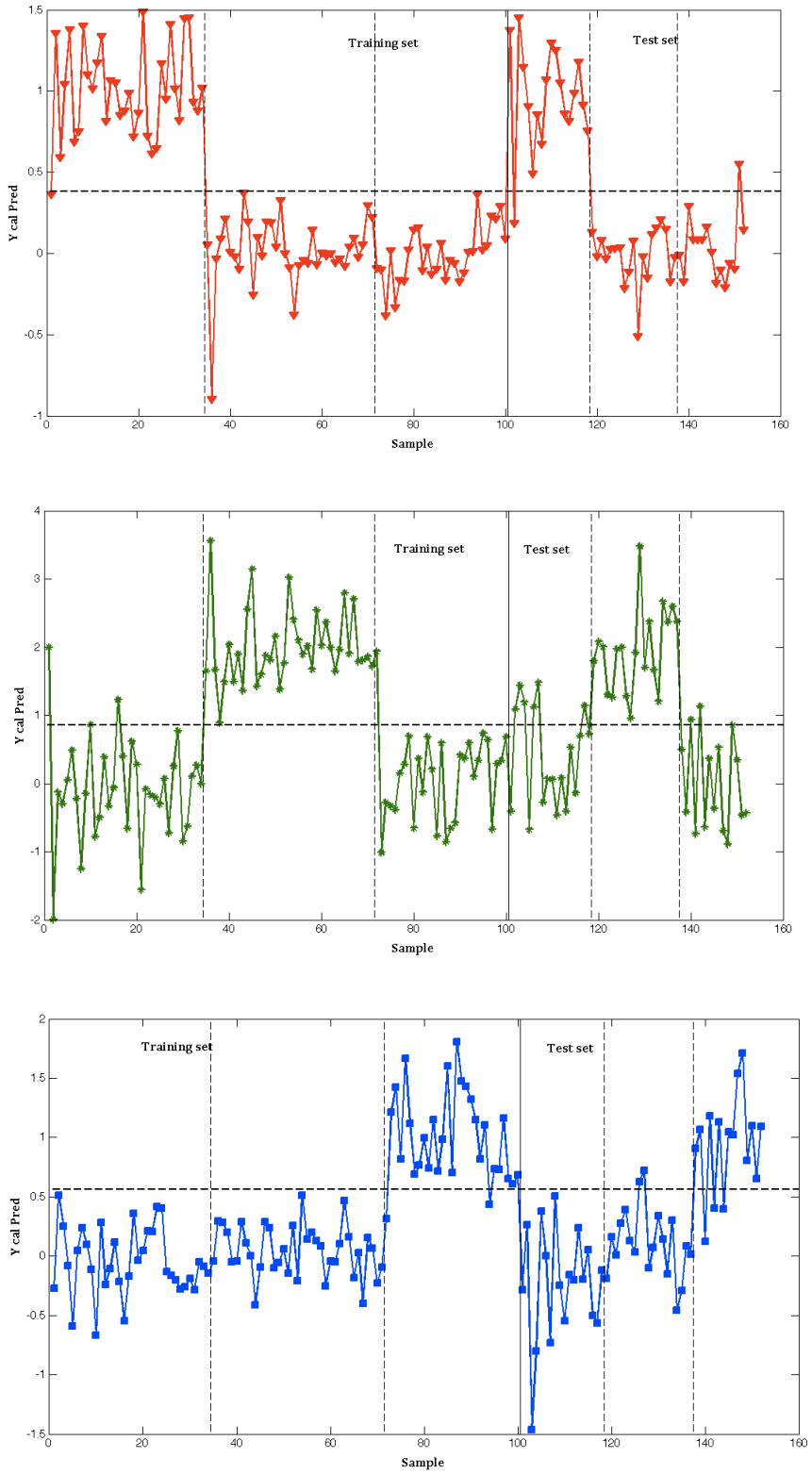


Figure 22: Lambrusco data set – ALS regression on fingerprinting matrix. Y prediction, Grasparossa Salamino and Sorbara class, 5 latent variables were selected.

	CL1 Grasparossa	CL 2 Salamino	CL3 Sorbara
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	97/100	100/98	96/100
TEST	95/97	100/77	80/95

Table 8: Lambrusco data set – ALS regression on fingerprinting matrix. Values of sensitivity and specificity for the three classes, in calibration and prediction, 5 latent variables were selected.

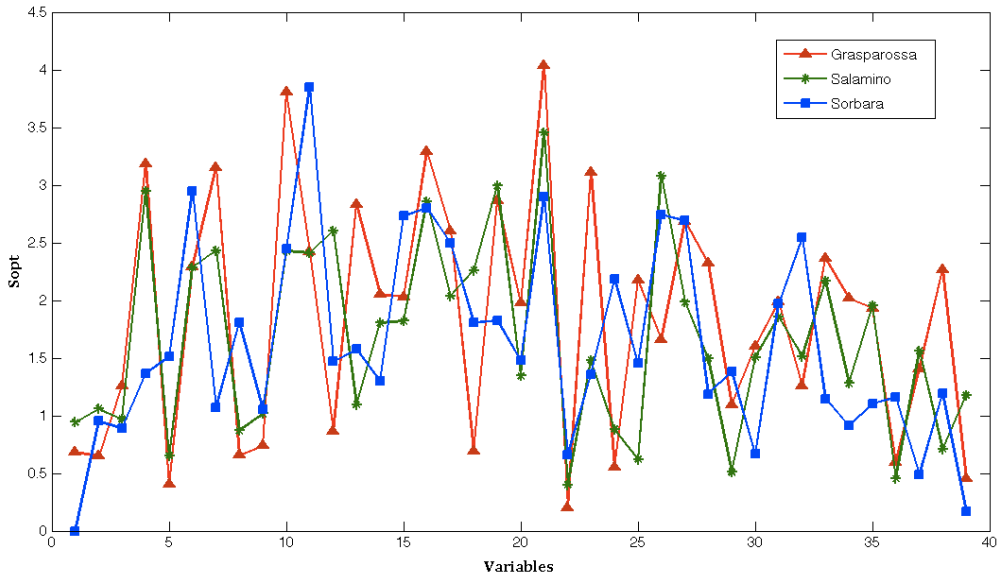


Figure 23: Lambrusco data set – ALS-regression. $\mathbf{S}^T$ optimized matrix, constrained components of the three models.

Figure 23 contains the optimized $\mathbf{S}^T$ components constrained with the class membership information in the iteration process, of each class model. This plot expresses the same information as a loadings plot and can be interpreted in the same way, but thanks to the Non-negativity constraint all the values are positives (they maintain the original meaning of peak areas and not abstract latent variable loadings). Some features are in common with the VIP plot and regression vectors of the PLS-DA model, but here it is highlighted the difference of the Sorbara class (blue line) which, except for six variables, presents always lower concentration of the species.

4.2.1.3 NSIMCA

The same data were also analyzed by N-SIMCA, method described in Section 2.3.3.2. To apply this method to the Lambrusco data set the entire chromatogram is considered and arranged in a 3-way structure with dimension 152x707x106 (sample x time x spectra). This array is a reduction of the original array (152x1211x106), obtained considering only the elution windows previously used for the resolution. Before the analysis the misalignment in the retention times direction (chromatograms) was corrected using the Icoshift algorithm; given the complexity of the chromatograms to have an effective shift removal the algorithm was separately applied in windows covering small regions of the retention times axis, taking care of having the boundaries of each window belonging to baseline (i.e. of not truncating or splitting a peak over more than one window). As explained in the data pretreatment Section 2.4.1 the shift correction was calculated on the chromatogram obtained summing the intensity at all the wavelengths and then it is applied to each spectral wavelength. This step is the most difficult one since these chromatograms, unlike those that will be object of the following section, is characterized by a strict peak coelution, in Figure 24, is reported how the data appear before and after the alignment procedure. Different pretreatments were applied in the three modes: mean-centering in the first, pareto-scaling in the second and the third, with the order 3,2,1.
The NSIMCA model is calculated using a PARAFAC as decomposition method, the number of components are selected considering best efficiency for a 'leave one out' cross validated results on a range of factors investigated from 4 to 10 for each class. The number of components was selected considering the best efficiency in 'leave one out' cross-validation on a range of factors investigated from 4 to 10. The number of components selected resulted to be 8 for each class. The sensitivity and specificity values reported in table, refers to the alternative SIMCA framework and to class limits calculated according to Pomerentsev criteria [12], which gave the best results.

The classification results obtain (Table 9) are dramatically worse with respect

to those reached in the discriminant approach, even if they can be considered acceptable for the Grasparossa and Sorbara class the model is not specific for the Salamino class. This data set of Lambrusco is a typical case in which the N-way approach is not as performant as the one implying the resolution step on chromatograms. This is due to the difficulty of the method of treating with strictly coeluted chromatographic signal. Here the data are reported only to compare the results obtained with the other approach.

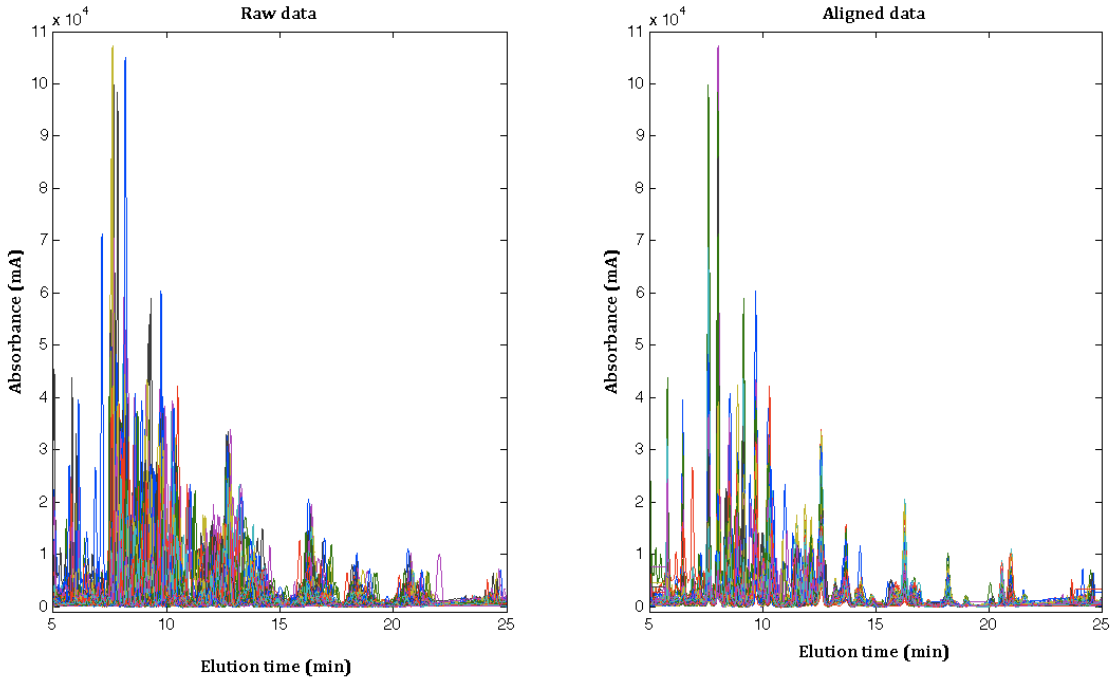


Figure 24: Lambrusco data set – 3way data. Raw data and aligned data with Icoshift.

NSIMCA	CL1 Grasparossa	CL 2 Salamino	CL3 Sorbara
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/51	92/17	97/71
CV	77/41	71/17	70/80
TEST	82/56	66/23	50/86

Table 9: Lambrusco data set – NSIMCA. Sensitivity and specificity for the three classes in Calibration, Cross-Validation and prediction for three different pretreatments dimensionality [8 8 8].

4.2.1.4 NPLS-DA

The array was analyzed also by mean NPLS-DA, maintaining the same data pretreatment. The number of latent variables was selected, as the minimum classification error, in 'venetian blind' cross validation with a split rate of six elements (8 LV). Samples are assigned to the class for which present the higher values of Y predicted. Results of the classification in terms of Y predicted and sensitivity and specificity for the three classes are reported in Figure 25 and Table 10.

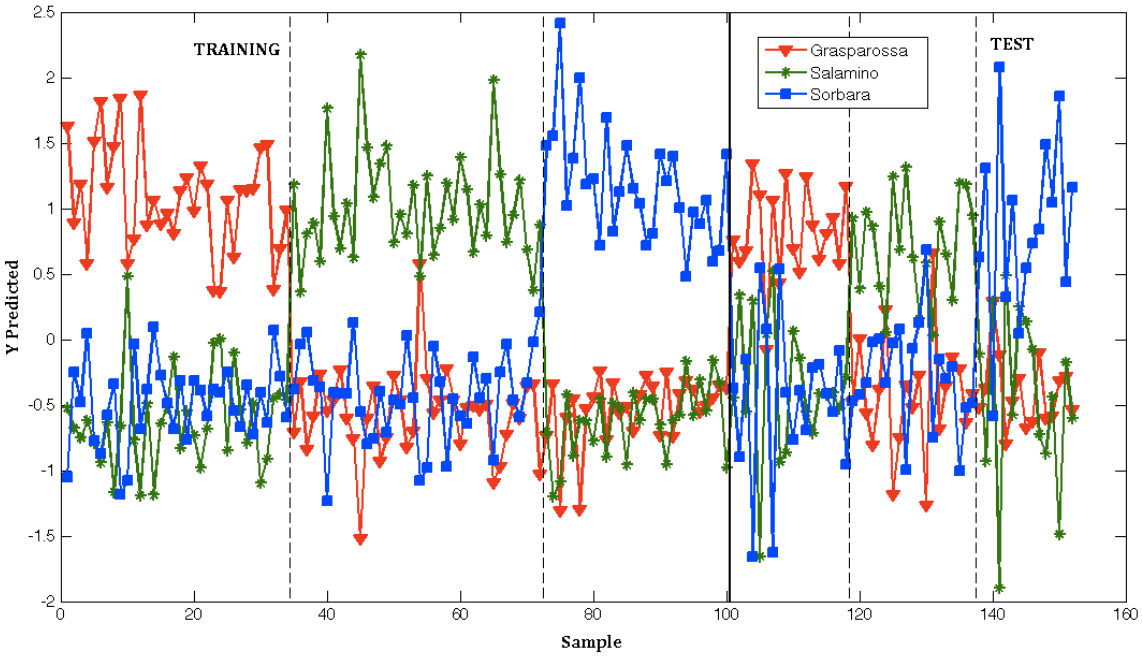


Figure 25: Lambrusco data set – NPLS-DA. Y predicted for the three classes (8LV) training and test set, 8 latent variables were selected.

NPLS-DA 8 LV	CL1 Grasparossa	CL 2 Salamino	CL3 Sorbara
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/98	97/98	96/100
TEST	89/94	84/94	80/99

Table 10: Lambrusco data set – NPLS-DA. Sensitivity and specificity for the three classes in Calibration and prediction, 8 latent variables were selected.

For this data set the discriminant approach seems to be the most appropriate, this can be attributed to the fact that, as explained in Chapter 3, the three DOP are produced from similar grapes cultivated in adjacent territories, and for the production regulation they can be also mixed. As a consequence the class are very similar and overlapped.

4.2.2 EVOO/WINE data sets classification

In this section the two data sets WINE and EVOO described in detail in Chapter 3 are used as benchmark to exploit the potentiality of N-way class modeling (NSIMCA) and discriminant classification (NPLS-DA) and MCR-ALS regression/discrimination. In particular these results obtained in multi-way and multiset classification are compared with those obtained with unfolding approaches. The simpler classification approaches, consisting in unfolding the data array, $\underline{\mathbf{X}}_{I,J,K,}$ to a matrix, $\mathbf{X}_{I,JK}$, and then applying standard bilinear classification techniques (PLS-DA and SIMCA) will be described first. Second part regards the multiset discriminant approach consisting in a step of peaks integration and resolution with MCR and the application of ALS-regression and PLS-DA on the fingerprinting matrix of peak areas. Finally he results of the n-way methods NSIMCA and NPLS-DA are reported.
The results are discussed in parallel for the two data sets.

The general pretreatments applied to the raw data of both WINE and EVOO data sets are listed below:
i. Baseline removal: Baseline contribution has been corrected by using the EBS method [13] (described briefly in Section 2.4.2).
ii. Misalignment correction: These are corrected with IcoShift [14] (described in Section 2.4.1), the data were first converted into a matrix considering the TIC signals, then after the alignment of the TIC data (Figure 26), the displacement applied for each sample is reproduced in each point in m/z direction).
iii. Reduction of matrices dimensionality: Retention times and masses with a variance near zero are removed to operate only on statistically sensible

information. Clearly this data reduction is applied independently on 3-way and unfolded data and as a consequence the deleted times and masses are not exactly the same for the two data structures so there are some differences in the reduction of the data set for the two approaches.

iv. Centering and scaling: Both data sets were mean-centered, unfolding approach, (or centered across the first mode, three way arrangement) and scaled testing different scaling methods (see Section 2.4.3); moreover, comparison with unscaled data was also considered.

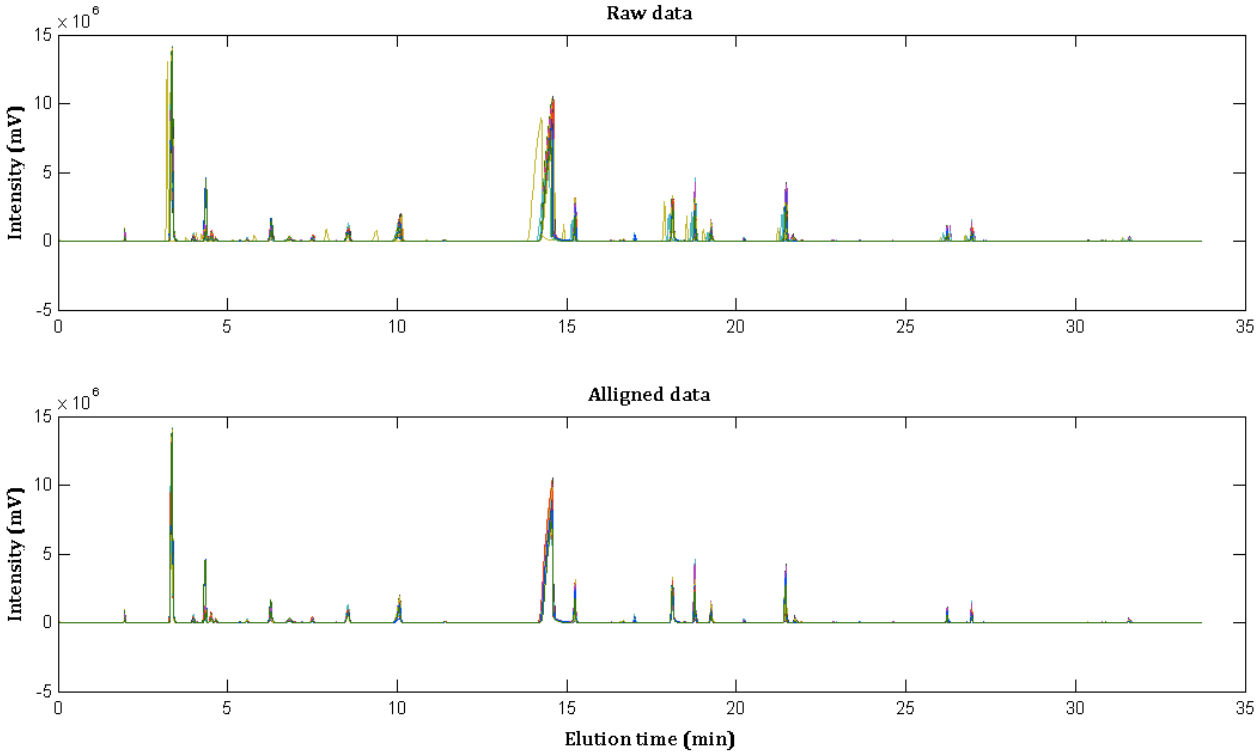


Figure 26: Peaks alignment of WINE data set chromatograms.

4.2.2.1 Unfolding

The PLS-DA results for WINE data set are reported in Table 11, where the samples are assigned according to the criteria of higher value of Y predicted and the effect of the different preprocessing can be assessed. Based on these outcomes auto-scaling and gave correct classification rates higher than 80% for the external test set samples, for all the three classes.

Auto-scaling	CL1 South America SENS/SPEC	CL 2 Australia SENS/SPEC	CL3 South Africa SENS/SPEC
TRAIN	100/100	100/100	100/100
CV	95/100	100/100	69/82
TEST	80/93	86/100	86/88

Mean-centering	CL1 South America SENS/SPEC	CL 2 Australia SENS/SPEC	CL3 South Africa SENS/SPEC
TRAIN	95/96	100/96	92/100
CV	85/93	69/96	69/88
TEST	90/86	57/100	86/76

Pareto-scaling	CL1 South America SENS/SPEC	CL 2 Australia SENS/SPEC	CL3 South Africa SENS/SPEC
TRAIN	100/100	100/100	100/100
CV	80/96	85/100	69/82
TEST	90/93	71/100	100/88

Table 11: WINE data set - PLS-DA classification after unfolding. Correct classification rates for the three discriminated categories in calibration, cross-Validation and prediction for three different pretreatments, in each cell first value auto-scaling (7LV), mean-centering (7LV), second value third value pareto-scaling (9LV).

The results of SIMCA analysis (the class distances and classification rule adopted were those described by Section 2.1.3.2, i.e. alternative SIMCA framework) on the unfolded WINE data, are summarized in Table 12.

Mean-centering LV 5/LV 4/LV 3	CL1 South America	CL 2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/69	100/48	100/30
CV	75/74	77/53	46/47
TEST	80/79	43/47	72/47

Auto-scaling LV 3/LV 3/LV 3	CL1 South America	CL 2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/50	100/48	100/73
CV	75/58	77/51	54/80
TEST	90/51	71/65	43/80

Pareto-scaling LV 4/LV 4/LV 3	CL1 South America	CL 2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/65	100/54	100/51
CV	75/66	77/62	69/64
TEST	100/79	57/59	43/65

Table 12: WINE data set – SIMCA on unfolded data. Sensitivity and specificity of the individual class models in Calibration, Cross-Validation and prediction, for the three preprocessing adopted.

Auto-scaling is the best choice in pretreatment for both PLS-DA (in Table 11 the results in bold) and SIMCA classification (results in Table 12).

EVOO data set classification is accomplished through the use of PLS-DA, model and SIMCA as described for WINE data set. Three different preprocessing strategies (mean-centering, auto-scaling and pareto-scaling) are evaluated and compared. The results of PLS-DA modeling are summarized in Table 13 and results of SIMCA approach are in Table 14.

EVOO Taggiasca C1	Mean-centering SENS/SPEC	Auto-scaling SENS/SPEC	Pareto-scaling SENS/SPEC
TRAIN	100/100	100/100	100/100
CV	100/100	100/100	100/100
TEST	100/100	100/100	100/100

Table 13: EVOO data set - PLS-DA classification after unfolding. Sensitivity ans specificity for the three discriminated categories in calibration, cross-validation and prediction, for the pretreatments meancenterin (LV 4), auto-scaling (LV2) and pareto-scaling (LV 4).

EVOO Taggiasca C1	Mean-centering SENS/SPEC	Auto-scaling SENS/SPEC	Pareto-scaling SENS/SPEC
TRAIN	92/100	100/100	100/100
CV	85/100	54/100	92/100
TEST	56/100	56/100	56/100

Table 14: EVOO data set - SIMCA modeling on unfolded matrices. Sensitivity and specificity in calibration, cross-validation and prediction for the pretreatments mean-centering (PC 2), auto-scaling (PC 1) and pareto-scaling (PC 1).

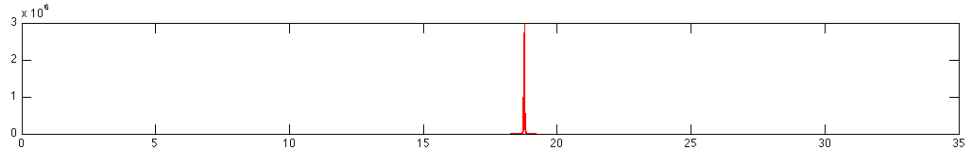
A perfect discrimination between the two classes is obtained, irrespective of the selected pretreatment for the PLS-DA method. However, when considering the model complexity, auto-scaling results in the highest parsimony, as only two latent variables are necessary.
For what concern the SIMCA results the model have a lower sensitivity on the prediction set, but considering that in general, dealing with authenticity issue, the focus is on a single category, and in this case where the aim is to have a model to assess ligurian EVOO, adopting a class modeling approach may be more appropriate.

4.2.2.2 MCR-ALS discrimination

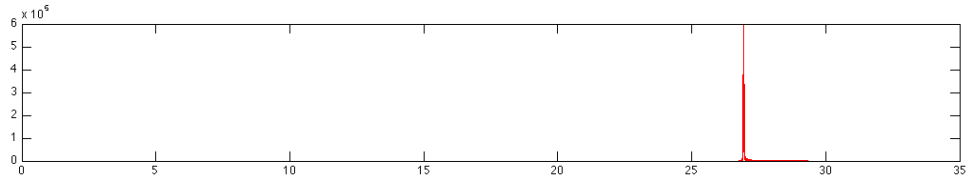
The discriminant approach, which makes use of the ALS-regression, is explained in detail in Section 2.2.2 and some information is given also in Section 4.2.2.2. For WINE and EVOO data set the chromatograms were acquired with a good resolution so the subdivision in windows of the retention times axis was not necessary and MCR was applied to the whole chromatograms, all the peaks were resolved and integrated in a single step obtaining the relative peak areas table.
WINE data set was firstly resolved with twelve components; this model presents an explained variance of 98.58 % and lack of fit of 10.55. This step allows reducing drastically the number of variables from $189*10^3$x200, rank of the multiset obtained from the unfolding of the array, to 70x12 the reduced peak area matrix. In Figure 27 in each subplot are reported the resolved component for all the samples.

4. RESULTS

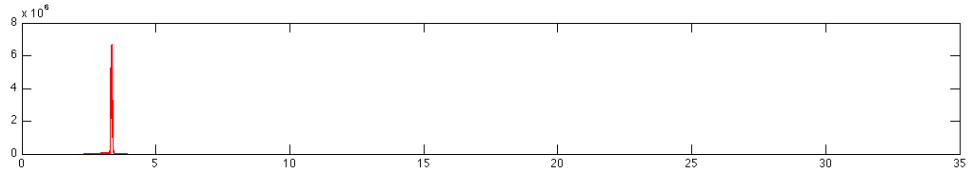
Comp1



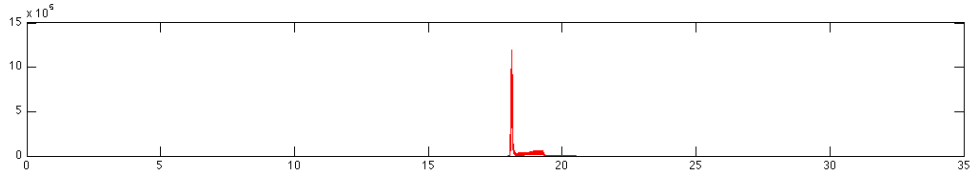
Comp2



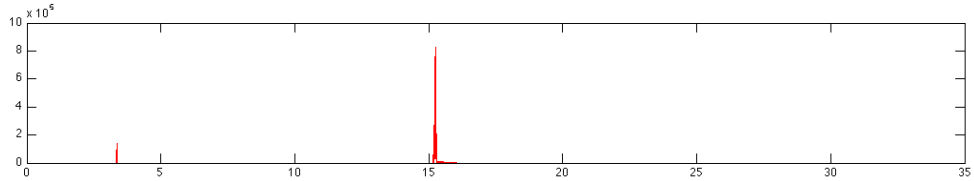
Comp3



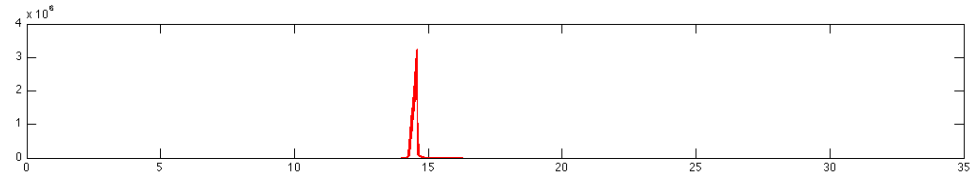
Comp4



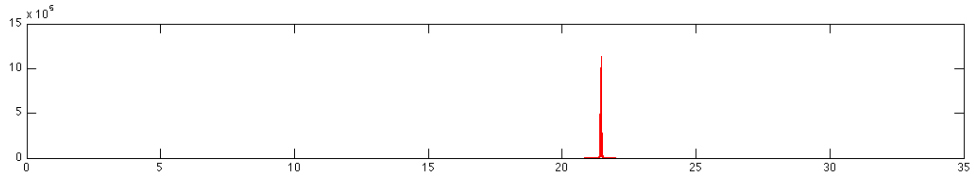
Comp5



Comp6



Comp7



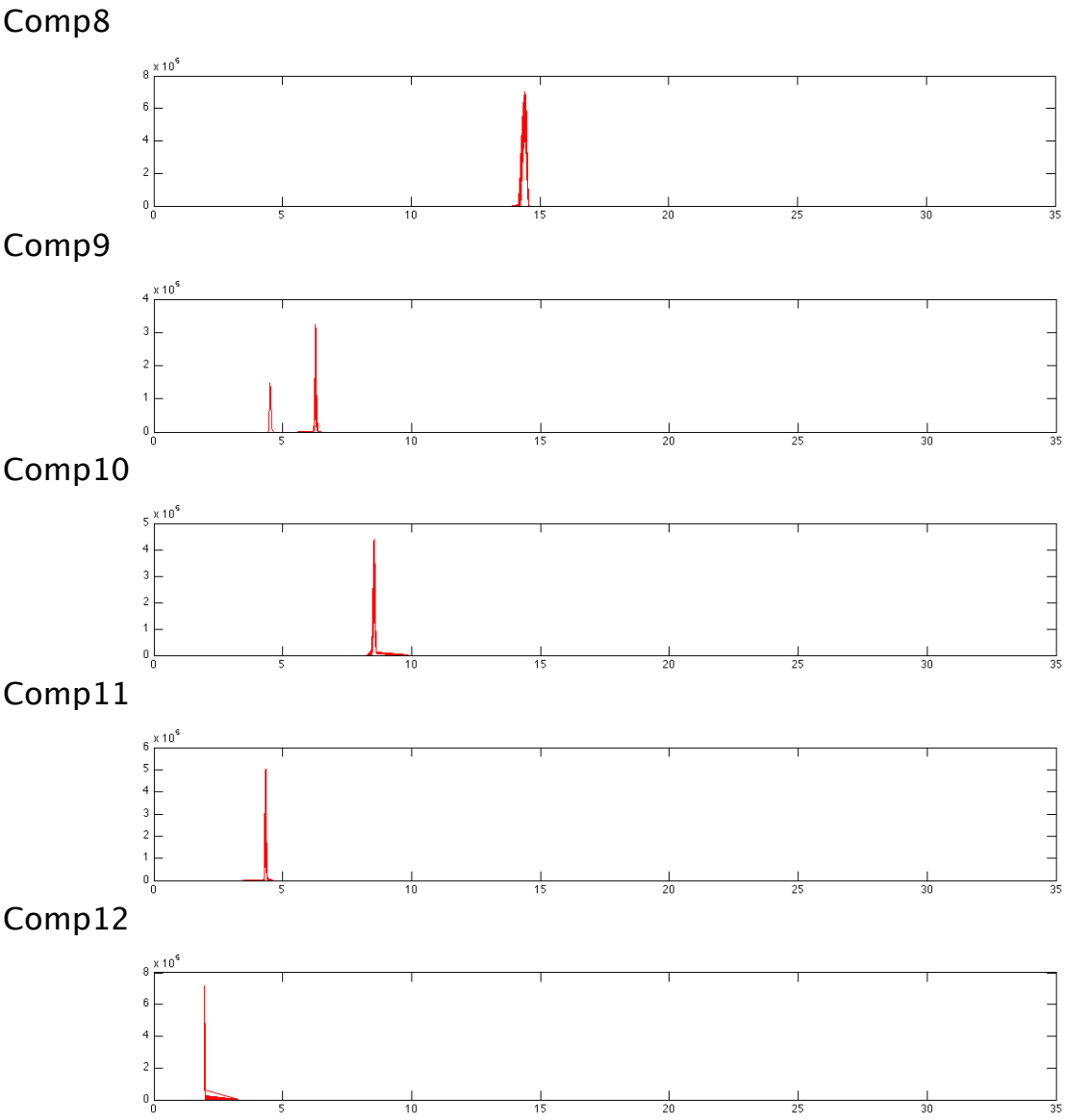


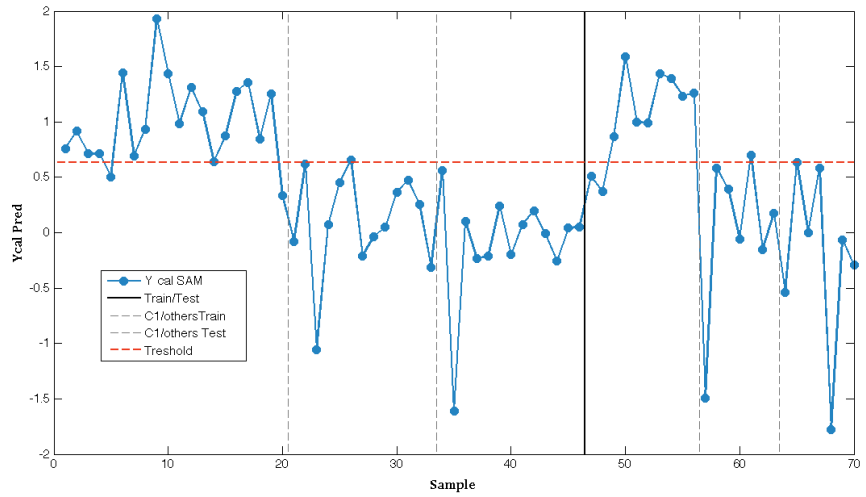
Figure 27: WINE data set – MCR resolution, 12 components resolved, for all the 70 samples.

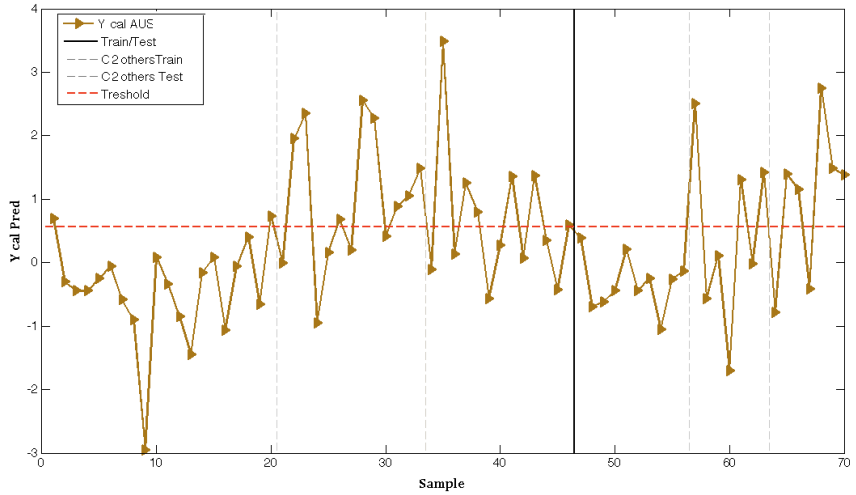
An ALS-regression with four components is calculated, the initial estimation was calculated like explained in Section 4.2.2.2 for Lambrusco data set (average of training samples). Normalization (per rows) and scaling (per columns) were applied as pretreatments. Figure 28 reports the Y predicted values for the three classes and the relative values of sensitivity and specificity are reported in Table 15.

The same fingerprinting table was also analyzed with PLS-DA obtaining the following results reported in Table 17 and shortly illustrated in Figure 29.

Same pretreatments were applied on the data and five latent variables were selected.

Figure 30 of the optimized S^T matrix show a similar fingerprinting for the three classes especially for class 2 and 3. Class 1 has higher values of all the species; in particular 2, 5, 7 and 10 are the variables, which more differentiate this class from the other two. The number of the variable identify also the subplot in which is reported the correspondent chromatographic peak. For these peaks in Figure 31 are reported the relative mass spectra obtained from the MCR resolution. The identification of the compound is not inherent of this work of thesis, but will be pursued in future research.





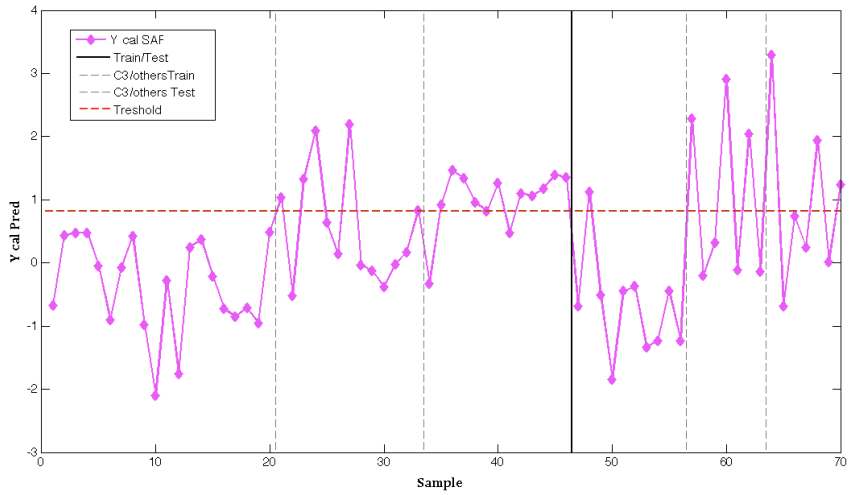


Figure 28: WINE data set – ALS-regression. Y prediction for the class South America, Australia and South Africa, training and test set. Vertical continuous line separate training samples from test samples and the dashed line the samples belonging to different classes. 4 components were selected for each class.

	CL1 South America	CL 2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	90/96	61/76	77/85
TEST	80/93	43/71	57/76

Table 15: WINE data set – ALS-regression. Values of sensitivity and specificity, in calibration and prediction 4 components were selected for each class, 4 latent variables were selected for each class.

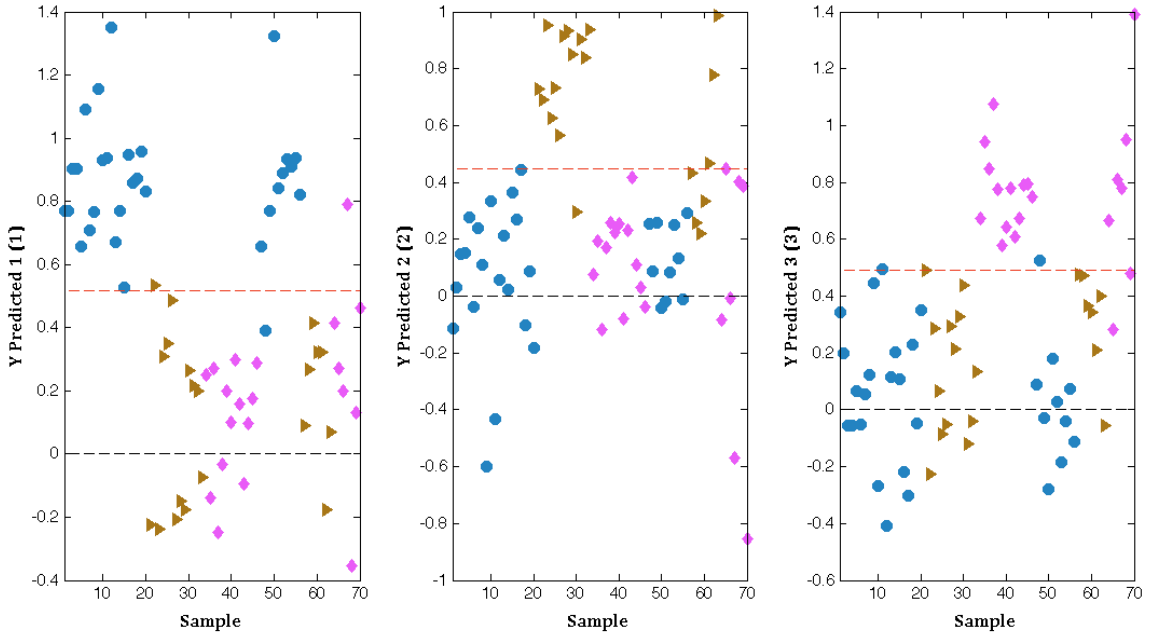


Figure 29: WINE data set – PLS-DA on peak areas. Y predicted values, for the three classes South America, Australia and South Africa, 5 latent variables were selected.

	CL1 South America	CL 2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/96	92/100	100/94
CV	90/92	77/85	85/88
TEST	90/92	43/100	71/94

Table 16: WINE data set – PLS-DA on MCR peak areas. Values of sensitivity and specificity, for the three classes modeled, in calibration and prediction, 5 latent variables were selected.

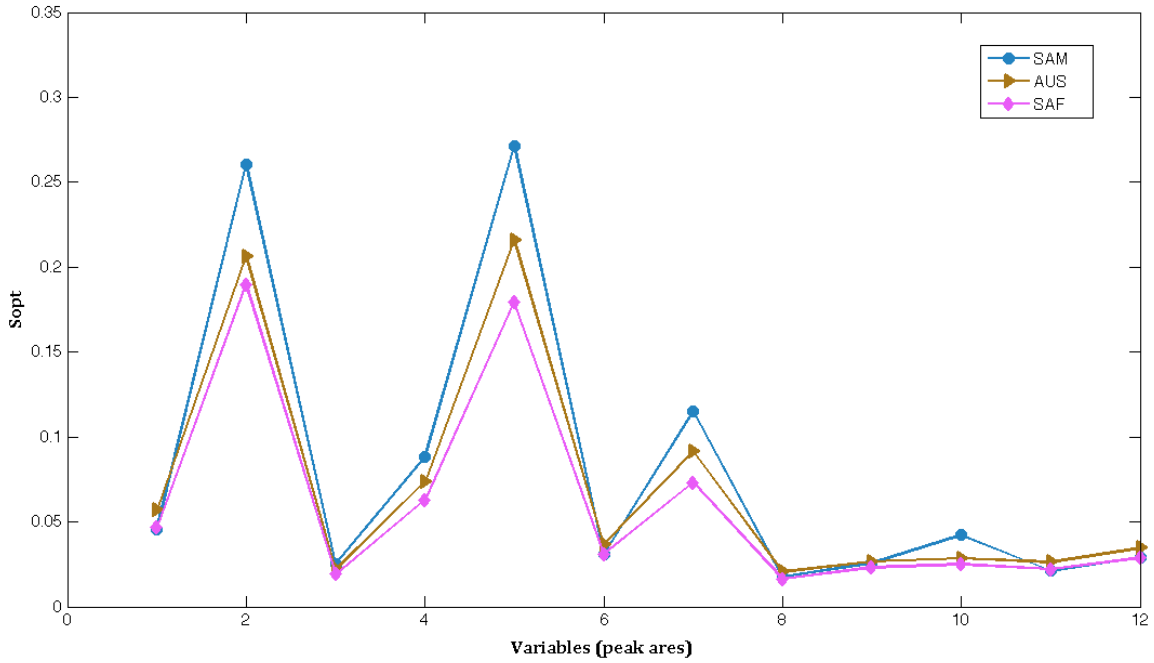


Figure 30: WINE data set – ALS-regression. $\mathbf{S}^T$ optimized matrix, constrained components of the three models.

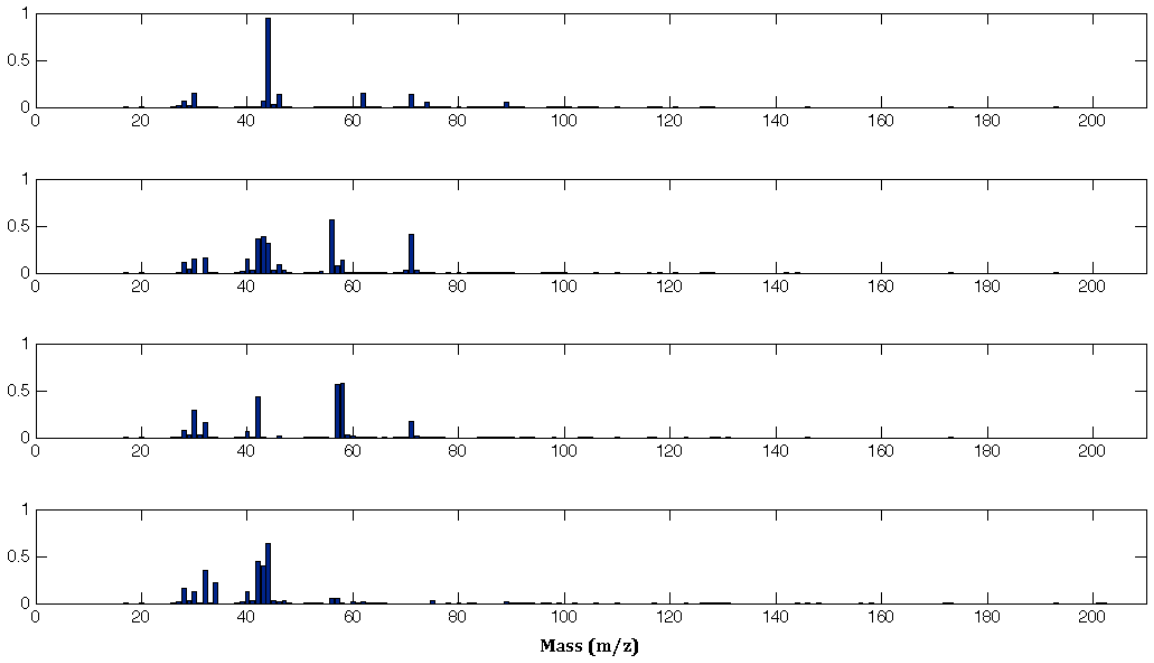


Figure 31: WINE data set – ALS-regression. Spectra profile of the variables 2, 5, 7 and 10 respectively, which differentiate the classes.

4. RESULTS

For EVOO data set the resolution step with MCR gives the values of area of ten chromatographic peaks, with a variance explained of 99.71 % and lack of fit 8.12 %. Matrix dimensionality passes from the multiset of 109728x216 to 72x10 matrix of the peak areas. ALS-regression was calculated modeling the class of Taggiasca olive oil on scales data, selecting five components and building the initial estimation as explained for the other cases.

	SENS/SPEC
TRAIN	92/100
TEST	100/95

Table 17: EVOO data set - ALS-regression. Values of sensitivity and specificity for the modeled class, 5 components were selected.

The class results are reported in Table 17 and illustred in Figure 32. A perfect classification is obtained. PLS-DA model is also calculated on the same data, obtaining perfect classification reported in Table 18 obtained with a for latent variables model.

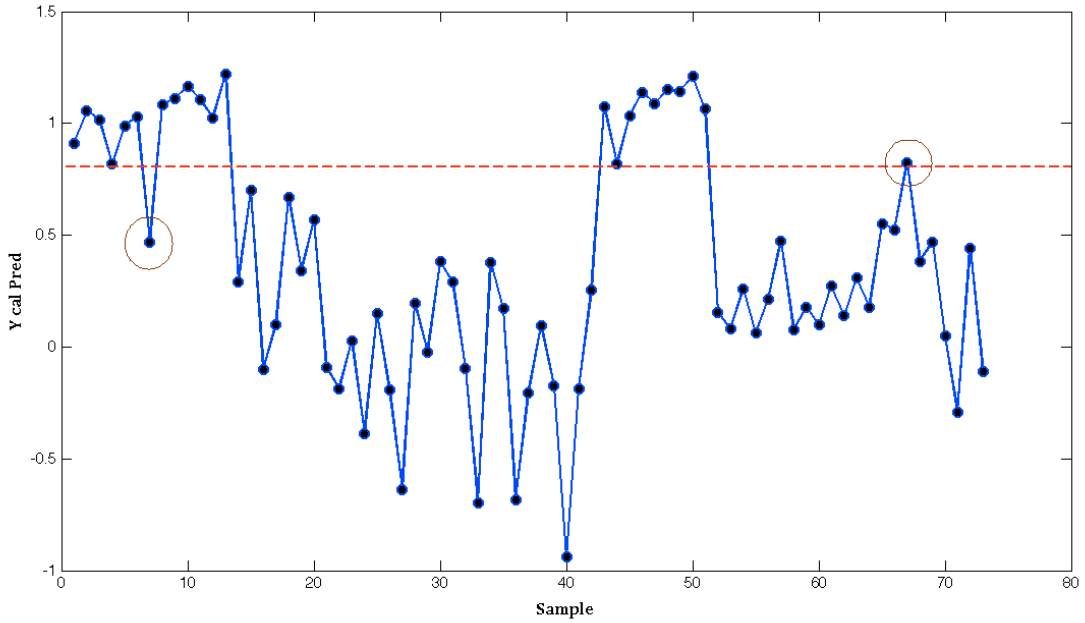


Figure 32: EVOO data set - ALS-regression. Values of Y predicted for the class modeled, training and test set, in evidence the confused samples, 5 components were selected.

	SENS/SPEC
TRAIN	100/100
TEST	100/96

Table 18: EVOO data set – PLS-DA on MCR peak areas. Values of sensitivity and specificity, for the class modeled, in calibration and prediction, 4 latent variables were selected.

4.2.2.3 NSIMCA / NPLS-DA

To choose the best pretreatment for the three-way data arrays of data sets WINE and EVOO an explorative data analysis step has been carried out in order to evaluate the possible pre-treatments and their order of application. In this case, since we are dealing with GC-MS, hence strict trilinearity cannot be assumed, Tucker3 has been preferred as decomposition technique. The different pre-treatments were studied considering the screening Tucker3 models and inspecting the relative scores plot (samples mode), and choosing the combination which gives better separation among the classes. Following centering and scaling combinations were tested:

mean-centering, std-scaling, std-scaling for mode 1, 2 and 3, respectively, (both order: 3 2 1 and 2 3 1 were considered);

mean-centering, pareto-scaling, pareto-scaling; order: 3 2 1 and 2 3 1;

mean-centering, no pretreatment , pareto-scaling; order: 3 2 1 and 2 3 1;

mean-centering, pareto-scaling, no pretreatment; order: 3 2 1 and 2 3 1.

In order to reduce the number of N-SIMCA models to calculate, the data are preliminary screened with a Tucker3 model, the distribution of the training samples has been evaluated separately for each class [15].

The mode 1 scores plot obtained by the Tucker3 screening model is shown in Figure 35. It can be seen that while all the three classes overlap, class 1 has lower dispersion with respect to class 3 and 2, which shows the highest one.

The combinations corresponding to the highest explained variance, for the same sum, are labeled and connected by a dotted line. Considering the variance trend and by inspecting the analogous plots for the other two categories, factors combinations from [3 2 2] to [6 6 6] were explored for

South America, and from [2 3 2] to [7 9 8] for the other two classes. As an example, the results for the Australia category of WINE data set are reported in Figure 34, which report the explained model variance versus the sum of the number of factors in each mode.

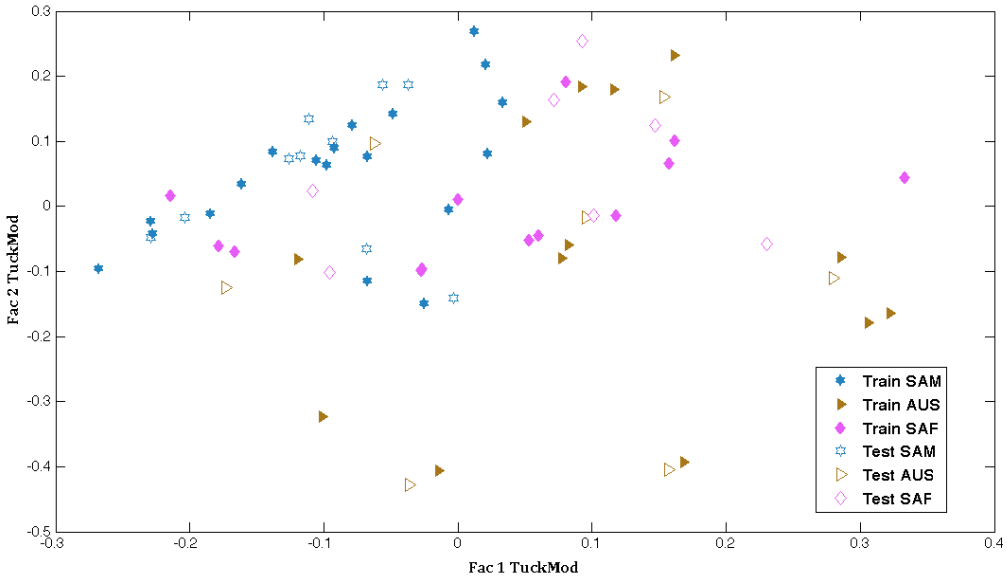


Figure 33: WINE data set – Score plot for factors 1 and 2 of the screening Tucker3 model.

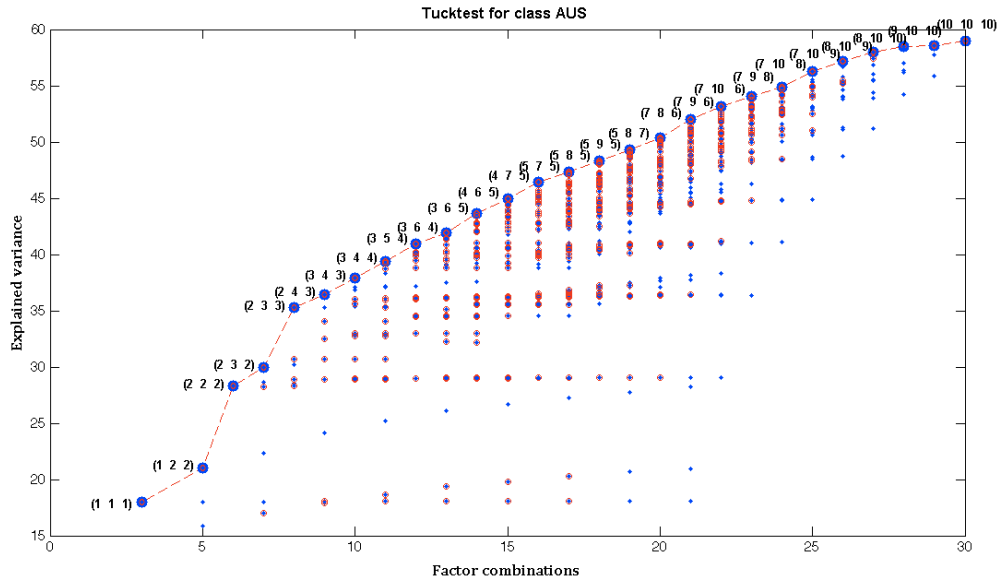


Figure 34: Tucker3 test for Australia class of WINE data set.

For WINE data set the pretreatment: mean-centering (mode 1), pareto-scaling (mode 2) and pareto-scaling (mode 3) in the order 2 3 1, resulted in the best visual appearance with respect to class separation for the plot of mode 1 scores. It can be seen that while all the three classes overlap, class 1 has lower dispersion with respect to class 3 and 2, which shows the highest one.

In assessing what is an appropriate N-SIMCA model, local models are calculated for the data of each class. Once these models are calculated, the best factor combinations for each class (i.e. the N-SIMCA model dimensionality) are chosen considering the best efficiency (geometric mean of sensitivity and specificity for each class) in cross-validation. When evaluating the model corresponding to the highest CV-efficiency in the orig-SIMCA approach it behaves with not sensitivity at all for the test set samples.

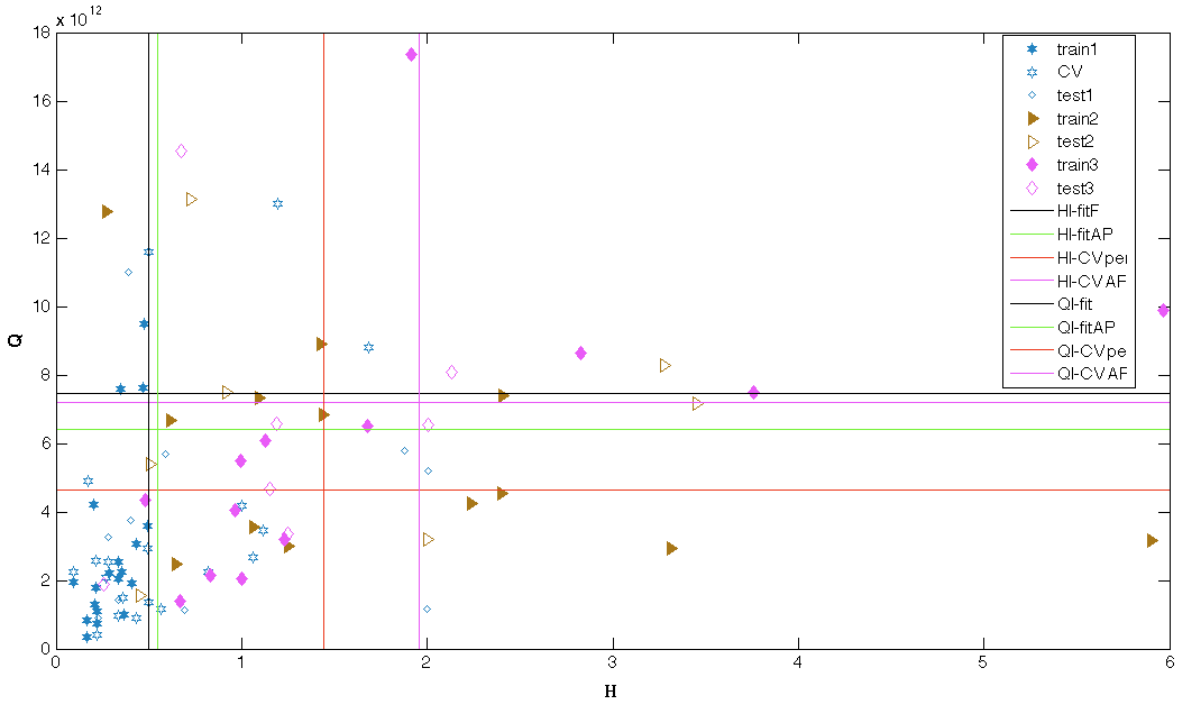
Table 19 reports, for each class, the values of Sensitivity and Specificity in calibration, cross-validation and prediction obtained for the chosen criteria and best combinations. These were selected according to higher efficiency, estimated in 'leave one out' cross validation for fit based classification rules, being the CV based classification criteria already calculated on projection of samples taken out in cross-validation run.

	CL1 South America	CL2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	95/88	100/76	100/70
CV	65/88	54/76	69/70
TEST	60/79	86/71	71/64
	Spec (CL2) 86	Spec (CL1) 60	Spec (CL1) 80
	Spec (CL3) 71	Spec (CL3) 86	Spec (CL2) 43

Table 19: WINE data set – NSIMCA. Sensitivity and specificity (total specificity and respect the single classes) for classification criterion, i.e. Hfit/Qfi, with the following model dimensionality South America [6 3 2], Australia [6 2 4] and South Africa [5 2 4].

As it may be expected from the sample distribution and class overlap observed by explorative Tucker3 model (Figure 33) class 1, showing the lowest dispersion, has the highest specificity value.

In Figure 35, H vs Q plots are shown for each class (training, CV and test samples) with the class acceptance limits calculated according to the different classification criteria in the alternative-SIMCA framework. For all classes, criteria in fit lead to best performing N-SIMCA models, if a best compromise between sensitivity and specificity is sought. Since the classes are rather heterogeneous, cross-validated samples (see legend for the symbol) are rather spread thus class limits in CV are larger and as a consequence specificity decrease. There is little difference when limits are calculated according to Pomerantsev criteria (green lines).



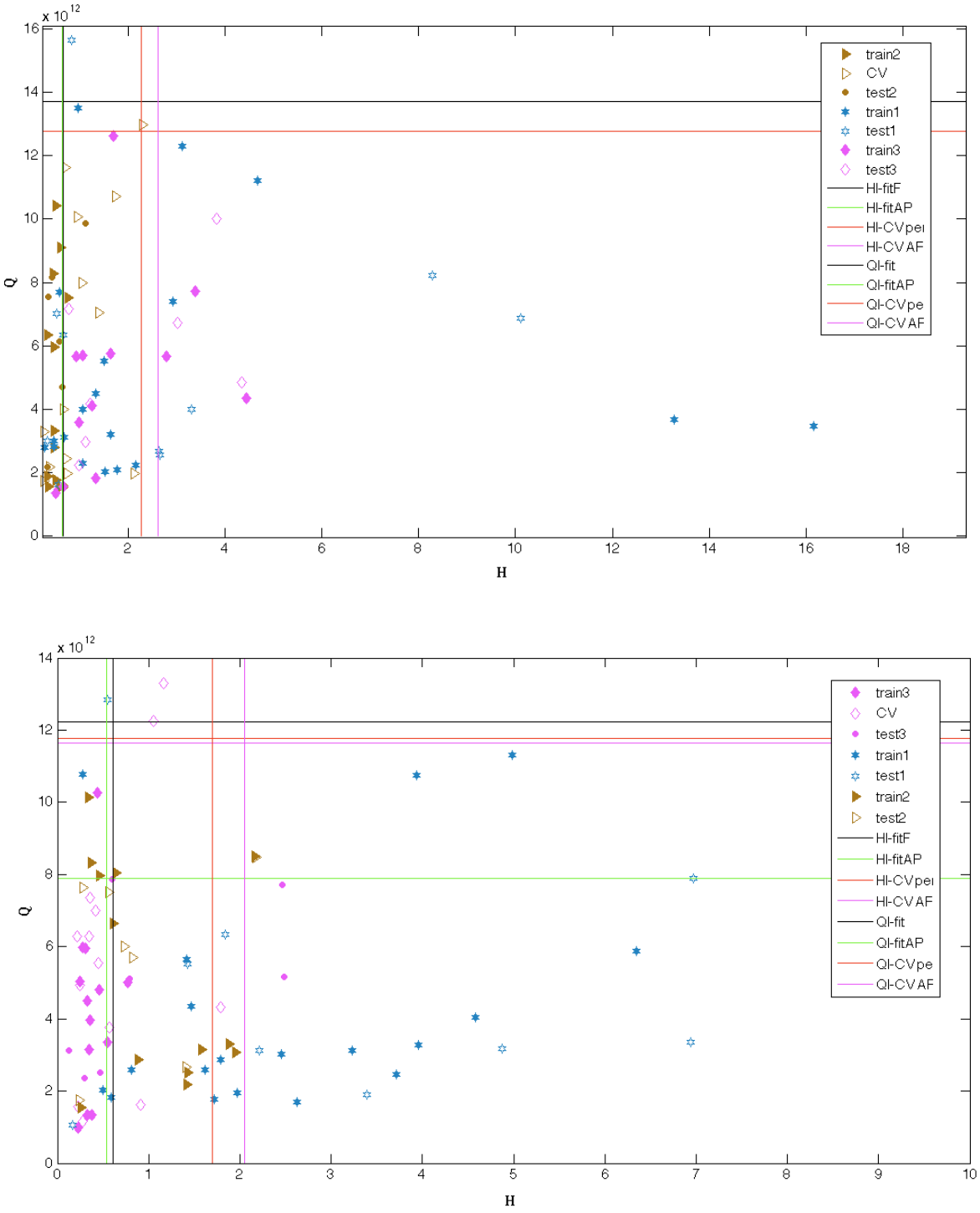


Figure 35: WINE data set – NSIMCA. Q/T plots for the classes South America, Australia, South Africa respectively. Different class boundaries are reported in different colors.

It is possible to inspect which are the chromatographic and mass spectral regions contributing to the category models by looking at the loadings plots of mode 2 and 3, respectively. In Figure 36 are reported the plot of the first factor for mode 2 and 3 for each class. In particular, in the top of the figure the mode 2 (Retention time) factor one loadings for the three classes are superimposed. Some Retention time regions are relevant for all classes but some are specific for specific classes. The mode 3 (masses) factor one loadings plot for each class are shown separately and also here some mass pattern is common and some are specific. A detailed discussion in terms of chemical components corresponding to these spectral features is beyond our present aims but it shows that loadings can be interpreted as in two-way data analysis.

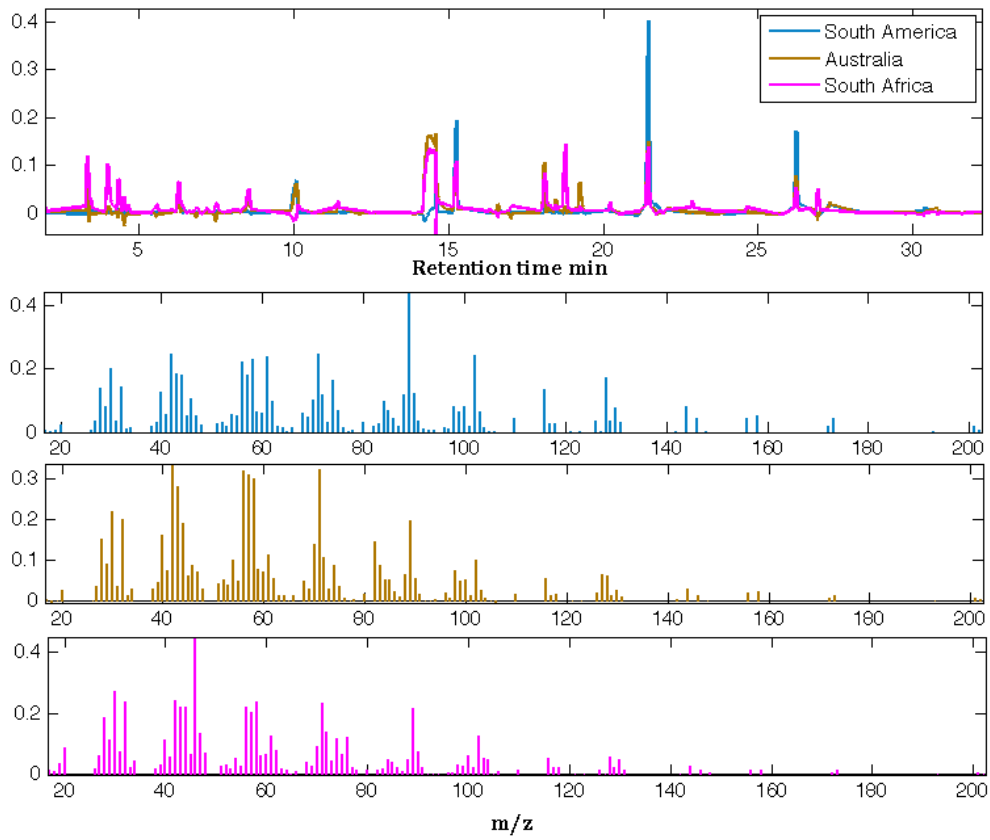


Figure 36: WINE data set – NSIMCA modeling on 3-way data. Loadings of mode 2 and 3 for the three classes. The first component [1 1 1] is reported since most significant

For the EVOO data set, after elimination of variables with almost null variance, both for the chromatographic and mass spectrum directions, the final three-way array as the dimensions 72x1039x161. Most promising results, in term of less overlap of the Liguria with the other classes, are obtained pretreating the data set with mean-centering in the first mode, block-scaling in the second mode and std-scaling in the third mode, with the order 3 2 1. We focused on the factors range [7 8 8] to [10 9 9] considering the explained variance and run N-SIMCA with these settings. Finally, considering the efficiency of the model in cross validation, the best performing classification criteria were relative to alternative-SIMCA framework for both fit and CV criteria; for both criteria the combination leading to higher efficiency in CV, corresponds to factors [9 9 8]. In Table 29 are reported the values of sensitivity and specificity for the Liguria class comparing the criteria in fit and cross-validation for the classification rule based on Pomerantsev limits definition for H and Q statistics, $Hlim_{fit}(AP)$, $Qlim_{fit}(AP)$, best performance is obtained by fit criteria, as also shown in Figure 37.

EVOO Taggiasca	**H_{fit}/Q_{fit} AP**	**H_CV/Q_CV AP**
	SENS/SPEC	SENS/SPEC
TRAIN	92/100	100/96
CV	92/96	92/97
TEST	100/100	100/77

Table 20: EVOO data set – NSIMCA modeling on the 3-way matrix. Values of sensitivity and specificity for the class Liguria: classification criteria 3 and 5 (according to Pomerantesev), both in calibration and cross-validation, [9 9 8] is selected as factor combination.

4. RESULTS

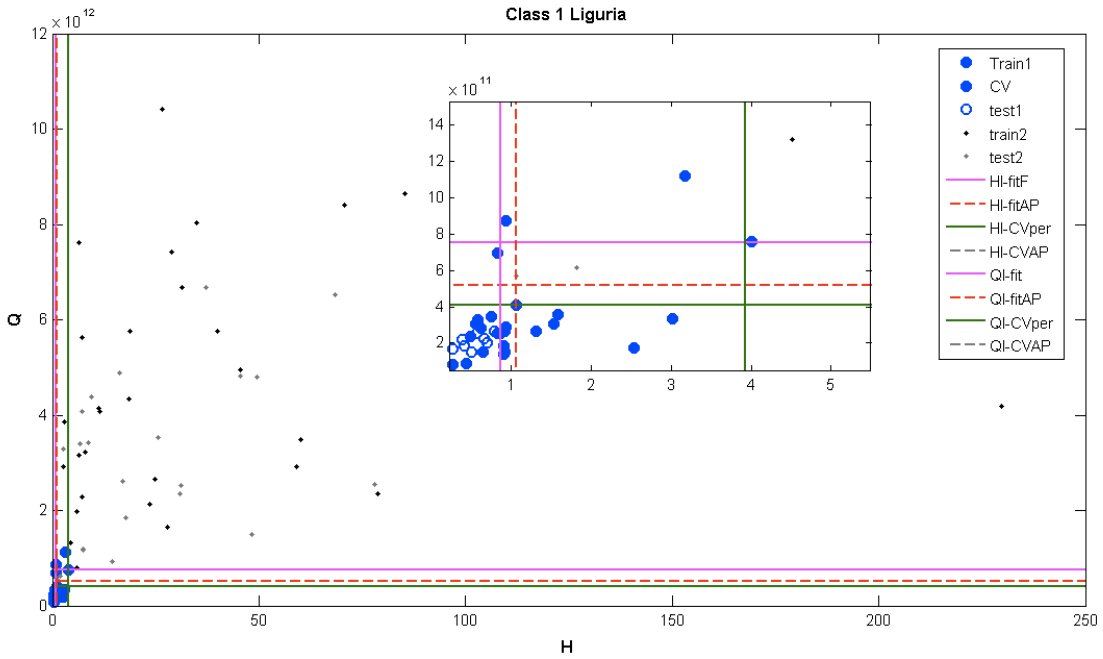


Figure 37: EVOO data set – NSIMCA modeling on the 3-way matrix. Q/T plot for class Liguria, [9 9 8] is selected as factor combination.

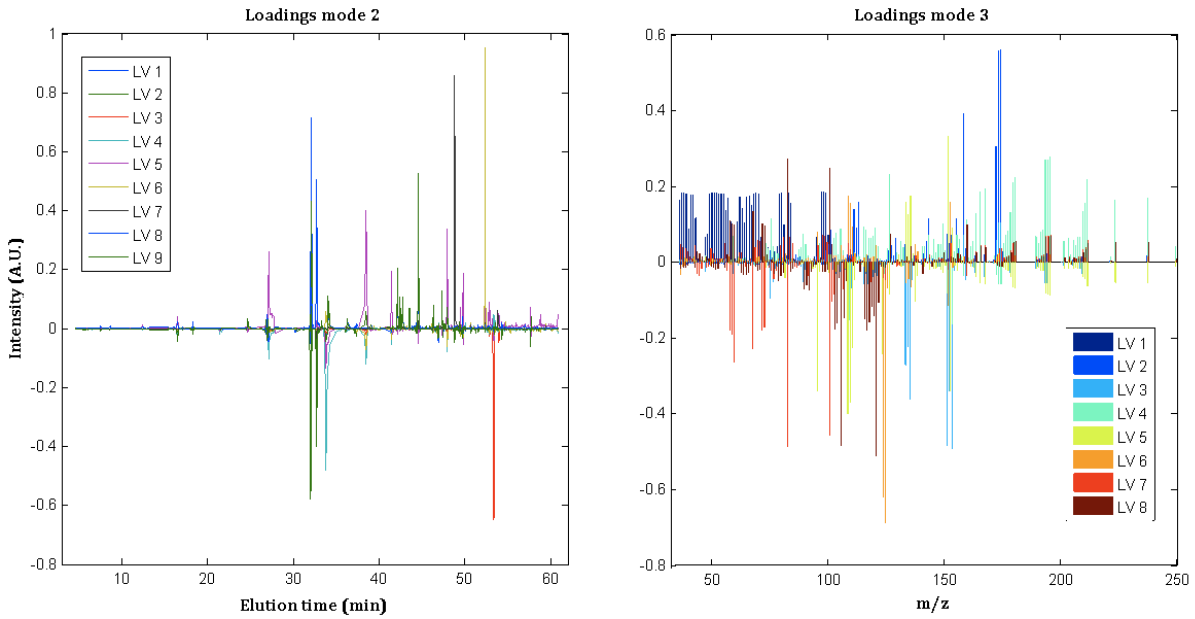


Figure 38: EVOO data set – NSIMCA modeling on the 3-way matrix. Loadings of mode 2 and mode 3 for the modeled class Liguria factors [9 9 8].

NPLS discriminant analysis (NPLS-DA) for WINE data set has been performed using the same data preprocessing described for the NSIMCA models 5 latent

variables (LV) model has been chosen considering the minimum values of the classification error in cross validation (using six-folds venetian-blind cross validation) The objects are assigned to the class for which the corresponding value of the dummy y-variable, as predicted by the model is higher. Table 21 reports the classification rates for training set, in cross validation and for the test set for the three classes.

Pareto-scaling	CL1 South America	CL 2 Australia	CL3 South Africa
	SENS/SPEC	SENS/SPEC	SENS/SPEC
TRAIN	100/96	100/100	92/100
CV	70/92	85/88	85/76
TEST	90/86	86/100	86/94

Table 21: WINE data set – NPLSDA. Sensitivity and Specificity for the three discriminated categories in calibration, cross-Validation and External Validation, 5 latent variables were selected.

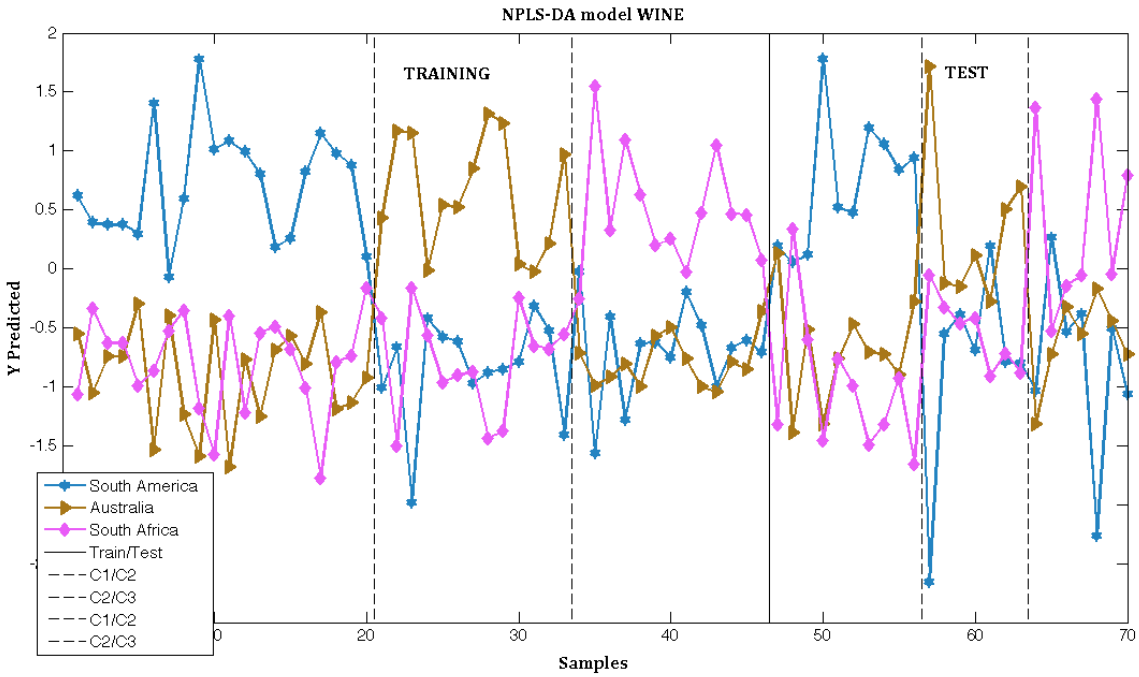


Figure 39: WINE data set – NPLS-DA. Y values recalculated and predicted for the training and test sets, respectively the predicted dummy y-variable versus samples for each class are plotted with different style. Vertical continuous line separate training samples from test samples and the dashed line the samples belonging to different classes. Results are satisfactory despite lower specificity, considering the starting class overlaps.

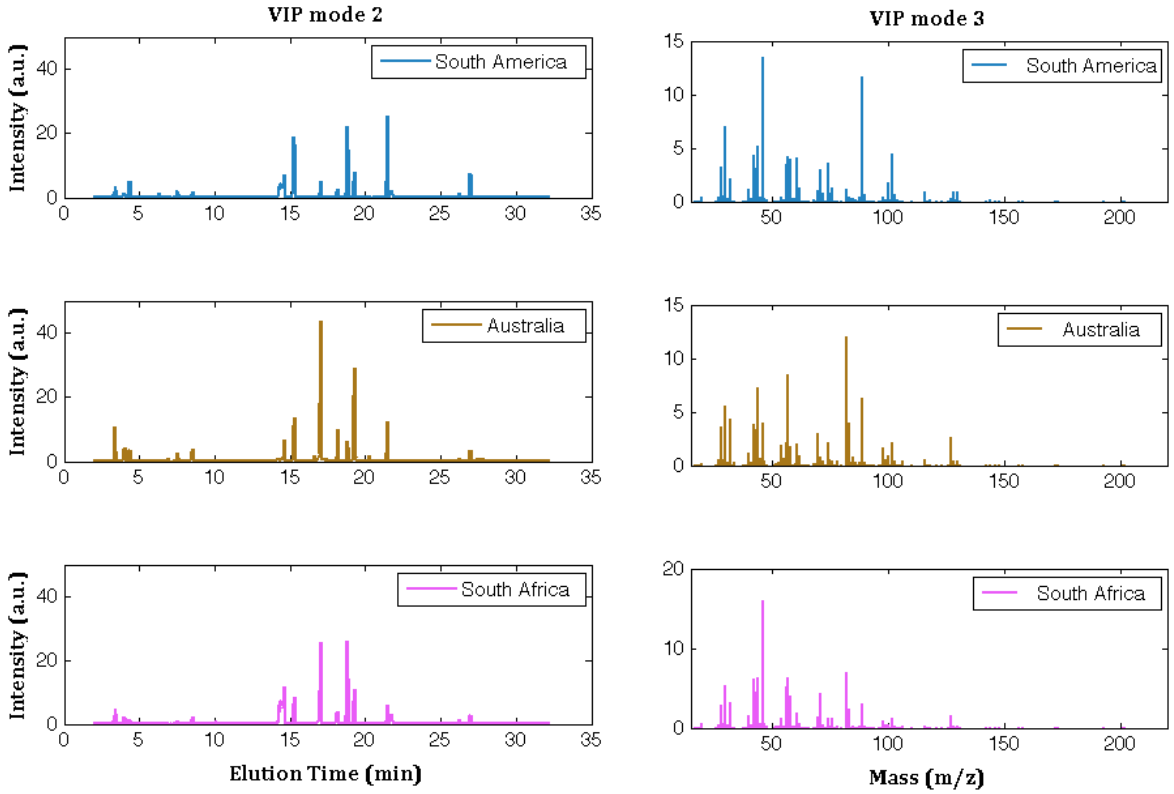


Figure 40: WINE data set – NPLS-DA. VIP scores of the two modes respectively for class South America, Australia and South Africa.

The NPLS-DA model dimensionality for EVOO data set has been chosen according to minimum classification error in 5-fold venetian blind cross-validation, 3 latent variables are selected. Results obtained for the training and test set are reported in Table 22 and Figure 41 that shows the values of Y predicted for each class. The model shows perfect sensitivity and specificity in calibration and prediction. The most influent signal regions are highlighted in the VIP plot, Figure 42. As an example the relevant contribution of retention time about 46 minutes together with mass patterns 53, 68, 79, 93, 107, 121 can be attributed to limonene, by interpreting jointly the VIP plots in the two modes some information can be gathered on the volatile pattern of the EVOO.

EVOO	SENS/SPEC
TRAIN	100/100
CV	100/100
TEST	100/100

Table 22: EVOO data set – NPLS-DA. Sensitivity and specificity in calibration, cross validation and prediction for the Taggiasca class, 3 latent variables were selected.

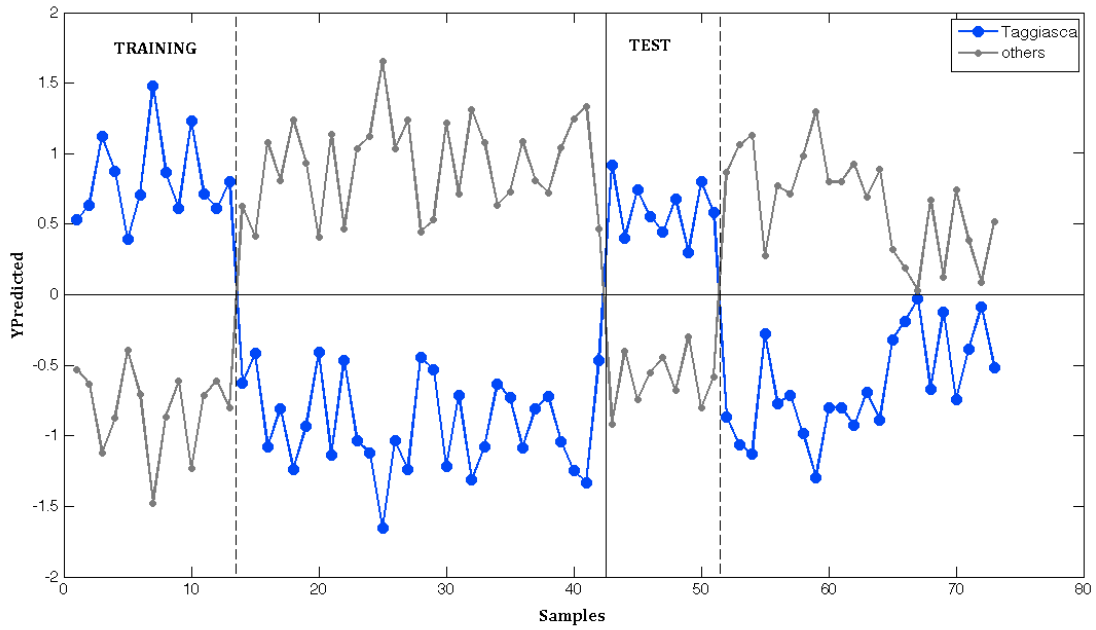


Figure 41: EVOO data set – NPLS-DA. Values of predicted Y, training and test set for Liguria EVOO and the others, 3 latent variables were selected.

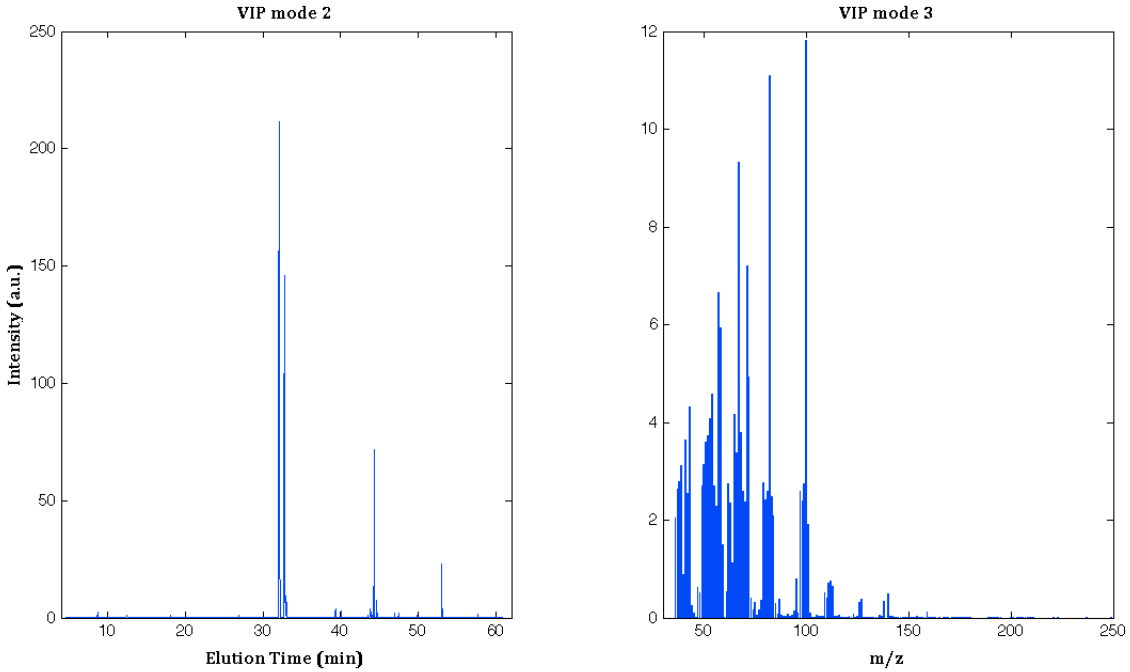


Figure 42: EVOO data set – NPLS-DA. VIP for the two modes, elution mode and spectra mode.

4.3 Final Remarks

4.3.1 Lambrusco data set classification

For what concern the Lambrusco data set, resolution of chromatographic peaks, with multivariate curve resolution – alternating least squares, has been demonstrated to be a successful approach, which allows optimizing the time and costs of analysis. Both qualitative information, related to the resolved spectra of compounds, and quantitative information, content of analytes for which reference standards are present, or in general represented by the possibility to integrate the chromatographic peaks in the resolved elution profiles, can be obtained in a unique step. Thus, it became possible to perform calibration also when the acquired signals are taken as a fingerprint in a faster and less resolved experimental conditions. Limits of detection and quantification obtained were comparable with those present in literature, where a similar instrumentation is used. This kind of resolution defined 'a posteriori' allows to optimize the time analysis without losing the information, but providing a further knowledge of the system, the possibility of separating qualitative information of spectra and quantitative information of peak areas. The versatility of the methods, which permits to apply specific constraints to solve specific problems, gives the possibility of obtaining optimum results also in situation of strict coelution. Summarizing a rapid and simple HPLC-DAD method is developed for the determination of phenolic compounds in wine also in presence of overlapping interferences. The method was also verified with an external test set acquired further in time. The use of resolved peak areas instead of the entire chromatographic run as fingerprinting information permitted to circumvent the fundamental problem of chromatographic alignment, always present when full chromatographic traces are used in authentication analysis.

Similar results are obtained also using PARAFAC to resolve the same elution windows, however in this case alignment in the chromatographic direction was needed, and anyway in some elution windows it is evident that the MCR

resolution permits to resolve more accurately the chromatographic peaks (Quercetin and p-Coumaric acid cases) thanks to the application of selectivity constraint.

As far as the discrimination of the three Lambrusco varieties is concerned, the results obtained by the different classification approaches are compared in Table 23 in terms of sensitivity and specificity. For each model are reported the dimensionality (number of latent variables or components), the applied pretreatments, and the dimensionality of the data set. The analysis of the fingerprinting information allows to distinguish among the three DOP of Lambrusco analyzed. The multiset approach presents a great vantage with respect to the multiway one, it permits the resolution and the analysis of chromatographic data, also if problems of misalignment are present; this is a common situation when one deals with real samples of complex nature (like in food analysis). The class modeling method, represented by NSIMCA, does not perform a good separation among the classes; similar values were obtained also calculating a SIMCA model on the matrix of peak areas (data not reported). On the contrary all the other discriminant approaches shows good results. This may be due to the fact that the classes overlap with no clear separation in latent variables space. Good results are obtained with discriminant approaches on peak areas, such as PLS-DA and ALS-regression showing that the resolution step maintains all the relevant information. The direct application of n-way discriminant analysis (NPLS-DA) furnishes results of similar quality. The Sorbara class is better modeled (higher sensitivity) by PLS-DA on the peak areas data set; in this case the resolution step gave a sensible improvement in the discrimination of this class, identifying the lower content of the compounds in this classes, like shown in the regression vector plot, for which the most important variables for the discrimination (VIP) presents negative values.

Lambrusco data set **original dimension** **152x1211x106**	CL1 Grasparossa	CL 2 Salamino	CL3 Sorbara
	SENS/SPEC	SENS/SPEC	SENS/SPEC
PLS-DA on peak areas 7LV (norm aut) [152x39]			
TRAIN	100/100	100/98	97/99
TEST	89/97	89/85	93/92
ALS-regression on peak areas LV* 5 5 5 (norm scal) [152x39]			
TRAIN	97/100	100/98	96/100
TEST	95/97	100/77	80/95
NPLS-DA on 3-way data LV 8 (aut psc psc, 321) [152x707x106]			
TRAIN	100/98	97/98	96/100
TEST	89/94	84/94	80/99
NSIMCA on 3-way data LV 8 8 8 (aut psc psc, 321) [152x707x106]			
TRAIN	100/51	92/17	97/71
TEST	82/56	66/23	50/86

Table 23: Lambrusco data set – Comparison of models obtained for each classification approach. LV* refers to MCR components.

4.3.2 WINE data set classification

The results obtained by the different classification approaches on the WINE data set are summarized in Table 24. The three modeled classes shows in the screening analysis an elevate overlap among the classes Australia and South Africa, while the class South America is distinguishable from the other two. In this respect, the values of sensitivity and specificity, obtained from the different approaches, reflect the initial situation observed. The performance of the discriminant methods is comparable for PLS-DA on unfolded data and NPLS-DA; the NPLS-DA method is more parsimonious, 5 LV instead of 7, with respect to unfolded PLS-DA, and gives better prediction results for the first class, i.e. the most homogenous one. Moreover the VIP scores of chromatographic and m/z mode are easily interpretable, while this will be much more troublesome in the case of unfolded data. In this data set, discriminant methods based on peak areas, both PLS-DA and ALS-regression, shows worst results for the discrimination of the classes Australia and South Africa. This means that some information useful for discrimination

is lost during the resolution step. This may be due to the lower chemical knowledge on this data set in term of identified compounds that prevented the use of selectivity constraints. However, also different criteria are used fro the assignation in PLS-DA on un folded data the higher value of Y predicted is chosen for the class assignation, while in PLS-DA on peak areas and ALS-regression the sample is assigned to a class if the values of Y predicted is higher of a threshold value (like implemented in the PLS-toolbox).

WINE data set **Original dimensions** **70x2700x200**	CL1 South America	CL 2 Australia	CL3 South Africa
PLS-DA on unfolded data **7 LV (aut) [70x21350]**			
TRAIN	100/100	100/100	100/100
TEST	80/93	86/100	86/88
SIMCA on unfolded data **LV 3 3 3 (aut) [70x21350]**			
TRAIN	100/50	100/48	100/73
TEST	90/51	71/65	43/80
ALS-regression on peak areas **LV* 4 4 4 (norm scal) [70x12]**			
TRAIN	90/96	61/76	77/85
TEST	80/93	43/71	57/76
PLS-DA on peak areas **LV 5 (norm aut) [70x12]**			
TRAIN	100/96	92/100	100/94
TEST	90/92	43/100	71/94
NPLS-DA on 3-way data **LV 5 (mnc psc psc) [70x684x99]**			
TRAIN	100/100	100/100	92/96
TEST	90/86	86/100	86/94
NSIMCA on 3-way data LV **[6 3 2] [6 2 4] [5 2 4] (mnc psc psc 321) [70x684x99]**			
TRAIN	95/88	100/76	100/70
TEST	60/79	86/71	71/64
	Spec (CL2) 86	Spec (CL1) 60	Spec (CL1) 80
	Spec (CL3) 71	Spec (CL3) 86	Spec (CL2) 43

Table 24: WINE data set – Comparison of the models obtained for each classification approach. LV* refers to MCR components.

Comparing the resolved components with the VIP scores of NPLS-DA model it can be observed that peak corresponding to the components number 2 4 7 (it is reported the relative mass spectrum is also reported) are the ones showing an higher values for the class South America in the VIP plot of the NPLS-DA model and the same compounds are also highlighted in the loadings of the second mode of NSIMCA model.
Regarding the class modeling approaches (SIMCA on unfolded data and NSIMCA) in this case, they suffer the excessive dispersion of the three classes, when compared with the discriminant ones that result in this case more performing. Models calculated with N-SIMCA with respect to the unfolded SIMCA, present improved values of specificity for all classes in calibration and prediction, equal sensitivity for the class 2 and improved sensitivity for class 3 in prediction; only in the case of class 1 test set sensitivity is lower. Moreover, N-SIMCA models always show a better balance between sensitivity and specificity.

4.3.3 EVOO data set

Comparing all the results for EVOO data set, reported in Table 25 it is clear that optimal discrimination of ligurian EVOO is accomplished by all approaches except for class modeling on unfolded data (SIMCA) that furnishes a model with no sensitivity for the test set. This highlights how beneficial it is to take into account the multi-way nature of the data array extracting the relevant information to model the class feature. In authentication tasks this asymmetric, one class, situation is very common and the possibility to use a class modeling approach is very relevant and it became feasible by N-SIMCA.
The discriminant method proposed based on MCR resolution and ALS-regression, gave in this case very close results moreover offering the advantage of resolution.

EVOO **Taggiasca C1** **Original dimension** **72x1514x216**	CL1 SENS/SPEC
PLS-DA on unfolded data **LV 2 (aut) [72x49644]**	
TRAIN	100/100
TEST	100/100
SIMCA on unfolded data **LV 1 (psc) [72x49644]**	
TRAIN	100/100
TEST	56/100
ALS-regression on peak areas **LV* 5 5 5 (scal) [72x10]**	
TRAIN	92/100
TEST	100/95
PLS-DA on peak areas **LV 4 (scal) [72x10]**	
TRAIN	100/100
TEST	100/96
NPLS-DA on 3-way data **LV 3 (mnc blk scal 321) [72x1039x161]**	
TRAIN	100/100
TEST	100/100
NSIMCA on 3-way data **Fac [9 9 8] (mnc blk scal 321) [72x1039x161]**	
TRAIN	92/100
TEST	100/100

Table 25: EVOO data set – Comparison of models obtained for each class classification approach. LV* refers to MCR components.

In conclusion, multi-way models even when perform similarly in classification, as in the case of discriminant approach, offer a better interpretability of the results, and in the case of class modeling perform better highlighting that taking into account the multi-way structure allows improvement of class characterization. To reduce the data set by MCR and then applying classification tools in general furnishes good results, in particular if PLS-DA is applied.

In general, MCR-regression presents many characteristics in common with the other discriminant techniques, such as the unique assignation of a

sample to a single class based on its predicted value, but at the same time the information on variable influence in the discrimination model is related to one class at time, and this feature make the approach also similar to a class modeling procedure, at least on interpretative ground.

Clearly the innovation consists in the possibility of constraining the output, in particular the matrix of spectra the application of non-negativity for the spectral mode, which is represented by peak areas.

References

[1] V. Cheynier; *Phenolic compounds: from plants to foods*. Phytochemistry Reviews (2012).

[2] M. I. Alarcón Flores, R. Romero-González, A. Garrido Frenich, J. L. Martínez Vidal; *Analysis of phenolic compounds in olive oil by solid-phase extraction and ultra high performance liquid chromatography–tandem mass spectrometry*. Food Chemistry (2012) 134, 4 2465–2472.

[3] J. J. J. van der Hooft, M. Akermi, F. Y. Ünlü, V. Mihaleva, V. Gomez Roldan, R. J. Bino, R. C. H. de Vos, J. Vervoort; *Structural Annotation and Elucidation of Conjugated Phenolic Compounds in Black, Green, and White Tea Extracts*. Journal of Agricultural and Food Chemistry (2012) 60, 36 8841–8850.

[4] C. Chang-Sheng, Z. Di, W. Ya-Qiang, L. Peng-Min, M. Feng-Wang; *Effects of fruit bagging on the contents of phenolic compounds in the peel and flesh of 'Golden Delicious', 'Red Delicious', and 'Royal Gala' apples*. Scientia Horticulturae (2012), 142, 13 68–73.

[5] B. Sun, A. C. Neves, T. A. Fernandes, A. L. Fernandes, N. Mateus, V. De Freitas, C. Leandro, M. I. Spranger; *Evolution of Phenolic Composition of Red Wine during Vinification and Storage and Its Contribution to Wine Sensory Properties and Antioxidant Activity*. Journal of Agricultural and Food Chemistry (2011) 59, 12 6550–6557.

[6] F. Buiarelli, F. Coccioli, R. Jasionowska, M. Merolle, A. Terracciano; *Determination of some stilbenes in Italian wines by liquid chromatography/tandem mass spectrometry*. Rapid Communications in Mass Spectrometry (2007) 21 2955–2964.

[7] F. Buiarelli, F. Coccioli, R. Jasionowska, M. Merolle, A. Terracciano; *Chromatographic analysis of trans resveratrol in Italian wines: comparison between FL, UV and MS detection*. Chromatographia (2006) 64 475–481.

[8] E. Gárcia-Beneytez, F. Cabello, E. Revilla; *Analysis of grape and wine anthocyanins by HPLC–MS*. Journal of Agricultural and Food Chemistry (2003) 51 5622–5629.

[9] B. Sun, A. C. Neves, T. A. Fernandes, A. L. Fernandes, N. Mateus, V. De Freitas, C. Leandro, M. I. Spranger; *Evolution of Phenolic Composition of Red*

Wine during Vinification and Storage and Its Contribution to Wine Sensory Properties and Antioxidant Activity. Journal of Agricultural and Food Chemistry (2011) 59, 12 6550–6557.

[10] HPLC stessa strumentazione dall'articolo

[11] G. L. La Torre, M. Saitta, F. Vilasi, T. Pellicanò, G. Dugo; *Direct determination of phenolic compounds in Sicilian wines by liquid chromatography with PDA and MS detection*. Food Chemistry (2006) 94 640–650.

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

[13] 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.

[14] 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.

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

[16] E. Salvatore, M. Bevilacqua, R. Bro, F. Marini, M. Cocchi; *Classification methods for multiway arrays as a basic tool for food PDO authentication*. In M. De La Guardia and A. Gonzalvez Illueca (Eds.) *Food protected designation of origin: methodologies & applications.* Wilson & Wilson's Comprehensive Analytical Chemistry, Volume 60, 2013. Elsevier B.V., in press.