UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

School of Graduate Studies
Multiscale Modelling, Computational Simulations and Characterization in Material and Life Sciences

Data Fusion to integrate data of different nature in food authenticity

Tutors:

PhD candidate: **Dr. Marina Cocchi**

Dr. Michele Silvestri **Prof. Andrea Marchetti**

School director: Prof. Ledi Menabue

XXVI cycle (2011 – 2013)

Table of Contents

"The true sign of intelligence is not knowledge but imagination"

A. Einstein

CHAPTER 1

Introduction

1.1. We are living in a data fused world

The interpretation of the world around us, the way by which we can assess the goodness of a dish, the beauty of a sunset or the appreciation for the latest movie is a brain assisted operation that comes from the acquisition of the information that the five senses give. Reality, as we know it, could be defined in a trivial way as a data fusion process, which allows evaluating, interacting with, operating in the world around. Tasting a good wine without smelling its aroma or appreciating its color does not deny the possibility to drink it, but certainly limits the quality of the experience. Human brain may be considered the best paradigm to explain and understand how different sources of information (smell, hearing, taste, touch, sight) can be fused together in order to achieve a unified picture, as comprehensive as possible, of the multisensor environment.

These considerations, which can appear on one hand trivial if referred to the natural way in which we are accustomed to perceive reality, are not straightforward if other kinds of problems are considered.

In many fields of science, in particular in chemistry, the holistic approach is often overlooked and the reductionistic one, which provides the separation and the study of simpler processes and constituents able to describe the phenomena under investigation, is preferred.

Data fusion methodologies may be read as tools able to operate in the most natural way, the holistic one, on data of different nature.

In recent years, the interest of the scientific community for "data fusion" has increased, and data fusion or data integration is establishing itself as a new discipline. Numerous definitions have been drawn since the fields of application span from sensor networks to robotics, from military applications to video and image processing to name a few, one of the most inclusive was reported by the Defence Science and Technology Organization (DSTO, 1994): *"Data fusion is a multilevel, multifaceted process*

dealing with the automatic detection, association, correlation, estimation, and combination of data and information from single and multiple sources."

The guidance principle, characteristic of all the data fusion frameworks, which can be transferred when dealing with chemical information, is that the simultaneous investigation of a given set of samples based on a multi-platform approach, hence considering data arising from different analytical techniques, gives the possibility to evaluate the complementary and/or verifying information and to enhance both sample description and correlation between the variables of different sources.

In the cases in which a deep *a priori* knowledge about the involved variables is not present, helpful information about the correlation/similarities among the sources of different nature can be emphasized and not only the descriptive capability related to the samples caused by the investigation of different features extracted from different analytical platforms.

In many cases, the characterization of the samples is a difficult task to face, in particular when complex matrices are involved. Hence, in fields such as metabolomics and biomarker discovery [1-3], toxicity studies [4-5], where very complex signals/data are considered, the data fusion approach is emerging as feasible methodology to extract the concealed information from difficult data structures.

When considering the field of food analysis [6], which usually contemplates the presence of complex matrices, interferences, additional uncontrolled variability due to biological, chemical and physical transformations, the necessity to unveil and understand the salient features of the samples able to characterize quality, authenticity, etc., renders the adoption of a data fusion strategy in many cases mandatory. In particular, in order to accomplish an extensive characterization, the analytical platforms usually adopted span from chromatographic to spectroscopic techniques, from elemental compositions to nuclear magnetic resonance signals to name a few.

Especially when authenticity and geographical traceability issues are considered, the samples under investigation can be very similar, hence, the predominant part of the information is common. The connotative features, able to distinguish the investigated

samples, are often veiled by the not informative and redundant part of the data. Therefore, the joint multi-platform evaluation of the problem enhances the possibility to discover minority information which can appear confused when the traditional reductionistic approach is adopted.

Since nature is not only multivariate but also *"data fused"*, without over presumption, the well-known statement of Harald Martens can be extended saying that:

"to have a lot of good data of different nature, without fully and jointly interpreting them, is like having a symphony orchestra in which each instrument is played one at a time"

1.2. Data fusion methodologies

Hundreds of works are present in literature covering many fields of application. Data fusion is emerging from the late 80s, when studies were deepened in sectors such as military applications [7], robotics [8], pattern recognition [9] and others.

Due to the versatility of the data fusion approach, the used inputs can vary in terms of number and typology of data or blocks of data. Therefore, different strategies of data fusion can be formulated depending on the goal to pursue and, more relevant, on the data to be fused. As example, citing the most common application of the data fusion methodologies, the involved variables can be EM or acoustic radiation used in military application for geolocation, tracking and targeting, or temperature, pressure and other used in robotics.

Since the complexity and the variability of fields in which data fusion methodologies are applied, several classifications of these techniques can be found [10]. The most widespread one, common also in chemometrics, distinguishes three main classes: a) low-level or concatenated data fusion, b) mid-level data fusion c) high-level or decision data fusion.

The classification is based on the level at which data are associated together, even if, since many facets can be considered, a rigorous assignment at a given class is not always possible in all the cases.

In this section, only the key concepts regarding the different data fusion typologies will be reported. For further information related to pros and cons, limitations and strength points, the next chapters will treat in detail the data fusion methodologies both from a theoretical and applicative point of view.

In low-level data fusion, the different information sources are arranged at the so call "data level", or rather, merging them as raw data without other kind of manipulation on the separate data sets. In many cases, low-level data fusion is a generalization of multivariate data analysis [11-12]. In fact, the data are arranged in a unique dataset as first step. Since the data of different nature to organize for the data fusion application can have different measurement units, magnitude and variability, one of the most important issue to face is the way in which variables, or blocks of variables, are scaled in the merging procedure. Several criteria are used in order to scale the different blocks of information; one of the most common is to associate to each block of variables the same variance. This procedure is known as block scaling. The low-level data fusion approach is feasible when the blocks of data present good similarities in terms of number of variables, since the scaling procedure can considerably affect the data structure, when the dimensionality of the various blocks is very different.

Mid-level data fusion methodologies are that kind of approaches oriented to the fusion of data after a preliminary data analysis. For this reason, mid-level data fusion is also known as features level data fusion. In mid-level data fusion, the original variables are separately investigated by means of different typologies of multivariate data analysis, such as PCA, multivariate curve resolution, PARAFAC or by the simple extraction of the most important variables via variable selection methods. The extracted features are then merged together using, also in this case, different scaling procedures depending on the goal to pursue and the characteristics of the data to be merged. When considering mid-level fusion based on the concatenation of the scores extracted by

decomposition methods, an important aspect to consider is that the number of variables involved in the fusion step is quite smaller with respect to the original one; hence, the variance related to each new variable is higher and in general is possible to handle a small subset of variables without a leak of information.

In the last years, high-level data fusion is emerging in fields such us multi-sensor analysis. High-level methodologies can be considered as operations that seek to process local decisions from multiple sensors/variables to achieve a joint decision. In chemometrics, high-level data fusion is not yet diffuse; only few applications can be found facing classification problems. The responses obtained by classification methods such as PLS-DA or SIMCA on the different blocks of information are combined together in order to create a new data set used to compute a final classification model. The advantage of this approach is that the fused based model outperforms the best separate model, or at least, results are the same [13]. The other side of the coin is that, using a high-level approach, the analysis of the fused results does not bring the information about the characteristic of the original data sets, or in other words, the characterization of the investigated samples is only maintained in the lower level of data analysis. Hence, even if the classification results can be improved, it is quite difficult to understand backwards which are the variables that most of all are involved in the differentiation of the samples belonging to the different modelled classes.

In the last years, in parallel with the development of the data fusion methodologies listed above, other techniques, called coupled data fusion methodologies, are establishing.

The main difference with respect to the classical data fusion approach is that the different sources of information are simultaneously evaluated by means of joint models. In coupled algorithms, a single model is computed, considering simultaneously all the sources of information. Therefore, several components, in analogy with the generality of decomposition methods, are extracted. Some of these components share the information of the samples common to the different blocks of

data, i.e. redundant or replicated among blocks, others, can be considered as peculiar of just one or few blocks. Depending on the data input structure, different algorithms can be applied. In the easiest case, the different sources of information are organized in a data table, or two-way data set, hence bilinear decomposition methods can be applied imposing an objective function in order to extract at the same time shared and peculiar components. A lot of coupled data fusion techniques are reported in literature, considering two-way datasets: SUM-PCA [14], SCA-P [15], multiple factor analysis [16], and STATIS [17], to name a few. In the recent years, new algorithms based on partial least squares [18], orthogonal-PLS [19-20], or kernel based [21] and kernel OPLS based [20] were introduced with the aim to work with multiblock data.

The only published method able to work with different order data presenting shared and unshared components, coupled-matrix-tensor-factorization CMTF [22] and ACMTF-OPT [23], will be discussed in detail in the next sections and an application will be shown in order to better highlight the potentiality of this methodology.

1.3. Data fusion in food authentication

In the last decades, different episodes have contributed to decrease the consumers' confidence towards food products and to draw the attention to what we eat, to the provenance and the production methods. The importance of words like quality and safety are increasing at many levels of the social perception, starting from consumers up to regulatory institutions. The European Community faced these topics introducing the product designations [24-29], with the goal to assess the links between the products and the territories of origin as an added value and to safeguard peculiar products made using traditional methods.

In council regulation 510/2006, the reported description of the designation of origin and geographical indication, related to PDO (protected designation of origin) and PGI (protected geographical indication), states:

"*The two types of geographical description are different. A PDO (Protected Designation of Origin) covers the term used to describe foodstuff which are produced, processed and prepared in a given geographical area using recognized know-how (such as Mozzarella di Bufala Campana). A PGI indicates a link with the area in at least one of the stages of production, processing or preparation (such as Turrón de Alicante). The link with the area is therefore stronger for PDOs."*

Traditional specialty guaranteed (TSG) is defined in Council Regulation 509/2006 as:

"*An agricultural product intended for human consumption or foodstuff with a traditional composition, or produced according to a traditional production method may become a traditional speciality guaranteed (TSG). This possibility encourages the diversification of agricultural production and has positive consequences in several areas*".

In addition, the Council Regulation 178/2002 sets out the basis for a new control method by defining the terms of traceability and production chain traceability. *"Traceability is defined as 'the ability to trace and follow a food, feed, food-producing animal or substance intended to be, or expected to be, incorporated into a food or feed, through all the stages of production, processing and distribution."*

To control all the mandatory requirements, necessary for the assignment of the quality marks, paper based traceability systems are usually adopted. These systems have proven to be, in some cases, not sufficient to protect the real authenticity of foodstuff and, in recent years, striking episodes came to the fore (blue Mozzarella, horse meat in pasta sauce, melamine in milk scandals, just to cite a few).

Beside this, studies aimed at obtaining warranties about the authenticity of food are establishing by means of the effort of the scientific community.

As an example, the geographical traceability issues were investigated also through European community funded projects [48] and several potential markers emerged as suitable to assess the provenance of a product.

In particular, the geographical traceability indicators can be classified in two main categories:

- ✓ ***primary or direct indicators*** are able to directly link some chemical characteristics of the territory of origin with the same ones measured in the final products. The most used are: the elemental composition, the stable isotope ratios of light elements (H, C, N, O and S) and the ratio of the relative isotopic abundances of radiogenic heavy elements (Sr, Pb, Nd);
- ✓ ***secondary or indirect indicators*** are variables related to compositional and chemical/physical characteristics of the food and to the transformation process. By means of an extensive characterization of the matrix, based on secondary indicators, it is possible to identify products with the same origin distinguishing them to all others similar. Spectroscopic and spectrometric techniques, such as IR, NMR, GC-MS, LC-MS, etc., are the most used to obtain signals related to the composition of the food matrix.

The characterization of food matrices and, in particular, their geographical traceability are affected by many sources of variability: even if a foodstuff is produced in a restricted area and in agreement with the production method defined by the production regulation, the seasonal variability and the local differences, present in the soils in which the products are cultivated, can give to the final products different characteristics that make difficult the determination of the origin, using a single or a restricted pool of indicators.

Even if a boundless literature is present regarding studies about geographical traceability using direct and indirect indicators on different products such as wine [30-31], meat [32-33], honey [34-35], dairy products [36], etc…, the data fusion methodologies are not so established, in particular, when primary and secondary indicators are evaluated at the same time [37-47].

In the food analysis context, due to the complexity of the investigated matrices and of some analytical outputs (such as many indirect indicators), it is not always possible to unveil the responses able to answer to the addressed problem from the analysis

performed on a single technique. For these reasons, the adoption of data fusion techniques is suited to face food analysis issues, by strengthening the assumptions obtained from the analysis of single techniques or achieving new ones, which could remain hidden when using a reductionistic approach. Moreover, the complete characterization of foodstuffs is possible only if the determination of several characteristics, related to the chemical composition, organoleptic properties or physical/chemical features, all of them based on different kinds of analysis, is performed.

As stated before, the development of geographical traceability models focused on the discrimination of products of different origin is a problematic issue to face.

TRACE, [48] a large-scale traceability study was conducted within the VII European Framework with the goal to assess the potentiality of both primary and secondary indicators for the geographical traceability of commodities, such as water, olive oil, meat, honey and cereals. TRACE conclusions highlighted the need to operate on more defined and restricted geographical areas, in order to assess the effectiveness of the geographical markers.

At national level, the AGER project [49], in which part of my work of thesis is involved, was designed with the objective to deeply investigate a restricted area and only two kinds of oenological products: Lambrusco wines and TRENTODOC wines.

The aim of the AGER project is to achieve all the information related to the production chain, starting from the characterization of the soils in which the vines are cultivated, through raw materials and intermediates of production, concluding with final products.

In next sections, applications of data fusion methodologies for the characterization of soils and for the varietal determination of Lambrusco wines, concerning the AGER project, will be presented in detail.

1.4. Aims and outline of the thesis

As briefly introduced in the sections before, data fusion methodologies cover a large number of fields of application, dealing with variables of different nature and with multiple goals to pursue.

In order to give to the reader the basis to understand the potentiality and applicability of data fusion techniques, the Thesis is articulated in a way that, first of all, the general concepts are introduced and then applications are reported, starting from simple to more complex cases.

In **Chapter 2,** the used multivariate data analysis methods are described. The first two sections describe decomposition methods, where particular attention is given to the argumentation of Multivariate Curve Resolution. The last section of the chapter concerns classification methods, applicable to two-way and multi-way datasets.

Chapter 3 is dedicated to the detailed description of data fusion methodologies. Each data fusion approach is examined in a way that pros and cons are highlighted, motivating the choice of the most suitable one, depending on the problem to face.

In **Chapter 4,** many of the described data fusion methodologies are presented in form of applications in the food analysis context, in order to handle with the common problems and the way to solve them. The first application described is related to a case of study aimed at developing a classification model for the discrimination of different commercial classes of Aceto Balsamico di Modena, using a low-level data fusion approach.

In the second section of the chapter, a mid-level data fusion approach is proposed for the characterization of soil samples of the District of Modena. The description of the variability of soil samples using different direct and indirect indicators gives the basis for the development of geographical traceability studies.

Varietal traceability of Lambrusco wines, a typical P.D.O. products of the District of Modena, is presented in third section, facing the problem with the use of both mid and high level data fusion approaches.

Prior to the conclusion and final remarks, in the last section of **Chapter 4**, an application of coupled data fusion based on Coupled Matrix Tensor Factorization focused on the description of Lambrusco wine samples by means of a multi-platform determination is proposed.

1.5. References

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

[2] J. Forshed, R. Stolt, H. Idborg, S.V. Jacobsson, *" Enhanced multivariate analysis by correlation scaling and fusion of LC/MS and 1H-NMR data"*, Chemometrics and Intelligent Laboratory Systems, 2007, 85, 179-185

[3] A. Smolinska, J.M. Posma, L. Blanchet, K.A.M. Ampt, A. Attali, T.Tuinstra, T.Luider, M.Doskocz, P.J. Michiels, F.C. Girard, L.M.C. Buydens, S.S. Wijmenga, *"Simoultaneous analysis of plasma and CSF by NMR and hierarchical models fusion"*, Analytical and Bioanalytical Chemistry, 2012, 403, 947-959

[4] R. Dyck, M.S. Islam, A. Zargar, A. Mohapatra, R.Sadiq, *"Application of data fusion in human health risk assessment for hydrocarbon mixtures on contaminated sites"*, Toxicology, 2013, 313(2-3),160-73

[5] W.Z. Chen, Y.F. Cui, Y. Y. He, Y. Yu, J. Galvin, Y.M. Hussaini, Y. Xiao, *"Application of Dempster-Shafer theory in dose response outcome analysis"*, Physics in medicine and Biology, 2012, 57(17), 5575-5585

[6] Edited by F. Marini, *"Chemometrics in food analysis",* Data Handling in Science and Technology (28), 2013

[7] D.L. Hall, J. Llinas, "*An introduction to multisensor data fusion, proceedings of the IEEE*", 85 (1997) 6–23.

[8] M. Brady, "*Special issue on sensor data fusion—foreword, International Journal of Robotics Research",* 7 (1988) 2–4.

[9] L.I. Kuncheva, "*Combining Pattern Classifiers"*, Wiley, Hoboken, New Jersey, 2004.

[10] D.Hall, J. Llinas, "*Hanbook of Multisensor Data Fusion",* 2001, CRC Press

[11] M. Cocchi, P. Lambertini, D. Manzini, A. Marchetti, A. Ulrici, "*Determination of Carboxylic Acids in Vinegars and in Aceto Balsamico Tradizionale di Modena by HPLC and GC Methods"*, Journal of Agricultural and Food Chemistry, 2002, 50(19), 5255-5261

[12] C. Durante, M. Cocchi, M. Grandi, A. Marchetti, R. Bro, "*Application of N-PLS to gas chromatographic and sensory data of traditional balsamic vinegars of Modena"*, Chemometrics and Intelligent Laboratory Systems, 2006, 83(1), 54-65

[13] S.Vao , "*A Fusion Method That Perfroms Bettcr Than Best Sensor*", USION'98 Internation Conference, 19-26

[14] A. K. Smilde, J. A. Westerhuis, S. de Jong, "*A framework for sequential multiblock component methods*", Journal of Chemometrics 2003, 17, 323-337

[15] H. A. Kiers, J. M. ten Berge, "*Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure*", British Journal of Mathematical and Statistical Psychology, 1994, 47, 109-126

[16] B. Escofier, J. Pagès, "*Analyses factorielles simples et multiples",* 3rd edition. Paris: Dunod; 1998

[17] H. L'Hermier des Plantes, B. Thièbaut, "*Etude de la pluviosité au moyen de la méthode S.T.A.T.I.S*",Revue de Statistique Appliquée 1977, 25, 57-81

[18] L.E. Wangen, B.R. Kowalski, "*A multiblock partial least squares algorithm for investigating complex chemical systems"*, Journal of Chemometrics, 1989, 3, 3–20

[19] L. Eriksson, M. Toft, E. Johansson, S. Wold, J. Trygg, "*Separating Y-predictive and Y-orthogonal variation in multi-block spectral data"* , Journal of Chemometrics, 2006, 20, 352–361.

[20] J. Boccard, D. N. Rutledge, "*A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion",* Analytica Chimica Acta, 2013, 769, 30– 39

[21] A. Smolinska, L. Blanchet, L. Coulier, K.A.M. Ampt, T. Luider, R.Q. Hintzen, S.S. Wijmenga, L.M.C. Buydens, "*Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis",* PLoS ONE, 2012, 7, e38163.

[22] E. Acar, T. G. Kolda, D. M. Dunlavy, "*All-at-once Optimization for Coupled Matrix and Tensor Factorizations*", KDD Workshop on Mining and Learning with Graphs, 2011

[23] E. Acar, A. J. Lawaetz, M. A. Rasmussen, and R. Bro, "*Structure-Revealing Data Fusion Model with Applications in Metabolomics",* Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC'13), 2013, 6023-6026

[24] EC Regulation 2081/1992, 14 July 1992

[25] EC Regulation 2082/1992, 14 July 1992

[26] EC Regulation 509/2006, 20 March 2006, on agricultural products and foodstuffs as traditional specialties guaranteed. This replaces EC 2082/1992

[27] EC Regulation 510/2006, 20 March 2006, on the protection of geographical indications and designations of origin for agricultural products and foodstuffs. This replaces EC 2081/1992

[28] EC Regulation 1898/2006, 14 December 2006, laying down detailed rules of implementation of Council Regulation (EC) 510/2006

[29] EC Regulation 628/2008, 2 July 2008, amending EC 1898/2006

[30] I. Arvanitoyannis, M. Katsota, E. Psarra, E. Soufleros and S. Kallithraka, "*Application of quality control methods for assessing wine authenticity: use of multivariate analysis (chemometrics)*" Trends in Food Science and Technology, vol. 10, pp. 321 - 336, 1999.

[31] M. Suhaj and M. Koreňovska, *"Application of elemental analysis for identification of wine origin"* Acta Alimentaria, vol. 34, pp. 393 - 401, 2005

[32] F. Schwägele, *"Traceability from a European perspective"* Meat Science., vol. 71, pp. 164 - 173, 2005

[33] K. Heaton, S. Kelly, J. Hoogewerff and M. Woolfe, *"Verifying the geographical origin of beef: the application of multi-element isotope and trace element analysis"* Food Chemistry., vol. 107, pp. 506 - 515, 2008

[34] E. Anklam, *"A review of the analytical methods to determine the geographical and botanical origin of honey"* Food Chemistry, vol. 63, pp. 549 - 562, 1998

[35] L. Cuevas-Glory, J. Pino, L. Santiago and E. Sauri-Duch, *"A review of the volatile analytical methods for determining the botanical origin of honey"* Food Chemistry, vol. 103, pp. 1032 - 1043, 2007

[36] R. Karoui and J. De Baerdemaeker, "*A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products"* Food Chemistry., vol. 102, pp. 621 - 640, 2007.

[37] C. Alamprese, M. Casale, N. Sinelli, S. Lanteri, E. Casiraghi, *"Detection of minced beef adulteration with turkey meat by UVvis, NIR and MIR spectroscopy"* Food Science and Technology, 2013 , 53 , 225-232

[38] E. Karoui, P. Botosoa, *"Characterisation of Emmental Cheeses Within Different Brand Products by Combining Infrared and Fluorescence Spectroscopies",* Food and Bioprocess Technology, 2013, 6(9), 2365-2375

[39] L. Vera, L. Acena, J. Guasch, R. Boque, M. Mestres, O. Busto, *"Discrimination and sensory description of beers through data fusion"*, Talanta, 2011, 87, 136,142

[40] C. Apetrei, I.M. Apetrei, S. Villanueva, J.A. de Saja, F. Gutierrez-Rosales, M.L. Rodriguez-Mendez, *"Combination of an e-nose, and e-tongue and an e-eye for the characterization of olive oils with different degree of bitterness"*, Analytica Chimica Acta, 2010, 663(1), 91-97

[41][M. Casale, C. Casolino, P. Oliveri, M. Foria, *"The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil"*, Food Chemistry, 2010, 118, 163-170

[42] S. Buratti, D. Ballabio, S. Benedetti, M.S. Cosio, *"Prediction of Italian red wine sensorial descriptors from electronic nose, electronic tongue and spectrophotometric measurements by means of Genetic Algorithm regression models"*, Food Chemistry, 2007, 100(1), 211-218

[43] M. Casale, C. Armanino, C. Casolino, M. Forina, *"Combining information from headspace mass spectrometry and visible spectroscopy in the classification of the Ligurian olive oils"*, Analytica Chimica Acta, 2007, 589(1), 89-95

[44] G. Downey, P. McIntyre, A.N. Davies, *"Geographic classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near-infrared spectroscopic data"*, Applied Spectroscopy, 2003, 57(2), 158-163

[45] J. Boccard, D.N. Rutledge, *"A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion"*, Analytica Chimica Acta, 2013, 769, 30-39

[46] D. Asfaha, C. Quétel, F. Thomas, M. Horacek, B. Wimmer, G. Heiss, C. Dekant, P. Deters-Itzelsberger, S. Hoelzl, S. Rummel, C. Brach-Papa, M. Van Bocxstaele, E. Jamin, M. Baxter, K. Heinrich, S. Kelly, D. Bertoldi, L. Bontempo, F. Camin, R. Larcher, M. Perini, A. Rossmann, A. Schellenberg, C. Schlicht, H. Froeschl, J. Hoogewerff and H. Ueckermann, *"Combining isotopic signatures of n(87Sr)/n(86Sr)*

and light stabel elements (C, N, O, S) with multi-elemental profiling for the authentication of provenance of European cereal samples", Journal of Cereal Science, vol. 53, pp. 170 - 177, 2011

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

[48] TRACE: *"Tracing food commodities in Europe"*, N° FP6-2003-FOOD-2-A 006942, 2005 – 2009

[49] AGER, Agroalimentare e Ricerca. *"New analytical methodologies for varietal and geographical traceability of oenological products"*, contract n. 2011 – 0285

CHAPTER 2

Multivariate Data Analysis Methods

2.1 Exploratory Data Analysis

When an experiment is set up, the goals that every chemist tries to achieve are obtaining the information able to solve the investigated problem and unveiling peculiar characteristics about samples. When dealing with data arising from analytical determination, often provided in form of data tables in which the number of involved variables is much higher than the number of observations, the helpful information can be hidden by interferences, noise, redundancy and unnecessary information.

Moreover, the fast technologic development of last decades allowed making available analytical solution capable to collect huge amounts of data in few minutes, if not seconds. High throughput methodologies (spectroscopy, chromatography, mass-spectrometry to name a few) are nowadays completely established and often, customized instrumentations are present on the market in order to face specific issues. Despite that, the combination of sample pretreatments and dedicated analytical platforms is not sufficient to extract from the collected data the necessary information for which the experiment was conducted, especially when complex matrices as food samples are investigated.

Even if, the one variable a time approach is still present in many works, the use of multivariate data analysis methods is fortunately becoming a routine for the representation, interpretation and modeling of complex data. For this reason, chemometrics reputation and diffusion constantly increased from the 70s, the years in which the discipline was born due to Svante Wold and Bruce Kowalski [1-3] and niche techniques such as Design of Experiment and exploratory multivariate data analysis are nowadays well known and applied.

Chemometrics, the multivariate analysis of data, faces a widespread group of tasks: data compression and visualization, categories assessment, predictive models for properties evaluation and constituent calibration, curve resolution, etc., and many techniques are also used, in order to extract and interpret the information present in the data of different sources in data fusion methodologies.

One of the most appreciated quality of the chemometrics approach is that, in general, makes possible a graphical representation of the complex analyzed data, by means of few readable plots that highlight patterns in samples, variables and their relationships.

Exploratory multivariate data analysis [4] is the first step allowing looking at multivariate data without imposing any a priori model and includes methods such as, Principal Component Analysis [5-12] for bilinear data matrixes and PARAFAC and Tucker for multi-linear data [13-18]. These are the most common techniques used to extract from the data information related to which are the most important variables, for outliers detection and for the evaluation of the redundancy and correlations in data structures,

Most of the chemometrics methods require a deep knowledge of mathematics and statistics to be fully understand, enough to scare nearly always the audience. In this chapter the basis of some common multivariate data analysis techniques are introduced in order to give to the reader the key concepts sufficient to understand the data fusion applications shown in the final chapters.

First of all, some decomposition methods, usually involved in data fusion both as exploratory and data reduction tools will be introduced, with particular attention on the description of Multivariate Curve Resolution. The last part of the chapter is focused on classification problems for two and N-way datasets.

2.1.1 Principal Component Analysis

PCA is a bilinear decomposition method that implies the projection of the original data matrix **X** onto a space of reduced dimensionality, i.e. a new set of *A* latent variables chosen so that the maximum variance of the dataset is retained and latent variables are orthogonal, i.e. uncorrelated, to each other.

The Principal Component decomposition can be schematized as:

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} = \sum_{a=1}^{A} t_a p_a^T + \mathbf{E}$$

A new space of coordinates, based on A unit vectors p_a is built, the graphical representation of the original set of variables and the new one computed by PCA is shown in Figure 2.1.

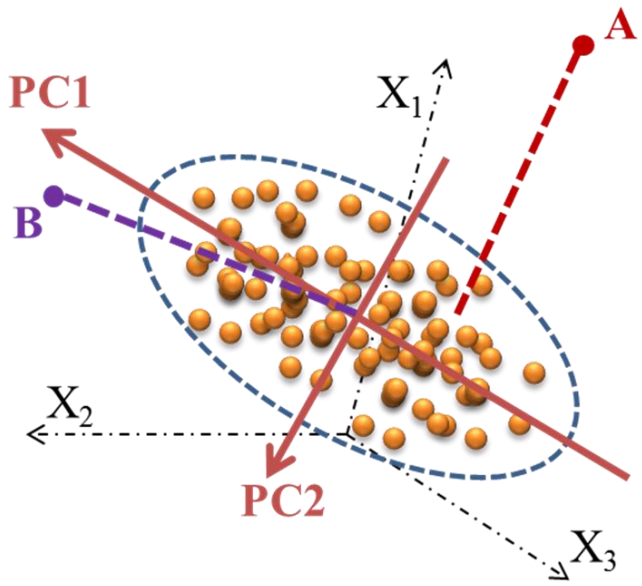

Figure 2.1. *Graphical representation of Principal component space (PC1 and PC2) obtained from an X matrix composed of three original variables (X_1, X_2 and X_3). The projection of samples "A" and "B" from original variables space to PCA space is shown; the red dashed line represent Q, the "A" residuals after projection, while the violet dashed line represent the distance in model space for "B" sample, i.e. how close is B once projected to the origin of the PCs space.*

T vectors [t_1,t_2,t_3,…t_A] are called scores and have length equal to the number of samples/observations, or in general, the number of row of the original data matrix **X**.

P vectors [p_1,p_2,p_3,…p_A] are called loadings, they are orthogonal and of length equal to the number of variables, or in general to the number of columns of the original data matrix **X**.

A corresponds to the number of Principal Components chosen for the computation of the PCA model and can be considered as the effective rank of the matrix **X**. Since *A* in almost all cases is very small with respect to the number of original variables, PCA is often used also as data reduction tool. On the other hand, the reduction of the number of variables involved in the decomposition does not allow to model the totality of the variance of the original data matrix, hence, an unexplained part of the information, ascribable to noise, is still present in residual matrix **E** of dimensionality equal to the original **X** data matrix.

The most widely used algorithms to compute PCA are Non-Linear Iterative Partial Least Squares NIPALS [2,12] and the Singular Value Decomposition [19-20].

The scores represent the coordinates of the samples in the space defined by the new orthogonal principal components, hence, if the scores are plot in a 2D or 3D representation in the so called scores plots, easy and readable information can be highlighted directly from the visualization of the graphical representations. In the scores plot the presence of groups of samples allocated in a restricted area of the new coordinates space can be identified, these share the same behavior, also the presence of anomalous samples lying at great distance with respect to all others and hence, with high probability of being outliers, can be revealed.

On the other hand, loadings represent the weight that the original variables bring through the decomposition of the original data onto the space of principal components, or better, the importance in determining the direction of principal components. Since PCs for definition are orthogonal, and the PC direction is computed in a way that the maximum of the raw data variance is explained, the higher is the weight of a variable in PCs space the higher is the variation of the same variable in the original space. Variables having equal or similar contribution to the PC space are correlated, or anti-

correlated if the sign is opposite, while variables with values close to zero can be considered almost constant.

From the joint interpretation of scores and loadings plot (possible also in a unique plot called biplot) is then possible to assess which are the most valuable variables that force the samples (in the scores plot) to be present in a given place in the principal component space. If the original data are in the form of a continuous series, as in the case of spectra, the interpretation of the loadings plot is quite uneasy. For this reason it is common practice to represent the loadings component-wise, in a way that they can be compared directly with the original signals, giving the possibility to identify the variables that most of all vary, the ones with the highest absolute loading values.

In order to evaluate how well samples are fitted by the PCA model, two statistics can be computed.

Q is a measure of the distance from the model (i.e. the higher Q is, the lowest the model fit for that sample) and is the sum of square of the residuals for a given sample, i.e. Q is the sum of squares of each row (sample) of **E**, for the *i*th sample in X

$$Q_i = \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}$$

where e_i is the *i*th row of **E.**

Q residuals statistic allow establishing if a sample is well fitted in the space of the principal component (in this case its Q value will be accepted according to Q confidence limits), or better, how high is the difference between the sample in the original space with respect to the space of the new set of variables.

In Figure 2.1 sample A is the one that has the highest Q value.

Another important statistics is Hotelling's T^2, which reflect the variation of each samples within the PCA model. In Figure 2.1 T^2 value (which is the sum of squares of scores over all components weighted by the variance explained by each component) of

sample B can be considered the distance of the sample from the center of the plane of principal component space.

T^2 is defined as

$$T^2{}_i = \mathbf{t}_i \lambda^{-1} \mathbf{t}_i{}^{\mathrm{T}}$$

where $\mathbf{t}_i$ refers to the i-th row of the scores matrix **T**, and λ is a diagonal matrix containing the eigenvalues (λ_1 through λ_A) corresponding to the *A* eigenvectors (principal components) retained in the model.

2.1.2 PARAFAC

In analytical chemistry many instrumentations give as output a three-way or higher-order arrays, i.e. data that have more than two sources of variability, as a general example a set of samples characterized by a set of variables acquired at different conditions. Many of the so called "hyphenated-techniques" such as GC/MS or HPLC/MS or others able to detect from the same set of samples more than one block of variables (time and mass/charge ratio from mass-chromatography or emission and excitation in Fluorescence) give as output a third order array, analogously when for the same set of samples variables are acquired at different times a third order array is obtained.

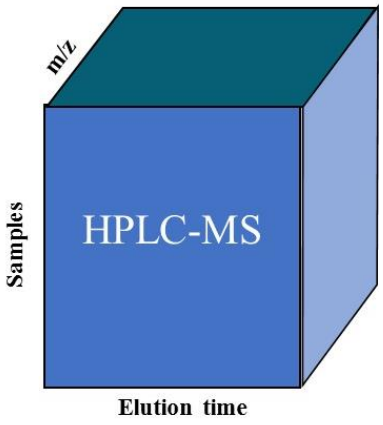

***Figure 2.2**. Schematic representation of a three-way array obtained by HPLC-MS analysis*

The order of an array indicates the number of modes it has, e.g. GC-MS recorded for a set of samples identify three modes: samples, chromatographic elution, mass fragmentation as shown in Figure 2.2. Higher order arrays are suitable for higher order decomposition methods as PARAFAC or TUCKER.

Three-way arrays can be analyzed by means of traditional bilinear-decomposition methods such as PCA after unfolding of the original structure along one of its mode; example will be presented in the next section for a clear explanation of Multiset Multivariate Curve Resolution.

Despite the possibility to analyze three way arrays by means of bilinear decomposition methods, when the assumptions of the N-way methods are fulfilled they are able to give the best results in understanding data structures because all the sources of variability are evaluated jointly.

Parallel Factor Analysis PARAFAC or (Canonical Decomposition CANDECOMP) can be described as one of the possible extensions of PCA when the analysis is performed on high-order arrays, only with respect to the fact that it decomposes the data array to a space of lower dimensionality and allows a graphical representation of the data, while the criteria to obtain the decomposition are different. PARAFAC decomposes the original data array (tensor) as a sum of triple outer product of vectors, when the order of the original data array is three.

The three sets of vectors are called loadings (also scores for the first mode) and the number of PARAFAC factor, *F*, has to be the same for each mode.

Given a three-way array of dimensionality IxJxK respectively for the first, second and third mode, using *F* PARAFAC factors the decomposition of the original data can be expressed as:

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

As in PCA the part of the data which cannot be modelled by the PARAFAC model with the chosen number of components is kept on the residual tensor E.

A(*I*x*F*), B(*J*x*F*) and C (*K*x*F*) are the loadings matrices of the three modelled modes while F is the number of PARAFAC factors.

PARAFAC factors are not imposed to be orthogonal contrary to PCA components, in fact the PARAFAC solution obtained using a given number of factors is unique, which doesn't mean true, but that no other solutions with the same fit of the data can exist.

To assess the correct number of factors is the most important step in building a PARAFAC model. Several strategies can be experienced in order to find the most suitable number of components. If *a priori*, the chemical rank of the investigated system is known, i.e. the number of chemical phenomena, e.g. the number of analytes in a mixture, and data are trilinear, the rank corresponds to the number of factors. On the other hand, a trial and error strategy can be followed using different number of factors and evaluating some important performance parameters of each PARAFAC model such as, the core consistency, the number of iteration to achieve the convergence and split-half analysis results [21]. Besides these parameters, the interpretation of the results is the most crucial point in order to confirm if the correct number of factors has been chosen, in fact, the solutions of the PARAFAC model must have not only a mathematical sense but also must correspond to a real chemical meaningful behavior, e,g, the third mode loadings (referring to the example of Figure 2.2 should have a profile that matches the mass fragmentation profile of each chemical species present in the samples. To better refine the PARAFAC model, constraints such as non-negativity of the loadings referring to concentration or spectra profiles may be introduced.

2.2 Multivariate Curve Resolution

As PCA, Multivariate Curve Resolution (MCR) should be considered a general-purpose factor analysis tool, and as in the case of Principal Component Analysis, the term MCR include more than one algorithm.

The first preliminary studies about MCR involved the resolution of multicomponents system according to the Lambert-Beer's law

$$A_\lambda = \varepsilon_{\lambda,X} b c_X + \varepsilon_{\lambda,Y} b c_Y$$

in which the absorbance values at the wavelength λ is given by the contributions of the concentrations (C_X and C_Y) the molar absorptivities ($\varepsilon_{\lambda,X}$ and $\varepsilon_{\lambda,Y}$) of the two species X and Y and by the constant common cell path length *b*.

Knowing only a singular absorbance value at a given wavelength the estimation of the concentrations of the two species cannot be achieved.

Approaching the problem with the collection of absorbance values at more than one wavelength and having knowledge of molar absorptivities a solution for the remaining variables can be obtained resolving two simultaneous Lambert-Beer's law expressions.

Generalizing the problem to all other analytical methods which can be described by linearly additive responses (chromatography, spectroscopy for example), the system of equations can be formulated as

$$X_1 = C_1 S_{1,1} + C_2 S_{1,2} + \ldots C_N S_{1,N}$$

$$X_2 = C_2 S_{2,1} + C_2 S_{2,2} + \ldots C_N S_{2,N}$$

$$\vdots$$

$$X_F = C_1 S_{F,1} + C_2 S_{F,2} + \ldots C_N S_{F,N}$$

Where X_F indicates the responses at a given instrumental parameter (for example at the *F*-th wavelength), *Cn* represents the concentration of *n*-th specie and $S_{F,n}$ an instrumental sensitivity factor (as molar absorptivity) at the *F*-th parameter for the *n*-th species.

In vector notation this becomes the well known

$$\mathbf{X} = \mathbf{c}\,\mathbf{S}^{\mathbf{T}}$$

If the model obtained is not able to describe the complete variance of the investigated system, as in PCA, the not modelled part of the data are usually indicated as residual matrix E, hence

$$\mathbf{X} = \mathbf{c}\,\mathbf{S}^{\mathbf{T}} + \mathbf{E}$$

When a series of mixtures samples is collected, the graphical representation of the bilinear decomposition can be illustrated as follow

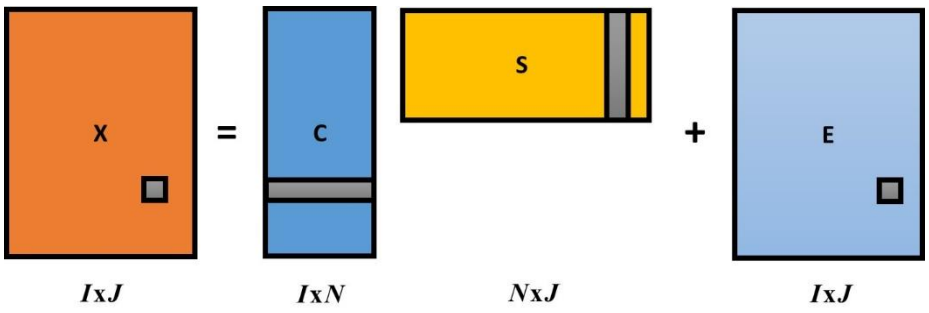

Figure 2.3 *Graphical scheme of bilinear decomposition based on MCR*

Since the purpose of MCR is to achieve from the original data a set of basis able to describe a chemical behavior of the investigated system, and due to the premise described above, the **C** matrix is called concentration matrix while the **S** one is called spectra matrix.

N is the number of MCR resolved components, e.g. the chemical species present in a mixture (but not only, an MCR component can as well depict the baseline or any other phenomenon present in the data, e.g. a temperature trend, etc.) and these components are not orthogonal, e.g. two species may have some part of the spectral profile in common. Moreover, MCR components are not sequential, hence MCR components are not ordered per variance explained, and the same variance can be explained by different components, it is commonly said that part of the variance is overlapped.

The most used multivariate curve resolution algorithms, such as MCR-ALS [22] and iterative target transformation factor analysis ITTFA [23-24] are based on the iterative calculation of **C** and $\mathbf{S}^T$ directly. These methods are based on the iterative calculation of **C** or $\mathbf{S}^T$ starting from estimates of the same matrixes

$$\hat{\mathbf{S}}^T = \mathbf{C}^+\mathbf{X} = (\hat{\mathbf{C}}^T\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}^T\mathbf{X}$$

$$\hat{\mathbf{C}} = \mathbf{X}\mathbf{S}^{T+} = \mathbf{X}\hat{\mathbf{S}}(\hat{\mathbf{S}}^T\hat{\mathbf{S}})^{-1}$$

where ^ indicates an estimates and $^+$ the pseudoinverse.

The iterative computation needs an initial estimates of the concentration or spectra matrixes, which can be a real spectra or concentrations profile, if known a priori, or evaluated by SVD or Evolving Factor Analysis EFA [25-26] or with interactive self-modeling analysis SIMPLISMA [27].

The aim of MCR methodologies is to achieve a resolution of the system able to extract profiles of "pure" species. In order to graphically understand the significance of **C** and $\mathbf{S}^T$ matrixes, without confusion on the terminology, a simulated set of mixture is here resolved and illustrated in Figure 2.4.

The **X** data matrix contains the signals of mixtures of three species having different concentration and spectra profiles.

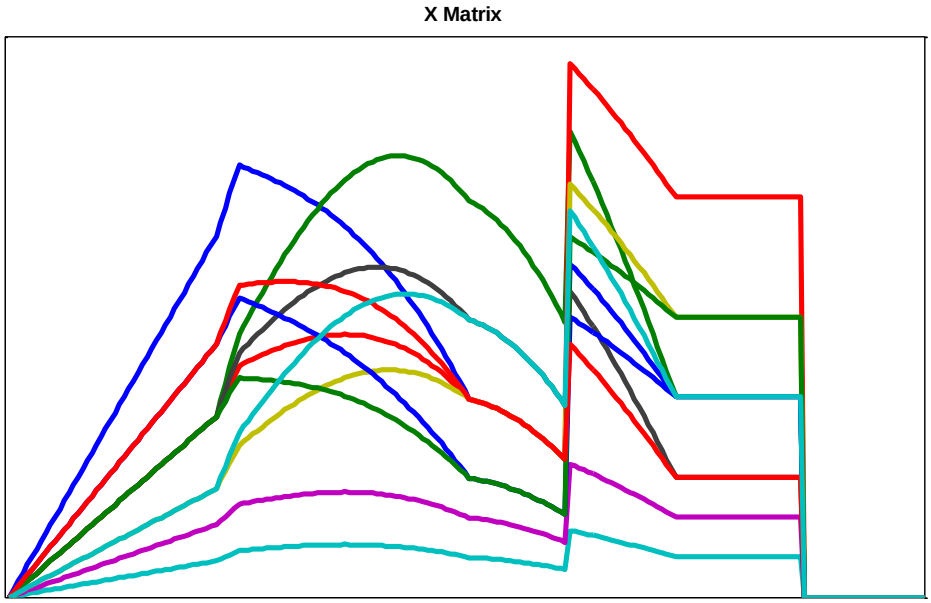

Figure 2.4 *Simulated signal of mixtures of three species*

Just inspecting the reported plot is not easy to explain how the contribution of spectra and concentrations can give, as addictive responses, the signals illustrated.

Applying MCR to the **X** dataset, without any other a priori information given as input, and using a PCA estimation of the concentration profiles, it is possible to obtain the results reported in Figure 2.5.

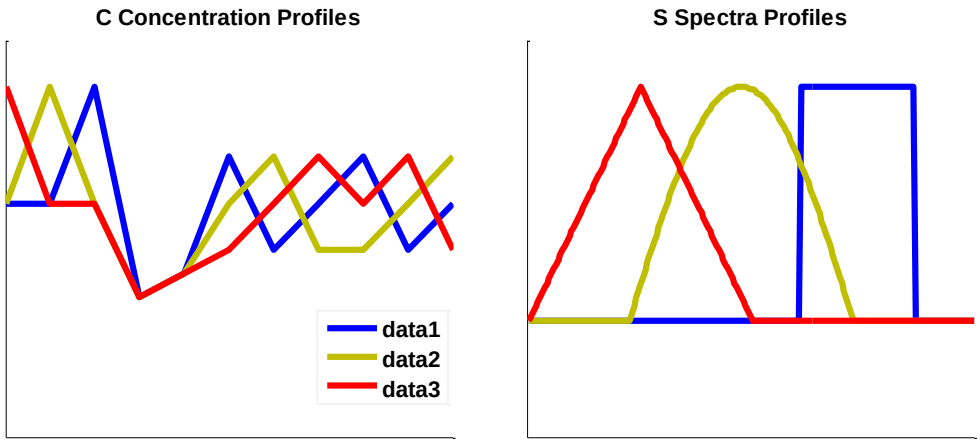

Figure 2.5 *Resolved concentration and spectra profiles*

In this easy case, the MCR model, based on three components, resolves three spectral profiles (Figure 2.5 right) the parabolic, triangular and rectangular shapes, whose **C** profiles (Figure 2.5 left) account for how much of each component is present in each sample. This MCR model provides a meaningful sense without the use of constraints. However, the MCR solution is not unique and many solutions with the same fit may be obtained up to rotation (rotational ambiguity), the same happens for PCA where orthogonality is used as constraint to reach a unique solution. Thus MCR need constraints to solve the rotational ambiguity. These constraints in order to obtain a chemical meaningful solution have to be derived by knowledge of the system, e.g. concentration and spectra profiles, in the case of hyphenated chromatography/UV spectroscopy, cannot be negative.

In general, implementable constraints can be applied both to concentration and spectra profiles and can be classified in two main groups:

a) ***soft constraints*** such as non-negativity, unimodality, and selectivity are applied iteratively at each step of the optimization on spectra and/or concentration profiles. They reflect a real behavior of the system. For example, *non-negativity* constraint is applied in each case in which the C matrix is referred to a concentration (molarity, ppm) or at spectra profile when the collected signals cannot be negative (absorbance). *Unimodality* is applied in order to force the signals to have only one maximum (as in elution profile for a chemical species). *Selectivity* constraints can be imposed both on concentration and spectra profiles, defining components per components if species are present or not, or if in a given region of the spectra the signal related to a specie have to be found or not. Soft constraints help the reduction of the rotational ambiguity of the modelled system;

b) ***hard constraints*** impose the resolved profiles to reflect a real characteristic with the use of equations or laws able to describe a given behavior. For example, a kinetic or equilibrium model can be applied to follow a chemical reaction. The uses of hard constraints helps the reduction of both rotational and intensity ambiguities.

MCR methodologies can be applied to high-order data [28-29] if an unfolding procedure is performed on the three-way (or higher order) dataset. Usually MCR is applied to matrixes in which the rows are referred to samples/observations and columns to variables. If a third mode containing another set of variables is present, the unfolding procedure can be performed column-wise or row-wise as illustrated in Figure 2.6.

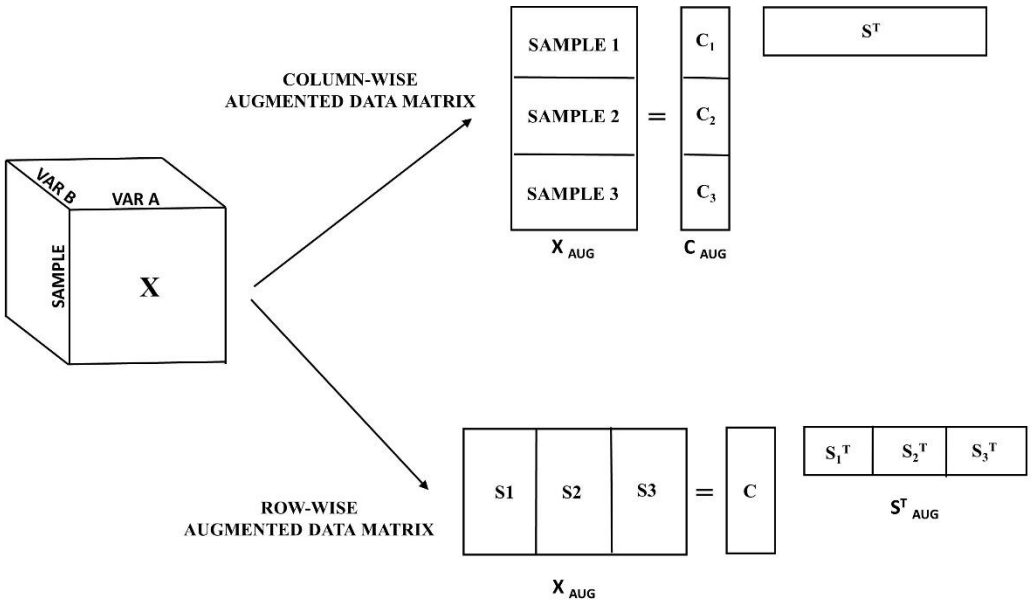

Figure 2.6 *Graphical representation of the bilinear model applied to different types of augmented data matrices.*

Augmentation of data matrix can also be performed adding, in the shared direction, other observations of sources of information. For example, if a set of samples is analyzed by means of other instrumental techniques, the new set of variables can be added via row-wise augmentation since the samples direction is shared.

In the application of MCR, both on bilinear data and multiset structures, as in the case of PARAFAC, the choice of the most suitable number of components, able to describe the system, is one of the most important aspects to consider.

Since MCR methods are focused on the development of a model able to extract the "pure" profile of concentration and spectra, the evaluation of the modelled components can be a suitable way to understand if the model is over or under fitted.

Usually, the rank of the model is equal to the chemical rank of the system but, if linear dependencies exist, model based on a lower number of components can give results which could appear better. To overcome the linear dependencies problem, (linearly dependent concentrations or spectra with the same shape), which produces a series of dyads (concentration and spectra) that cannot be referred to a real contribution but to a linear combination of them, the augmentation of data matrix can be followed in order to add to the investigated system a new source of information not linearly dependent with others.

2.3 Classification Methods

All methods described in previous sections, may be used in data fusion as variable reduction tools as well as to obtain from the X data matrix information about similarities or differences among samples, or to recognize the most important variables able to characterize the investigated system.

In all classification methods, the target of a study is to establish whether a set of samples can be classified or not to one or more classes on the basis of a series of measured responses. The class membership of a group of samples can be known *a priori* or not. To the latter case belong all classification techniques called unsupervised classification methods such as Cluster Analysis (CA) [30] and k-Nearest-Neighbours (kNN) [31].

The term "*class*" indicates a group of sample that share a specific characteristic/feature, e.g. the same ageing, being healthy subject, etc... It is assumed that on the basis of the analysis of a group of responses able to detect, directly or indirectly, any differences among samples, it could be possible to attribute the class membership. In food analysis is common to model a series of samples [32-38] belonging to the same family, for example a set of wines, trying to define the class membership on the basis of the ampelographic composition, or the geographical origin, or by means in general of traceability indicators.

Supervised pattern recognition methods on the other hand use known class membership to build a calibration model which, once assessed, will be used to predict the class membership of future investigated samples. The most common used supervised methods are Soft Independent Modelling of Class Analogies SIMCA [39], Linear Discriminant Analysis LDA [40] and its Partial Least Squares- based derivation, PLS – Discriminant Analysis PLS-DA [41-42]. Here PLS-DA and its extension to n-way data, i.e. NPLS-DA [43] will be briefly mentioned.

PLS-DA is the extension, on classification problem, of the well know Partial Least Squares regression which is a technique used to face calibration tasks, and work using a "PCA-like" projection of the descriptors matrix, **X,** using the criterion of finding directions of maximal covariance with the response **y** vector (PLS-1), or with the the latent variables space of the responses matrix, **Y,** (PLS-2 Regression), when more responses have to be modeled. The definition of the direction of the latent variables is obtained in a way that the covariance between the two blocks **X** and **Y** is maximized, by imposing a least squares fit between the **X** and **Y** scores component wise (PLS inner relationships).

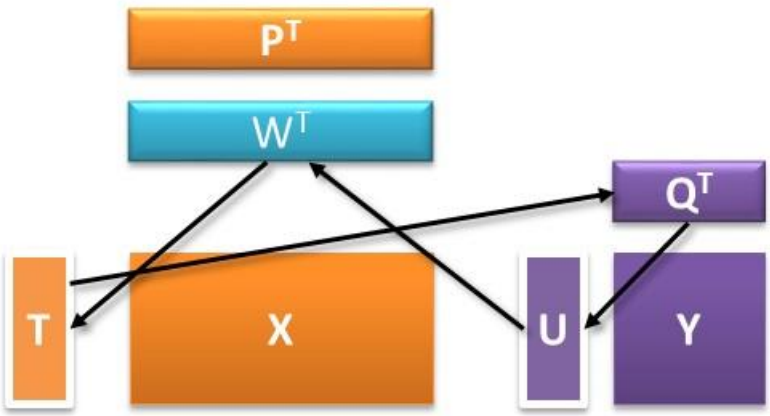

Figure 2.7 *Schematization of the decomposition of the X and Y matrixes performed by PLS*

PLS can be calculated by iterative algorithms such as NIPLAS [1,6] or by using SVD on the covariance XY matrix i.e. SIMPLS algorithm [44].

In PLS both X and Y are decomposed in a PCA-like way:

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{T} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{Q}^{T} + \mathbf{F}$$

The PLS inner relationships is achieved by using a set of weights that contains the contribution of each variable in **X** to model the responses **Y** for each component.

The weight matrix **W** contains the structure in **X** which maximize the covariance between **T** scores and **U** scores.

PLS-DA is based on the same assumption of PLS-R, but the responses matrix, **Y,** is replaced by a class membership vector (PLS-DA1 for two classes) or matrix (PLS-DA2 for three or more classes) containing as many y's as classes in the form of dummy indexes holding for each sample a code indicating belonging or not to a given class.

The response matrix is usually built codifying the samples belonging to the class with ***+1*** and the samples not belonging to the class with ***-1 (***or with ***+1*** and ***0)***.

In order to determine the membership to a class of a given sample a threshold has to be defined. In this Thesis was adopted as classification rule according to which samples are assigned to the class for which the predicted y-value is higher, e.g. if the predicted vector of responses for an unknown sample is [-0.7 0.7 0.2] (in the case of a three classes problem), it will be assigned to class two. Other threshold criteria can be used, based on example on Bayesian inferences [45-46], as the one adopted in the well know Eigenvector's Matlab package PLS-Toolbox™.

In analogy with PLS-DA, NPLS-DA is the derivation of NPLS for classification problems [43,47-48].

NPLS-DA works on higher order arrays, e.g. an **X** array of dimensionality $I \times J \times K$ and one **Y** array, containing the class indexes, of dimensionality $I \times M$.

In NPLS the X array is decomposed in a Tucker-like model:

$$\mathbf{X} = \mathbf{T}\mathbf{G}_{\mathbf{x}}\left(\mathbf{W}_{\mathbf{k}} \otimes \mathbf{W}_{\mathbf{j}}\right)^{\mathbf{T}} + \mathbf{E}_{\mathbf{x}}$$

where **T** are the first mode scores and $\mathbf{W_j}$ and $\mathbf{W_k}$ are the weight matrices for the second and third mode respectively.

$\mathbf{G_x}$ defines the core array of dimensionality $F \times F \times F$, where F is the number of factors used to build the model

$$\mathbf{G}_{\mathbf{x}=}\ \mathbf{T}^{+}\mathbf{X}\ ((\mathbf{W}_{\mathbf{k}})^{+} \otimes\ (\mathbf{W}_{\mathbf{j}})^{+})^{\mathbf{T}}$$

Analogously the **Y** block is decomposed as:

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{E}_{\mathbf{Y}}$$

As in PLS the weights are calculated in a way that the covariance between the scores of X and Y is maximized.

2.4 References

[1] Kowalski, B.R., "*Chemometrics: Views and propositions"*. Journal of Chemical Information and Computer Science, 1975, 15(4), 201-203.

[2] Wold, H., "*Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach*". In: Gani, J. (ed.): Perspectives in probability and statistics. Applied Probability Trust, Sheffield, England, 1975.

[3] Wold, H. "*Soft modelling, the basic design and some extensions*". In: Joreskog, K.G., Wold, H. (eds.): Systems under indirect observation II, North-Holland, Amsterdam, 1982.

[4] Li Vigni M., Durante C. , Cocchi M., "*Chapter. 3 Exploratory Data Analysis"*, In F. Marini (Ed.) Chemometrics in Food Chemistry, Vol. 28 Data Handling in Science and Technology series, Elsevier 2013

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

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

[7] Joliffe I.T., "*Principal Component Analysis*", 2nd Edition, New York, Springer-Verlag, 2002

[8] Massart D.L., "*Handbook of chemometrics and qualimetrics: part A*", in: Massart D.L., Vandeginste B.G.M., Buydens L.M.C. , De Jong S., Lewi P.J., Smeyers-Verbeke J., Data handling in science and technology series, part A, vol. 20. Amsterdam: Elsevier; 1998, 519–56 [chapter 17]

[9] Esbensen K.H., Geladi P., "*Principal component analysis: concept, geometrical interpretation, mathematical background, algorithms, history, practice"*. In: Brown S.D., Tauler R., Walczak B., editors. Comprehensive chemometrics: chemical and

biochemical data analysis, Vol. 2. Amsterdam: Elsevier Ltd.; 2009, 211–27 [chapter 2.13].

[10] Wold S, Esbensen KH, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory and System, 1987, 2 , 37–52.

[11] Smilde A., Bro R., Geladi P., "*Models for two-way one-block data analysis: component models", i*n: Smilde A, Bro R, Geladi P, editors. Multi-way analysis with applications, multiway analysis in the chemical sciences. Chirchester: John Wiley & Sons, 2004, 35–45.

[12] Wold, H. "*Nonlinear estimation by iterative least square procedures*". In: David, F.N. (ed.): Research papers in statistics, Festschrift for J. Neyman, Wiley, New York, 411-444, 1966.

[13] Smilde A., Bro R., Geladi P., "*Three-way component and regression models", i*n: Smilde A, Bro R, Geladi P, editors. Multi-way analysis with applications, multiway analysis in the chemical sciences. Chirchester: John Wiley & Sons, 2004, 57–86.

[14] Harshman, R.A. "*Foundation of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis*". UCLA Working Papers in Phonetics, 1970, 16, 1-84.

[15] Carroll, J.D., Chang, J. "*Analysis of individual differences in multidimensional scaling via N-way generalization and Eckart-Young decomposition*", Psychometrika, 1970, 35, 283-319.

[16] Bro, R. PARAFAC. "*Tutorial and application"*. Chemometrics and Intelligent Laboratory Systems, 1997, 38(2), 149-171.

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

[18] Tucker, L. R., "*Some mathematical notes on three-mode factor analysis*". Psychometrika, 1966, 31, 279.

[19] Eckart, C., Young, G. "*The approximation of one matrix by another of lower rank*". Psychometrika, 1936, 1, 211-218.

[20] Golub, G.H., Reinsch. C. "*Singular value decomposition and least squares solutions"*. Numerische Mathematik, 1970, 14(2), 403-420

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

[22] R. Tauler, "*Multivariate Curve Resolution Applied to Second Order Data*", Chemometrics and Intelligent Laboratory System, 1995, 30, 133–146

[23] P.G. Gemperline, "*A Priori Estimate of the Elution Profiles of the Pure Components in Overlapped Liquid Chromatography Peaks Using Target Factor Analysis",* Journal of Chemical Information and Modelling, 1984, 24, 206–212

[24] B. G. M. Vandeginste, W. Derks, G. Kateman, "*Multicomponent Self-Modelling Curve Resolution in High-Performance Liquid Chromatography by Iterative Target Transformation Analysis"*, Analytica Chimica. Acta, 1985, 173, 253–264

[25] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbulher, "*Evolving Factor Analysis of Spectrophotometric Titrations: Forget About the Law of Mass Action? ",* Chimia, 1985, 39, 315–317

[26] M. Maeder, "*Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks*". Analytical Chemistry, 1987, 59 (3), 527–530

[27] W. Windig, "*Spectral data files for self-modeling curve resolution with examples using the Simplisma approach*", Chemometrics and Intelligent Laboratory Systems, 1997, 36(1), 3-16

[28] J. Saurina, R. Tauler, "*Strategies for Solving Matrix Effects in the Analysis of Triphenyltin in Sea-Water Samples by Three-Way Multivariate Curve Resolution*", Analyst, 2000, 125, 2038–2043.

[29] E. Pere´-Trepat, E, S. Lacorte, R. Tauler, "*Alternative Calibration Approaches for LC-MS Quantitative Determination of Coeluted Compounds in Complex Environmental Mixtures Using Multivariate Curve Resolution*", Analytica Chimica Acta, 2007, 595, 228–237.

[30] T.Hastie, R. Tibshirani, J. Friedman, "*Hierarchical clustering. The Elements of Statistical Learning (2nd ed.)*", Springer, New York, USA, 2009.

[31] M.A. Sharaf, D.L. Illman, B.R. Kowalski, "*Chemometrics*", John Wiley & Sons, New York, USA, 1986.

[32] 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

[33] G. Papotti, D. Bertelli, R. Graziosi, M. Silvestri, L. Bertacchini, C. Durante, M. Plessi, *"Application of One- and Two-Dimensional NMR Spectroscopy for the Characterization of Protected Designation of Origin Lambrusco Wines of Modena"*, Journal of Agricultural and Food Chemistry, 2012, 61(8), 1741-1746

[34] M.A. Brescia, I.J. Kosir, V. Caldirola, J. Kidric, A. Sacco, *"Chemometric classification of Apulian and Slovenian wines using 1H NMR and ICP-OES together with HPICE data",* Journal of Agricultural and Food Chemistry, 2003, 51(1), 21-26

[35] R Bucci, A.D. Magrí, A.L. Magrí, D Marini, F Marini, *"Chemical authentication of extra virgin olive oil varieties by supervised chemometric procedures*", Journal of Agricultural and Food Chemistry, 2002, 50 (3), 413-418

[36] F. Marini, A.L. Magrı, F. Balestrieri, F. Fabretti, D Marini, "*Supervised pattern recognition applied to the discrimination of the floral origin of six types of Italian honey samples*", Analytica Chimica Acta, 2004, 515 (1), 117-125

[37] R. Vitale, M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, "*A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics",* Chemometrics and Intelligent Laboratory Systems, 2012, 121, 90-99

[38] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, R. Acquistucci, R. Francisci, "*Classification of 6 durum wheat cultivars from Sicily (Italy) using artificial neural networks",*Chemometrics and Intelligent Laboratory Systems, 2008, 90 (1), 1-7

[39] S. Wold, "*Pattern recognition by means of disjoint principal component models*" Pattern Recognition, 1976, 8, 127-139.

[40] R.A. Fisher, "*The Use of Multiple Measurements in Taxonomic Problems*". Annals of Eugenics 1936, 7, 179–188.

[41] M. Barker, W. Rayens, "*Partial Least Squares for Discrimination*", Journal of Chemometrics, 2003, 17(3), 166-173.

[42] G. Musumarra, V. Barresi, D.F. Condorelli, C.G. Fortuna, S. Scirè, "*Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis*",Journal of Chemometrics, 2004, 18(3), 125-132

[43] R. Bro, "*Multiway calibration. Multi-linear PLS,*" Journal of Chemometrics, 1996, 10, 47–61

[44] S. De Jong, "*SIMPLS: An alternative approach to partial least squares regression",* Chemometrics and Intelligent Laboratory Systems, 1993, 18(3), 251-263

[45] B.Wise, N. Gallagher, R. Bro, J. Shaver, W. Windig, R. Koch, "*PLS_Toolbox for use with MATLAB™*", Eigenvector Research, Inc., Wenatchee, USA, 2006.

[46] G. Barnard, "*Studies in the History of Probability and Statistics: IX. Thomas Bayes' Essay Towards Solving a Problem in the Doctrine of Chances*", 1958, Biometrika, 45, 293–295.

[47] A. K. Smilde, "*Comments on multilinear PLS*". Journal of Chemometrics, 1997, 11, 367-377.

[48] S. de Jong, "*Regression coefficients in multilinear PLS*", Journal of Chemometrics, 1998, 12, 77-81.

CHAPTER 3

Data Fusion Methodolgies

3.1 Introduction

In this chapter, the description of the methodologies belonging to the data fusion's family will be provided, with particular attention to the application to chemical data.

Each class of methodologies is treated separately, in order to point out pros and cons, which have to be considered when looking for the most suitable data fusion strategy.

Moreover, in this way we would like to clearly illustrate the methodologies, differentiating them depending on the level at which the fusion of data occurs.

This kind of classification is the most common in chemometrics, even if others are present in literature and many terms are used interchangeably and often bring to misunderstandings, especially for the distinction between mid and high-level data fusion techniques.

Several general considerations, such as data preprocessing and merging procedures, will be discussed, in order to provide to the reader almost all the basis to address data fusion applications.

3.2 Low-level data fusion

Low-level data fusion methodologies [1-5], also known as concatenated data fusion, represent the easiest way to deal with data arising from different sources of information.

In fact, data are arranged together by a simple concatenation of the blocks of variables, which are the outputs of different analytical techniques, following the shared sample direction, as depicted in Figure 3.2.1.

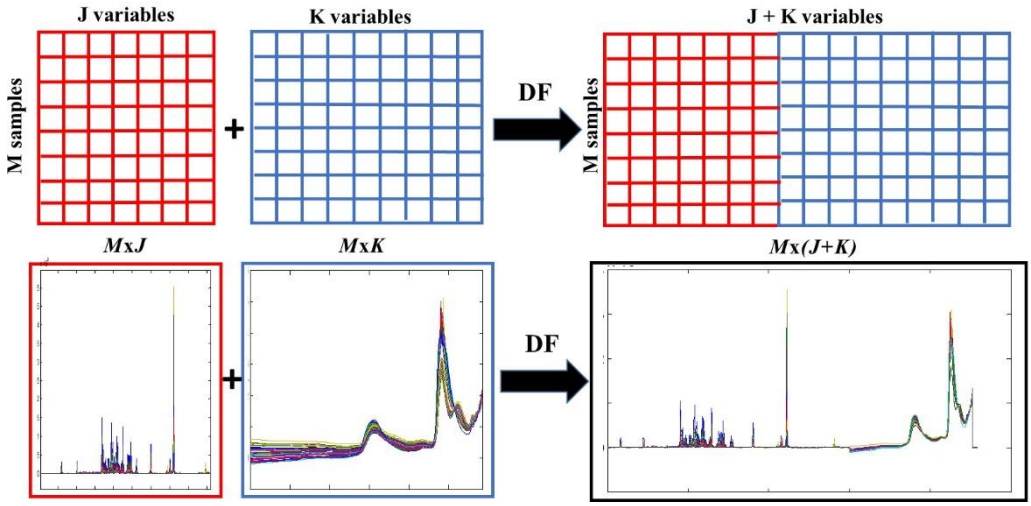

Figure 3.2.1 *Schematization of low-level data fusion*

Since the merging procedure is performed at the so called "*data level*" the most important task to face prior to the fusion step is the preprocessing of the data block.

It is possible to distinguish the preprocessing in two steps, the first one is oriented to treat the data as usually done for exploratory data analysis [6-7]. In this step, each block is treated separately in order to resolve problems related to misalignment, or to remove noise from the data or just to normalize and scale the variables. The common preprocessing such as Mean Center, Autoscale, Standard Normal Variate (SNV) [8-9], Multiplicative Scatter Correction (MSC) [8-10], Weighted Least Squares Baseline Correction (WLS) [10], Derivatives [11-12] and all kinds of normalization can be used.

The second step of data preprocessing is instead related to the concatenation of the blocks of data and involves all the aspects related to weighing and scaling. In low-level

data fusion the dimensionality (or size) of the original data has to be considered since it can deeply affect the results. The most common strategy for merging variables arising from different analytical techniques is the so called “block scaling” which imposes the same variance to each block. If the dimensions, e.g. the number of variables, are very different, without a suitable scaling, prior to the merging procedure and statistical data analysis, the largest block of data could dominate and results can be attributable to “size” effects. Generally, low-level data fusion methodologies are recommended when the variables which have to be fused are, in such a way, similar, for example they are all punctual variables (e.g. metal concentrations) or spectra (e.g. UV and NIR signals).

Regardless of all the aspects described above, the raw data matrixes can be expressed with different measurements units, hence, a normalization of each individual variable can be considered, and autoscaling or Pareto scaling may be needed. Autoscaling is generally not suggested when data in form of signals are considered, but becomes almost mandatory when individual variables are used.

Adopting autoscaling in order to solve the issue of the different measurement units consists of imposing to each individual variables to have mean equal to zero and variance equal to one.

On the other hand, if the same variance is imposed to each variable of each individual block, in the case where one or more of the original datasets have a number of variables much lower with respect to the others, the possibility to attribute too much importance to the these variables is another aspect to take in account.

Once that the merging procedure, considering both individual preprocessings of datasets and weighting for the concatenation, is concluded the augmented data matrix can be analyzed by means of multivariate data analysis tools. As example, an application of low-level data fusion will be presented in Chapter 4 with the aim to classify different commercial classes of Balsamic Vinegars from Modena using two spectroscopic techniques, namely Near- (NIR) and Mid- (MIR) Infrared Spectroscopy.

A particular case of low-level data fusion is the extension of Multivariate Curve Resolution known as multiset-MCR [13-14]. By means of row-wise unfolding (see Figure 2.6) an augmented matrix, built by merging variables arising from different analytical platforms, can be analyzed by means of MCR. In this particular case, the resolved spectra profiles consist of the merged contributions of the resolved spectra of the separate techniques and share the concentration profiles. This approach can be used to extract in a unique solution contributions which are shared between different blocks of variables enhancing both the possibility to have information related to concentration of the singular resolved profiles and correlation among data of different sources.

In general, multivariate data analysis of data sets containing different classes of punctual variables can be formally attributed to the family of low-level data fusion, hence, several examples of exploratory data analysis (in many different fields) can be found in literature even if not explicitly referred to as data fusion.

Since in low-level methodologies the different data are directly concatenated, without any kind of manipulation, with the only exception of preprocessings, all the models which can be obtained contain the "*raw*" information of the original sets of variables. In general, when a model based on low-level data fusion is used, the backward investigation able to explain characteristics linked directly to the original variables results easy and fast and not different to the one which is conducted when a not fused data table is elaborated. As an example, if considering a PCA model obtained from two concatenated datasets in which spectroscopic signals are present, the obtained loadings vectors contain the information of both the spectroscopic signals, hence, the same considerations which could be extracted by separate models can be highlighted (such as the most informative wavelengths, invariant regions etc.). Moreover, since the two blocks of variables are modelled jointly in a unique solution, also the correlation between variables belonging to different blocks should be easily individuated, and more, the characterization of samples (scores vectors) reflects the fact that two blocks of variables are involved giving the possibility to better describe and easily unveil

hidden features. A main drawback, on the other hand, can be that the noise level is also increased: since the raw data sets are fused “as is”, also their noise content is added.

3.3 Mid-level data fusion

Mid-level data fusion, also known as hierarchical or feature level data fusion, differs from low-level because prior to the merging procedure the original data are analyzed separately and a step of features extraction or data reduction is performed.

A schematization of the mid-level fusion strategy is proposed in Figure 3.3.1

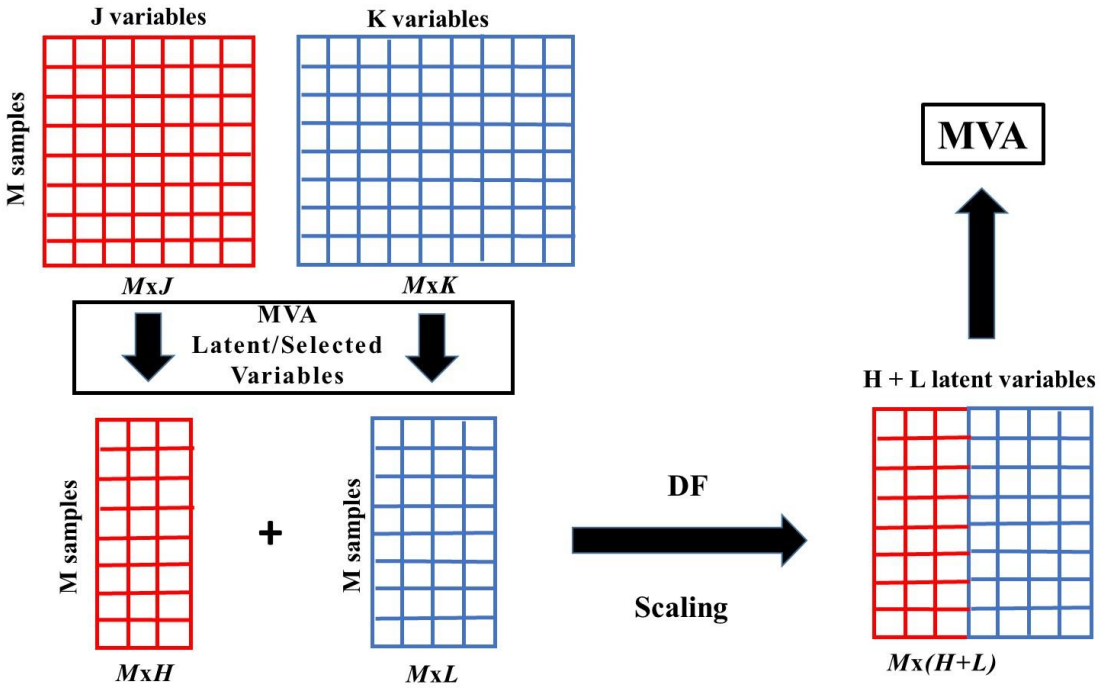

Figure 3.3.1 *Schematization of mid-level or feature-level data fusion. MVA is the acronym for Multivariate Analysis. H and L are the extracted features or components for each data set respectively.*

In mid-level fusion the data arising from different sources of information are merged after a preliminary data manipulation on raw original variables, which can be performed using many different data analysis strategies.

Once that the features are extracted and concatenated, the resulting fused dataset can be analyzed by means of multivariate data analysis tools depending on the goal to pursue (exploratory analysis, classification, etc).

The way in which the features to be fused are obtained is the most important task to face. Two different typologies of data analysis performed on raw signals in order to extract features from original data can be defined:

i) ***Variables selection tools*** are used to detect from the whole set of original variables the most important and informative ones in order to solve the investigated problem

ii) ***Features extraction tools*** are used to analyze the original data projecting them in a new latent variables space, or, in more general terms, data reduction is applied and a set of components or factors is obtained. The information obtained, related to the samples (scores, concentration profiles, first mode loadings, etc.) are fused instead of original variables

The main target of variable selection methods is to select, hence reduce, the number of original variables in order to improve modeling and make the interpretation of the results easier, disregarding all non-informative variables.

Well-established algorithms are present in literature, often used in calibration and classification issues. Since, as described in Chapter 2, in each kind of data the informative part is accompanied with noise, redundancy and invariant parts, the possibility to select a subset of variables which are the most suitable to solve the investigated problem should help the modelling and evaluation of results.

Variable selection methods are usually adopted when continuous signals are investigated (spectra) and are based on different approaches such as genetic algorithm (GA) [15-17], Jack-Knife Cross Validation [18], i-PLS [3,19], Wavelet [20-22] to name a few. With the exception of wavelet-based methods, they may also be applied to select punctual variables, as well as other methods such as Uninformative Variables Elimination (UVE) [23]. Moreover, variable (or signal region) ranking methods can be

adopted to retain a subset of relevant variables, i.e. Variable Importance in Projection (VIP) [24-25] or Selectivity Ratio (SR) [26].

The selected subsets of original variables are arranged through the common samples direction for each of the investigated data matrixes. The obtained fused dataset, containing the most informative variables belonging to each block, can then be analyzed by means of multivariate data analysis techniques such as PCA, PLS-DA, SIMCA and others.

As in the case of low-level data fusion, the merging strategy has to be evaluated in order to avoid that a given subset of variables, depending on its dimensionality, dominates the results of the analysis. The block scaling procedure is the most suited if *a priori* knowledge is not present, or if the intention to give higher weight to a block which is considered more informative is not desired.

Since the fused block contains a subset of selected original variables, the variance which each of them carries out is more or less the same, or better comparable with the variance that each variable of the original blocks explains. This is a crucial point when, in order to build a fused dataset, *via* mid-level fusion, selected variables and extracted latent ones are combined together. The backwards considerations, which can be extracted when a mid-level data fusion approach is applied, are quite similar with respect to the ones obtained via low-level fusion. The interpretability of the results, since the fused variables are related directly to the original ones, is readable and easily analyzable and the possibility to unveil hidden correlations among variables arising from different sources of information can be obtained.

In some cases, the variable selection could not be the most suitable way to proceed. In fact, it is assumed *a priori* that variables which are not relevant in the separate datasets do not carry any information that can be synergic when these are fused together. Moreover, all the limits and drawbacks of the adopted variable selection strategies will affect every further data analysis. The use of techniques which work in projection spaces becomes mandatory in some cases, for example when the data to be fused have

different orders (matrixes and tensors). The use of decomposition methods allows overcoming the above underlined problems and, at the same time, to reduce dramatically the number of variables involved in the merging procedure. As a matter of fact, they take advantage of the correlation structure of raw data, and data reduction is not achieved at the expense of the information carried by the original variables: only noise is filtered. Obviously, due care in components selection should be used. When the data to be fused are characterized by having different dimensionality several multivariate tools can be used to extract the features related to the shared direction. When a group of samples is analyzed by means of different analytical platforms, depending on the kind of outputs that each individual techniques provides, the shared samples direction can be combined as reported in Figure 3.3.2.

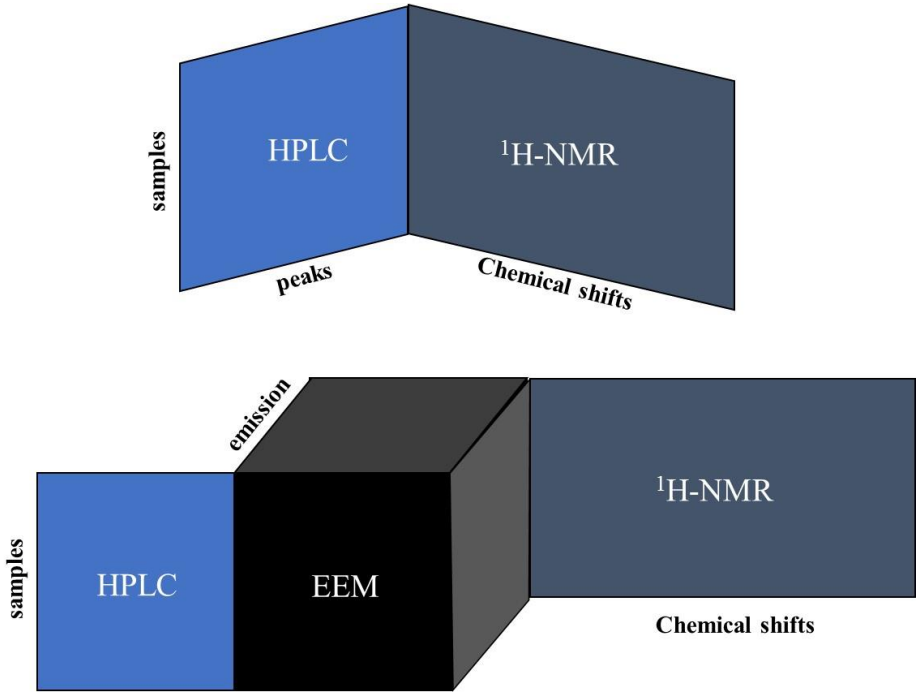

Figure 3.3.2 *Example of the organization of outputs of different analytical techniques which share samples direction*

The mid-level fusion methodologies which involve the use of MVA for the extraction of latent variables from the original blocks are also known as hierarchical data fusion

[27-28]. In fact, two sub-levels of multivariate analysis can be defined. In the lower level the original data are used to build a model from which some new latent variables are obtained, whilst, in the upper level, the merged latent variables are modelled at the same time in a unique step of MVA.

Several techniques can be used to model, in the lower level, the original variables, and the merged data table in the upper one. Some authors adopted PCA as a modelling tool for the two levels of multivariate analysis, using interchangeably the terms CPCA (Consensus Principal Component Analysis) or H-PCA (Hierarchical Principal Component Analysis) to describe this specific mid-level fusion framework [27-29].

The adoption of the most suited multivariate analysis tool is a crucial step in the development of the mid-level strategy.

Depending on the typologies of data, several techniques can be used such as PCA and PLS [27-29], Multivariate Curve Resolution [30-31], LDA [4], PARAFAC [32] and by means of wavelet transform [33-34].

The choice of the most suitable data analysis technique for modelling the data in the lower level have to be done by taking into consideration the aim to pursue in the upper level.

In fact, depending on the model adopted in the lower level, the extracted features may contain information that could result relevant or not.

If the aim of the data fusion application is to unveil some correlation among variables and to improve the characterization of samples, exploratory tools such as PCA can be adopted. If, on the other hand, the knowledge about the system allows resolving in an analytical way, by means of MCR, the pure spectra and concentration profiles, additional information can be highlighted, since the extracted components reflect a chemical meaningful behavior.

When mid-level data fusion is adopted for classification purposes, PLS-DA scores can be chosen to build the fused data set. In this way, since the new set of variables contains

the information related to class memberships, the classification results should improve with respect to the use of exploratory tools. On the other hand, a higher validation effort (in terms of samples and/or double cross-validation schemes) is required, since PLS-DA is a supervised method, while PCA is not.

An important aspect to consider, when dealing with variable reduction techniques for the construction of a mid-level fusion framework, is the readiness to reconstruct from the upper level of MVA the salient information related to the original blocks of variables.

With respect to low-level data fusion, the merged variables involved are able to characterize the original system but are not directly referable to the initial variables, since obtained by projection onto a new space of coordinates. The backward interpretation of the results obtained in the upper level is easy when the lower level of data analysis is performed without forcing the modelling procedure. In general, the easier is the explanation of the results on the lower level, the easiest should be the one obtained in the upper level. For example, interpreting the results obtained from the analysis of merged scores of PCA models should be easier with respect to the analysis of fused PLS-DA scores, in which both weights and loadings plots have to be taken into consideration.

As in low-level and mid-level variables selection data fusion, the way in which the new sets of variables are scaled, when concatenated, can dramatically modify the results.

With respect to the fusion of original variables (or a subset of selected variables), the latent ones obtained in the lower-level modelling are able to explain a great part of the variance of the raw blocks of data. Moreover, even if different multivariate analysis are performed on the separate data sets, the number of extracted principal components (or latent variables or concentration profiles) do not differ too much in terms of dimensionality. Depending on the adopted MVA the number of components should reflect a characteristic of the data which can be traced (not strictly but as a general explanation) to the "rank" of the system. Generally, the extracted variables to be merged

have similar scale, hence, the risk to impose to the most numerous block an higher weight in modelling is overcome.

For this reasons, a trial and error approach can be followed, considering also scaling procedures of the blocks forbidden when dealing with low-level fusion. In Chapter 4, for example, the different blocks of variables will be scaled without imposing the same variance to each block, but autoscaling separately the merged extracted variables.

3.4 High-level data fusion

High-level data fusion methodologies represent a widespread class of techniques particularly established in the analysis of multiple sensors (multisensors fusion) in fields such as process control, robotics, military applications [35-37].

High-level data fusion is also known as decision-level fusion: in fact, the combination of the information related to the different sources is not performed at the data level but combining the so called "identity declarations" obtained by the individual sets of variables.

The application of high-level methodologies is strictly exploited in classification problems where a super-classification index is computed using several sources of information as schematized in Figure 3.4.1

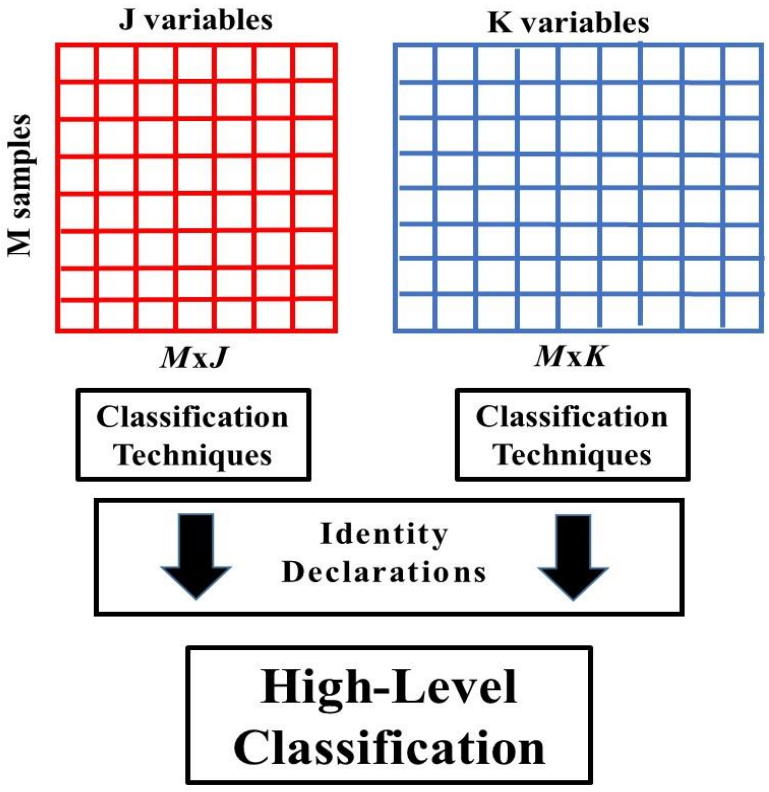

Figure 3.4.1 *Schematization of high-level or decision-level data fusion*

In order to apply high-level methodologies, the original sources of information (sensors, analytical and experts outputs) have to be modelled by means of classification techniques able to associate each sample to a given class, also called "*identity declaration*".

As in the case of mid-level data fusion, two levels of analysis are involved in decision-level fusion. The upper level consists in the combination of the first level identity declarations in order to produce the "*super-identity declaration vector"*.

Because the merging procedure involves only classification indexes, high-level data fusion can be applied without considering dimensionalities, structures, dissimilarities and problematics related to the original data.

The choice of the most suitable sets of sensors or data inputs strongly influences the possibility or not to achieve good classification performances, and represents the most time consuming part of the high-level framework as stated by Steinmetz [38].

In process control, a field in which high-level fusion is well established and used, the criteria adopted for the choice of the best set of sensors are based on *i)* economic aspects, *ii*) the evaluation of the complementarity or the redundancy of the sensors which will be used in the system, *iii*) the nature of the problem to be solved.

On the other hand, if the main goal of the fusion approach is to provide good classification performances by using different groups of variables, which characterize the investigated samples, the adoption of the most suitable way to identify the super-identity declaration could be a challenging step.

The easiest way to proceed is to consider the majority vote criterion [39]. In this approach the class membership in the final step is assigned considering which is the identity declaration resulting from the majority of the models obtained in the lower level of modelling.

More complex methods are available, the most widely used being the Bayes approach based on probabilistic responses [17,33-41]. Others are based on possibility (or

evidence theory), like the Dempster–Shafer theory [42], which is a generalization of Bayesian techniques applied to data with a high level of uncertainty.

Even if the use of high-level data fusion allows obtaining better results [43-44] when compared to the classification performances from separate datasets, it is quite impossible to obtain information able to explain intrinsic characteristics present in the raw data-level just analyzing backwards the results achieved in the fused model.

3.5 Coupled Methods for Data Fusion

All the data-fusion strategies described in the previous sections were characterized on the level in which the fusion process occurs, namely at data level, at features level or at decision level. In these approaches, with the only exception of low-level data fusion, the original blocks of variables are modelled using separate multivariate analysis tools and the outputs (features or identity declarations) are used to build the upper level of modelling.

The idea behind coupled methodologies is quite different with respect to other kinds of data-fusion approaches and deals with the simultaneous resolution of the investigated system, in which at least two matrixes are analyzed, in such a way that the different sources of information share the same features.

The general framework can be separated in six steps [45]:

i) choice of ***data*** to be analyzed
ii) ***pre-processing*** steps included in the method
iii) general mathematical ***model*** for the data
iv) ***objective function*** for the evaluation of model parameters
v) identification of ***constraints*** to obtain a unique solution
vi) ***algorithmic strategy*** to derive the model parameters

Obviously, the ***data*** involved in the data fusion has to provide information able to characterize/solve the problem. At least two data tables that share a common direction

(samples/rows or variables/columns) have to be considered. For the ***pre-processing*** the considerations made on low, mid and high-level data fusion are still valid. The most important aspect to take into account is to avoid the possibility that the results of the fusion process are not comprehensive, but dominated by only one block of variables. Weighting and scaling procedures aimed to remove offsets and scaling differences should be applied as described in the previous sections.

As mentioned above, ***the model*** built in the coupled framework has to describe the system in a way that a part of the results is shared between the different sources of information. Assuming to have *K* data matrixes $\mathbf{X}_K$, containing $\mathbf{I}_K$ rows (samples) and $\mathbf{J}_K$ columns (variables), if *K* PCA models are computed for the *K* distinct datasets using *R* principal components, the results can be formulated as

$$\boldsymbol{W}_K \mathbf{X}_K = \mathbf{T}_K \mathbf{P}_K^T + \mathbf{E}_K$$

$\boldsymbol{W}_K$ ($\boldsymbol{W}_K \geq 0$) is a pre-specified weight for the *k-th* block, $\mathbf{I}_K \times R$ and $\mathbf{J}_K \times R$ represent the matrixes of scores and loadings, whilst $\mathbf{E}_K$ ($\mathbf{I}_K \times \mathbf{J}_K$) contains the not modelled part of the original data.

In order to couple the data, the common direction is constrained for the *K* blocks. If the data tables share the samples mode, the system can be expressed as:

$$\boldsymbol{W}_K \mathbf{X}_K = \mathbf{T} \mathbf{P}_K^T + \mathbf{E}_K$$

or, if the variables mode is shared

$$\boldsymbol{W}_K \mathbf{X}_K = \mathbf{T}_K \mathbf{P}^T + \mathbf{E}_K$$

In order to evaluate the parameters of the model, a common ***objective function*** is introduced under the restriction that either $T_1 = ... = T_K = T$ (common object mode) or $P_1 = ... = P_K = P$ (common variable mode).

$$\min_{\mathbf{T}_k, \mathbf{P}_k} \sum_k \left\| \mathbf{X}_k - \mathbf{T}_k \mathbf{P}_k^T \right\|^2$$

Since rotational ambiguity may lead to not unique solutions, ***identification constraints*** can be implemented, such as imposing orthogonal axes in the direction of highest variance for scores and loadings. In this case, in order to minimize the objective function, the ***algorithmic strategy*** should be solving the problem by using singular value decomposition (SVD).

In literature many methods are proposed which can be defined as a specialization of the general framework. All methods are based on the same objective function but differences in data (shared rows or columns), pre-processing, model and identification of constraints may appear and bring to different results. The aim of this section is not to provide a deep investigation of the methodologies used to couple data matrixes arising from different techniques. The most salient aspects of established strategy such as SUM-PCA [46] ,unrestricted PCovR [47], multiple factor analysis MFA [48-49] , STATIS [50] and SCA-P [51] are summarized in Table 3.5.1

Table 3.5.1 *Summary of simultaneous component methods*

Method	**Common Mode**	**Pre-processing**	**Matrix Weights**	**Identification Constraints**
SUM-PCA	Samples	Auto-scaling of variables and block scale of blocks	$W_K = 1$	$\mathbf{T}^T\mathbf{T}=\mathbf{I}$
Unr. PCovR	Samples	Auto-scaling of variables	Minimize CV errors	$\mathbf{T}^T\mathbf{T}=\mathbf{I}$
MFA	Samples	Auto-scaling of variables	Inverse of largest singular value	$\mathbf{T}^T\mathbf{T}=\mathbf{I}$
STATIS	Samples		Compromise weights	$\mathbf{P}_{conc}{}^T\mathbf{P}_{conc}=\mathbf{I}$
SCA-P	Vaiables	Auto-scaling of variables	$W_K = 1$	$\mathbf{T}_{conc}{}^T\mathbf{T}_{conc}=\mathbf{I}$

A novel algorithm, Coupled Matrix Tensor Factorization (CMTF) [52], developed for social sciences purposes and applicable in chemometrics, was formulated for the joint analysis of structures of data with different dimensionalities (matrixes and tensors).

In order to jointly consider different data structures, the objective function described for coupled methods involving 2-way data has to be reformulated with the use of decomposition methods able to describe N-Way data

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \| \mathcal{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \|^2 + \left\| \mathbf{Y} - \mathbf{A}\mathbf{D}^{\mathsf{T}} \right\|^2$$

The objective function, to be minimized, describes in a unique solution the $\mathbf{Y}^{IxM}$ matrix, by bilinear parameters **A** and **D** (scores and loadings), and the tensor $\mathbf{X}^{IxJxK}$ with the loadings provided by a PARAFAC/CP model **A**, **B** and **C** (for the first, second and third mode respectively). The shared mode (samples mode) is described by the same set of vectors A^{IxR} where *R* is the number of resolved components.

The assumption intrinsic in the objective function of CMTF imposes that all resolved component are shared between tensor and matrix. Generally, when fusing different sources of information, the possibility that both are explainable by the same set of components is quite uncommon, since each technique can identify characteristics of the samples which cannot be detected by others (for example the fusion of NMR and HPLC signals in metabolomics).

For this reason, a new algorithm ACMTF [53] for the detection of shared and unshared components, which makes use of constraints able to highlight components which characterize at the same time both matrixes and tensors and components which have to considered independent, was introduced. In order to achieve the resolution of both shared and unshared factors, weights, related to the rank-one components in the third-order tensor and matrix, are introduced:

$$\min f_2(\lambda,\sigma,\mathrm{A},\mathrm{B},\mathrm{C},\mathrm{V}) = \left\|\mathcal{X} - [\![\lambda;\mathrm{A},\mathrm{B},\mathrm{C}]\!]\right\|^2 + \left\|\mathrm{Y} - \mathrm{A}\Sigma\mathrm{V}^{\mathsf{T}}\right\|^2 + \beta\|\lambda\|_1 + \beta\|\sigma\|_1$$

$$\text{s.t. } \|\mathrm{a}_r\| = \|\mathrm{b}_r\| = \|\mathrm{c}_r\| = \|\mathrm{v}_r\| = 1, \quad \text{for } r = 1,\ldots,R.$$

where λ^{Rx1} and σ^{Rx1} are the weights introduced in the new formulation, Σ^{RxR} is the diagonal matrix with entries of σ in the diagonal and $\beta \geq 0$ is a penalty parameter.

The 1-norm penalties are used to sparsify the weights λ and σ so that the unshared components will have norms equal to or close to zero for one of the two datasets.

In order to identify which are the common or the unshared components the weights vectors have to be analyzed. As examples, the weight vectors λ=[1 0 1] and σ =[1 1 1] indicate one individual component in the matrix, the weight vectors λ=[1 1 1] and σ =[0 1 1] indicate an unshared component in the tensors, the weight vectors λ=[1 0 1] and σ =[1 1 0] indicate one common and one individual component in each dataset.

In addition to the methods previously described, a full new family of coupled methodologies based on the well-known partial least squares algorithm, mainly used to deal with very complex data (metabolomics, proteomics) was introduced. In these contexts, algorithms such as multiblock partial least squares MBPLS [54], and its extensions for classification purposes such as Consensus multiblock OPLS-DA [55] and kernel based OPLS-DA (KOPLS-DA) [55] are emerging for their capability to handle megavariate data in which the informative part, related to the investigated responses, is often veiled by the redundant/not informative part.

3.6 References

[1] D. Cozzolino, H.E. Smyth, K.A. Lattey, W. Cynkar, L. Janik, R.G. Dambergs, I. Leigh Francis, M. Gishen, "*Combining mass spectrometry based electronic nose, visible–near infrared spectroscopy and chemometrics to assess the sensory properties of Australian Riesling wines*", Analytica Chimica Acta, 2006, 563(1-2),319-324

[2] P.M. Ramos, I. Ruisánchez, K.S. Andrikopoulos, "*Micro-Raman and X-ray fluorescence spectroscopy data fusion for the classification of ochre pigments*", Talanta, 2008, 75(4), 926-936

[3] C.V. Di Anibal, M.P. Callao, I. Ruisánchez, "*^{1}H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices",* Talanta, 2011, 84 , 829-833

[4] C. Alamprese, M. Casale, N. Sinelli, S. Lanteri, E. Casiraghi, "*Detection of minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy*", Food Science and Technology, 2013, 53, 225-232

[5] C. Pizarro, S. Rodríguez-Tecedor, N. Pérez-del-Notario, I. Esteban-Díez, J.M. González-Sáiz, "*Classification of Spanish extra virgin olive oils by data fusion of visible spectroscopic fingerprints and chemical descriptors*", Food Chemistry, 2013, 138, 915-922

[6] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L. Buydens, *"Breaking with trends in pre-processing?"*, TrAC Trends in Analytical Chemistry, 2013, 50, 96-106

[7] J. Gabrielsson, J. Trygg, J. *"Recent developments in multivariate calibration"*, Critical Reviews in Analytical Chemistry, 2006, 36(3-4), 243-255.

[8] Q. Guo, W. Wu, D.L. Massart, "*The robust normal variate transform for pattern recognition with near-infrared data*", Analytica Chimica Acta. 1999, 382, 87–103

[9] H. Martens, T. Næs, "*Multivariate Calibration*", John Wiley & Sons, New York, NY, 1989

[10] H. F.M. Boelens, R.J. Dijkstra, P.H.C. Eilers, F. Fitzpatrick, J. A. Westerhuis, "*New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection",* Journal of Chromatography A, 2004, 1057(1), 21-30.

[11] H.M. Heise, R. Winzen, "*Chemometrics in Near-Infrared Spectroscopy*", In: Siesler, H. W.; Ozaki, Y.; Kawata, S.; Heise, H. M. (Eds.) Near-Infrared Spectroscopy, Ed. Wiley-VCH, 2002, 147-149

[12] A. Savitzky, M.J.E. Golay, "*Smoothing and Differentiation of Data by Simplified Least Squares Procedures*", Analytical Chemistry, 1964, 36, 16-27

[13] E. Peré-Trepat, R. Tauler, "*Analysis of environmental samples by application of multivariate curve resolution on fused high-performance liquid chromatography–diode array detection mass spectrometry data*", Journal of Chromatography A, 2006, 1131 (1-2), 85-96.

[14] R. Tauler, A. de Juan, M. Maeder, "*Multiset Data Analysis: Extended Multivariate Curve Resolution"* In *"Comprehensive Chemometrics"*, edited by R. Tauler, B-Walczak, S.D. Brown, 2009,Elsevier.

[15] R. Leardi, "*Genetic algorithms in chemometrics and chemistry: a review*", Journal of Chemometrics, 2001, 15, 559-569

[16] W. Siedlecki, J. Sklansky, International Journal of Pattern Recognition and Artificial Intelligence, 1988, 02, 197-220

[17] S. Roussel V. Bellon-Maurel, J.M. Roger, P. Grenier, *"Fusion of aroma, FT-IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties"*, Chemometrics and Intelligent Laboratory Systems, 2003, 65, 209– 219

[18] H. Martens, M. Martens, " *Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)*". Food Quality and Preference, 2000, 11, 5-16

[19] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, "*Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy*", Applied Spectroscopy, 2000, 54(3), 413-419.

[20] M. Cocchi, R. Seeber, A. Ulrici, "*A. Multivariate calibration of analytical signals by WILMA (wavelet interface to linear modeling analysis)*", Journal of Chemometrics, 2003, 17 (8-9), 512-527

[21] A. Ulrici, G. Foca, C. Durante, A. Marchetti, L. Tassi and M. Cocchi, "*Multivariate analysis of analytical signals to decipher relevant chemical information*", in New Trends in Analytical, Environmental and Cultural Heritage Chemistry, M. P. Colombini and L. Tassi (Ed.), Transworld Research Network, Trivandrum (India), 2008, chpt 5. pp. 77-136. ISBN: 978-81-7895-332-8.

[22] M. Cocchi, C. Durante, G. Foca, M. Li Vigni, R. Leardi, A. Ulrici, "*Efficient variables selection in multivariate analysis of signals by coupling fast wavelet transform and genetic algorithms*". VI Colloquium Chemometricum Mediterraneum, St. Maximim La Sainte-Baume, France, 5-7 Septemper 2007.

[23] V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, C. Sterna, "*Elimination of Uninformative Variables for Multivariate Calibration*", Analytical Chemistry 1996, 68, 3851-3858.

[24] S. Wold, E. Johansson and M. Cocchi, PLS: "*Partial Least Squares Projections to Latent Structures*", in Hugo Kubinyi (Editor), 3D QSAR in Drug Design: Theory, Methods and Applications. ESCOM Science Publishers, Leiden 1993, pp. 523-550.

[25] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, "*Assessing feature relevance in NPLS models by VIP*", Chemometrics and Intelligent Laboratory Systems, 2013, 129, 76–86

[26] T. Rajalahti, R. Arneberg, F. S. Berven, K.M. Myhr, R. J. Ulvik, O. M. Kvalheim, "*Biomarker discovery in mass spectral profiles by means of selectivity ratio plot*", Chemometrics and Intelligent Laboratory Systems, 2009, 95, 35–48.

[27] S. Wold, N. Kettaneh, K. Tjessem, "*Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection*", Journal of Chemometrics, 1996, 10, 463-482

[28] J.A. Westerhuis, T. Kourti, J.F. Macgregor, "A*nalysis of multiblock and hierarchical PCA and PLS models",* Journal of Chemometrics, 1998, 12, 301-321

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

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

[31] S. Mas, R. Tauler, A. de Juan, "*Chromatographic and spectroscopic data fusion analysis for interpretation of photodegradation processes*", Journal of Chromatography A, 2011, 1218 .9260– 9268

[32] M. Silvestri, A, Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, "*A mid level data fusion strategy for the varietal classification of Lambrusco PDO wines*", Chemometrics and Intelligent Laboratory Systems, Submitted 2014.

[33] P.M. Ramos, I. Ruisanchez, "*Data fusion and dual-domain classification analysis of pigments studied in works of art*", Analytica Chimica Acta, 2006, 558, 274–282

[34] Y. Liu, S.D. Brown, "*Wavelet multiscale regression from the perspective of data fusion: new conceptual approaches*", Analytical and Bioanalytical Chemistry, 2004, 380, 445-452

[35] D.L. Hall, J. Llinas, "*An introduction to multisensor data fusion, proceedings of the IEEE*", 85 (1997) 6–23.

[36] M. Brady, "*Special issue on sensor data fusion—foreword, International Journal of Robotics Research",* 7 (1988) 2–4.

[37] L.I. Kuncheva, "*Combining Pattern Classifiers"*, Wiley, Hoboken, New Jersey, 2004.

[38] V. Steinmetz, F. Sevila, V. Bellon-Maurel, "*A Methodology for Sensor Fusion Design: Application to Fruit Quality Assessment*", Journal of Agricultural Engineering Research, 1999, 74, 21-31

[39] L.I. Kuncheva, "*Combining Pattern Classifiers*", Wiley, Hoboken, New Jersey, 2004.

[40] A. Dromigny, Y.M. Zhu, "*Improving the dynamic range of real-time X-ray imaging systems via bayesian fusion*", Journal of Nondestructive Evaluation, 1197,. 16, 147-160.

[41] R. Chatila, In Support de cours de stage: ''*Pour les syste`mes multicapteurs: la fusion de donnees*'', Polytechnique, Palaiseau, F, 1996, 1 – 38.

[42] G. Shafer, "*A Mathematical Theory of Evidence*", Princeton Univ. Press, Princeton, NJ, 1976.

[43] M.J. van der Laan, E.C. Polley, A.E. Hubbard, "*Super learner, Statistical Applications in Genetics and Molecular Biology",* 6 (2007).

[44] T.G. Doeswijka, A.K. Smilde, J.A. Hagemana, J.A. Westerhuis, F.A. van Eeuwijk, "*On the increase of predictive performance with high-level data fusion"*, Analytica Chimica Acta, 2011, 705 , 41– 47

[45] K. Van Deun, A.K. Smilde, M.J. van der Werf, Henk A.L. Kiers, I. Van Mechelen, "*A structured overview of simultaneous component based data integration*", BMC Bioinformatics, 2009, 10, 246-261

[46] A. K. Smilde, J. A. Westerhuis, S. de Jong, "*A framework for sequential multiblock component methods*", Journal of Chemometrics 2003, 17, 323-337

[47] S. de Jong, H.A.L. Kiers, "*Principal covariates regression: Part I. Theory*", Chemometrics and Intelligent Laboratory Systems, 1992, 14, 155-164.

[48] B. Escofier, J. Pagès, "*Analyses factorielles simples et multiples",* 3rd edition. Paris: Dunod; 1998

[49] B. Escofier, J. Pagès, "*Methode pour l'analyse de plusieurs groupes de vaiables: Application à la caractérisation de vins rouge*", Reveu de Statistique Aplliquée, 1983, 31, 43-59

[50] H. L'Hermier des Plantes, B. Thièbaut, "*Etude de la pluviosité au moyen de la méthode S.T.A.T.I.S*",Revue de Statistique Appliquée 1977, 25, 57-81

[51] H. A. Kiers, J. M. ten Berge, "*Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure*", British Journal of Mathematical and Statistical Psychology, 1994, 47, 109-126

[52] E. Acar, T. G. Kolda, and D. M. Dunlavy, "*All-at-once Optimization for Coupled Matrix and Tensor Factorizations*", 2011, KDD Workshop on Mining and Learning with Graphs.

[53] E. Acar, A. J. Lawaetz, M. A. Rasmussen, R. Bro, "*Structure-Revealing Data Fusion Model with Applications in Metabolomics*", Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC'13), 2013, 6023-6026

[54] L.E. Wangen, B.R. Kowalski, "*A multiblock partial least squares algorithm for investigating complex chemical systems*", Journal of Chemometrics, 1989, 3, 3–20

[55] J. Boccard, D. N. Rutledge, “A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion”, Analytica Chimica Acta, 2013, 769, 30– 39

CHAPTER 4

Data Fusion Applications

4.1 Low-level data fusion application: characterization of Aceto Balsamico Tradizionale di Modena (ABTM) and Aceto Balsamico di Modena (ABM)

4.1.1 Introduction

The Aceto Balsamico Tradizionale di Modena (ABTM) is one of the most established and valuable foodstuff produced in Italy. It is characterized by high density and viscosity, which give to the product the typical syrupy aspect.

The production of this PDO vinegar [1] is strictly regulated by a production disciplinary [2] which defines the production methodology, the allowed raw materials and the geographical area of production (District of Modena).

ABTM is produced starting from a unique raw material, the so-called "*mosto cotto*" (cooked must) obtained by cooking fresh must in open vessels until the initial volume is reduced to one third with respect to the original one. In the cooking phase, chemical and physical/chemical transformations, important for the production of typical aromas and color, occur (partial caramelization of sugars, furfurals etc.)

The production protocol imposes the use of only seven varieties of grape, cultivated in the District of Modena, in order to produce the fresh must, namely, Lambrusco, Ancellotta, Trebbiano, Sauvignon, Sgavetta, Berzemino and Occhio di Gatta.

The cooked must undergoes different steps of transformation which can be summed up in *i*) fermentative transformations *ii*) maturation *iii*) ageing, which occur in set of wooden casks named "*Batterie*".

A typical "*batteria*" consists of a set of five up to eight wooden casks (the kind of wood is important in imparting particular aroma to the final product) labelled with numbers in decreasing order from the biggest (70-90 L) to the smallest (10-20 L).

Two main fermentative transformations occur in the first period of production: the alcoholic fermentation (provided by Saccharomyces yeasts) transforms part of the sugars present in ethanol, which is then oxidized in the acetic fermentation to give acetic

acid (by means of Acetobacters). During maturation and ageing phases complex chemical and chemical/physical transformations occurs by means of enzymes produced by the microorganisms present in the casks.

The ageing of ABTM is a dynamic process extended in a long time period (decades) in which, by means of a particular technique called "*rincalzi*", every year freshly cooked must is added to the biggest flasks while little amounts of "aged" product is used to refill the smaller tanks. From the smallest casks, every year, a little part of the whole content is taken and bottled for commercialization. The whole process is shown in Figure 4.1.1.1.

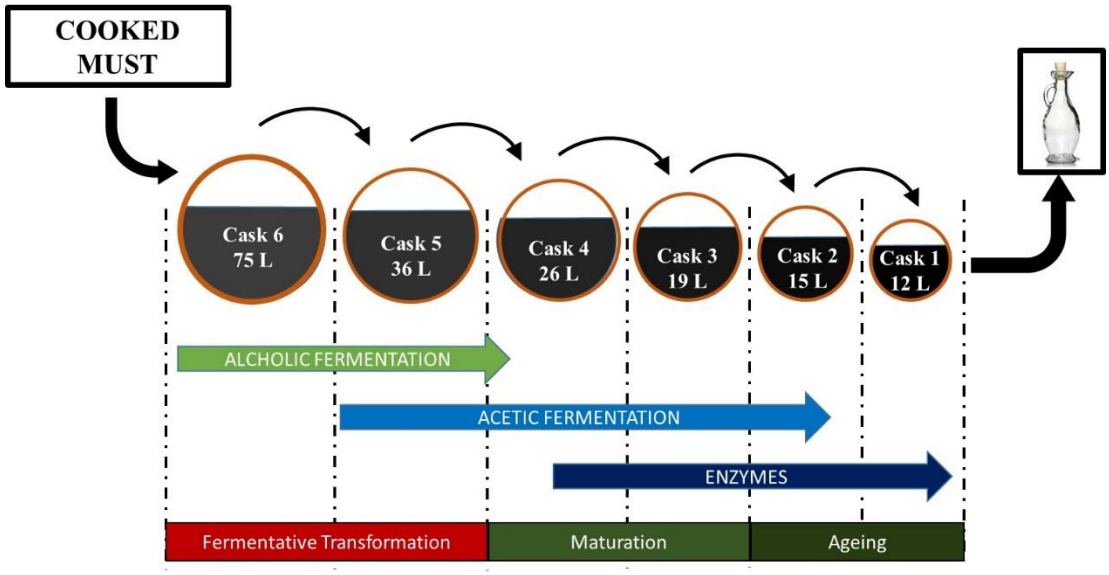

Figure 4.1.1.1 *Schematic representation of a six casks batteria*

For the commercialization of ABTM vinegar, an ageing period of at least 12 years is required by the production protocol, and more, the product has to be approved by a Panel of Master Testers which attribute a scoring variable whose values range from 0 up to 400.

Depending on the score given by testers and the ageing period, ABTM can be granted with the denomination of Extravecchio (score > 255, ageing > 25 years), the most valuable, or Affinato (score > 229, ageing > 12 years). The designation to one of the two classes strongly influences the price of the product, hence, the possibility to distinguish in an objective way the membership of an ABTM to Extravecchio or Affinato class is an important goal to attempt.

Besides, Aceto Balsamico Di Modena PGI (ABM) is also nationally and internationally known, the production and large-scale commercialization (93 million of liters per years) is very high and massive with respect to ABTM and is not difficult to find ABM in the supermarkets all over the world, moreover, the price of ABM is much lower than ABTM one.

In 2009, Aceto Balsamico di Modena received from the European Commission the Protected Geographical Indication [3], as in the case of ABTM a production regulation [4] indicates the allowed raw materials (concentrated grape must, wine vinegar and a certain amount of cooked grape must) and production methodologies.

ABM is defined as the product obtained from grape musts (fermented, cooked, concentrated) coming from grapes of Lambrusco, Sangiovese, Trebbiano, Ancellotta, Fortana, and Montuni. Wine vinegars and a small amount of caramel (2% max) are also allowed.

Fermentation and ageing are conducted, at least for 60 days, in wooden casks and mandatory characteristics have to be reached prior to commercialization: minimum total acidity 6%, density equal or higher than 1.06 g mL^{-1} (at 20°C).
Since prices of ABM span on the market from 3 €/L up to 60 €/L a commercial classification can be introduced, aimed to distinguish the most valuable (Fascia Alta) from the cheaper ones (Fascia Bassa).

In this section a low-level data fusion strategy, based on infrared spectroscopic signals (NIR and MIR) is proposed for the classification of ABTM and ABM samples.

4.1.2 Samples Description

The investigation was conducted on eighty-eight vinegar samples belonging to the families of Aceto Balsamico Tradizionale di Modena PDO and Aceto Balsamico di Modena PGI.

In addition, MIR and NIR spectra were acquired for six samples belonging to a "*batteria*" of production of ABTM with the intent to discover information related to the ageing and maturation of the products.

In Table 4.1.2.1 the investigated samples, classified for typology and class membership are reported

Table 4.1.2.1 *Vinegar samples analyzed by means of NIR and MIR spectroscopy*

Samples	**Number**
Aceto Balsamico Tradizionale di Modena ABTM	**42**
Extravecchio (EXT)	*21*
Affinato (AFF)	*15*
Batteria (BATT)	*6*
Aceto Balsamico di Modena ABM	**52**
Fascia Alta (FAA)	*26*
Fascia Bassa (FBA)	*26*

4.1.3 Results and Discussion

The instrumental settings used for the acquisition of the spectra are described in Appendix I. A total of 1763 data points were acquired on Mid-infrared region (from 4000 cm^{-1} to 600 cm^{-1}), whilst, 3035 data points were acquired for Near-infrared signals (10000 cm^{-1} to 4150 cm^{-1}), giving two datasets, namely "*MIR dataset"* and "*NIR dataset*" of dimensionalities 94 x 1763 and 94 x 3035 respectively.

In Figure 4.1.3.1 the MIR and NIR raw signals of two ABTM and ABM samples are reported.

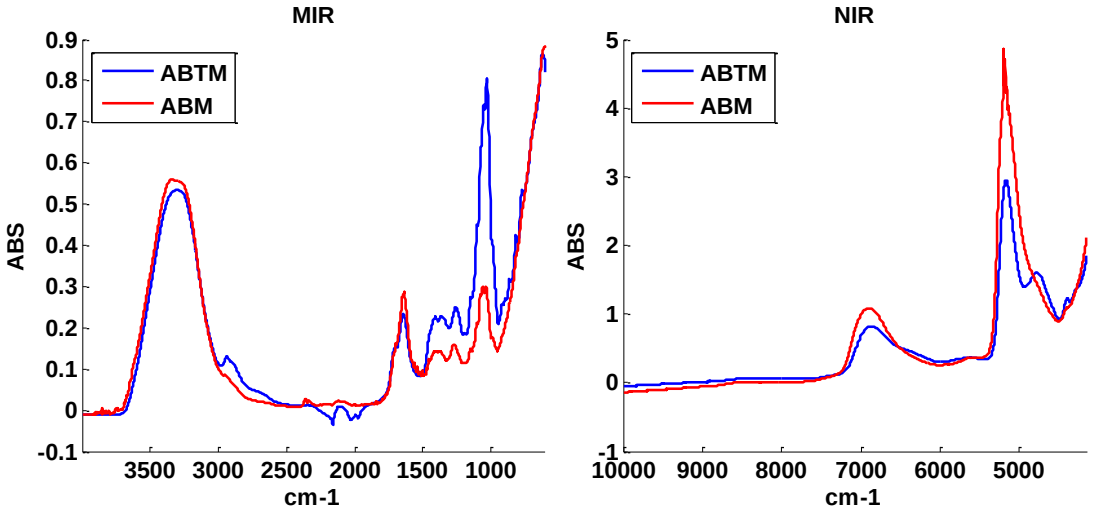

Figure 4.1.3.1 *Raw MIR and NIR signals of an ABTM samples (blue) and ABM samples(red)*

The differences between ABTM and ABM are quite clear both in NIR and in MIR signals. In particular, concerning the MIR spectra, higher absorbances can be noticed in the "fingerprint region" at around 1000cm^{-1} for ABTM samples. In the fingerprint spectral region of MIR signals, several vibrations, characteristic of more than one class of compounds can be found, in particular C–O stretching of sugars and alcohols, or all bending and stretching vibrations due to aliphatic chains.

When considering the NIR spectra, the most evident differences which can be highlighted are the different shape and intensity of the region from 5500 cm^{-1} to 4500

cm^{-1} and the intensities in the band around 6800 cm^{-1}. These regions are characterized by water overtones (5150 cm^{-1} and 6800 cm^{-1}) and combination bands of O-H stretching and C–O and C–C stretching (under 5000 cm^{-1}).

The principal difference between ABM and ABTM is the water content (in the former case spans from 55% up to 70% and in the latter from 25% up to 40%) that influences the concentration of all the solubilized species present in vinegars.

Prior to data analysis and the fusion step, the signals were preprocessed in order to remove baseline shifts by means of Weighted Least Square [5-6] and to cut noisy/not informative regions. The results of pretreatments are reported in Figure 4.1.3.2 and Figure 4.1.3.3

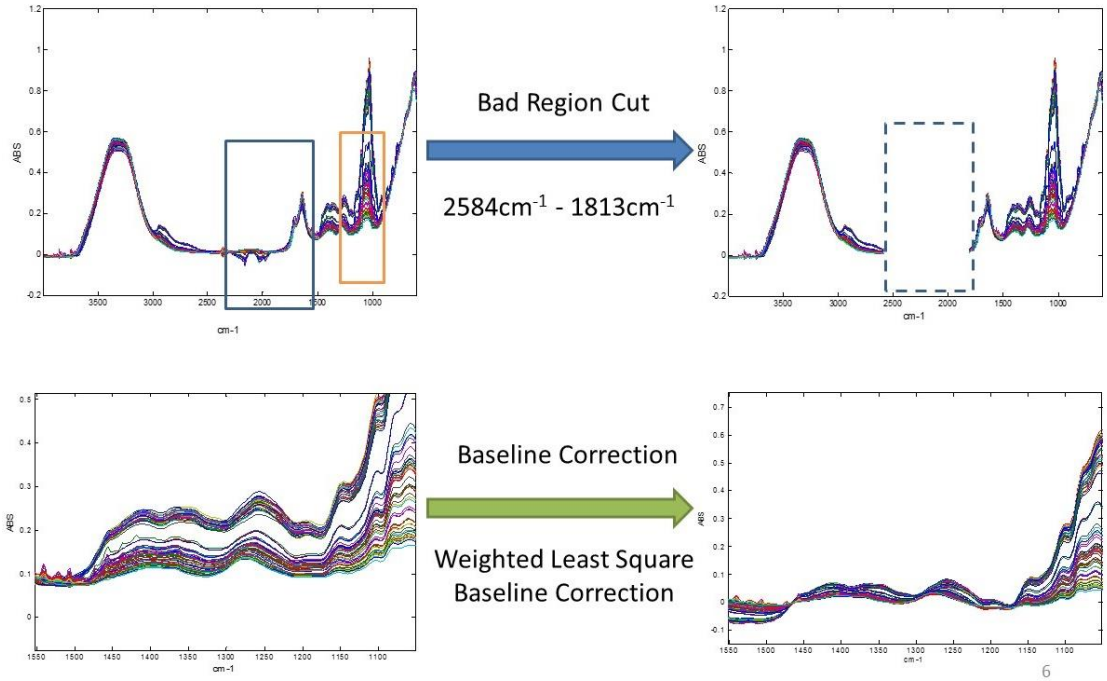

Figure 4.1.3.2 *Preprocessings performed on MIR spectra*

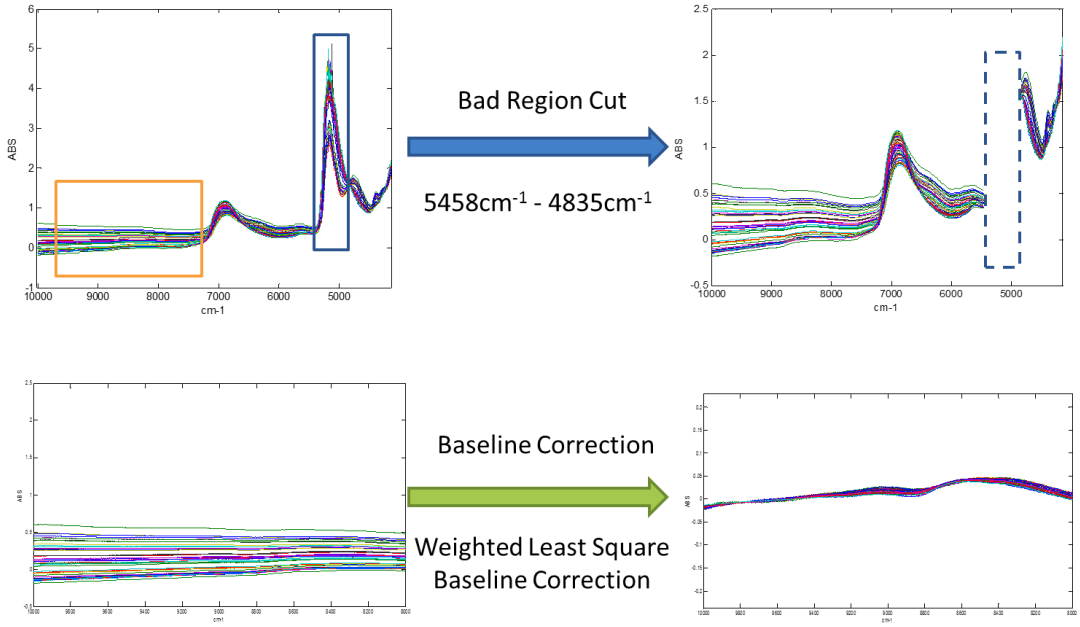

Figure 4.1.3.3 *Preprocessings performed on NIR spectra*

The concatenation of the two datasets was performed by scaling each block in order to give them the same variance (block scaling) since the dimensionalities after preprocessing of the separate data tables were different: the NIR spectra contains about twice of the variables present in MIR spectra. In Figure 4.1.3.4 a comparison of the concatenation with and without block scaling and mean centering is reported.

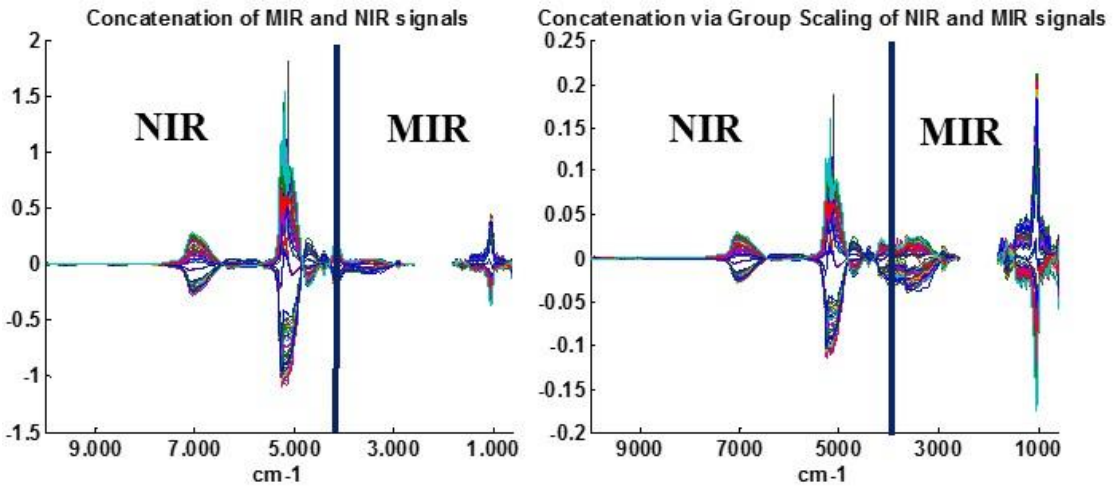

Figure 4.1.3.4 *Comparison between simply concatenation and group scaling*

The effectiveness of group scaling emerges from the inspection of the figure, in particular the contribution of the MIR spectral region results increased whilst the one of NIR spectral region reduced. The adoption of group scaling allows giving each block the same importance, in a way that an equal amount of information (variance) is brought to the final model by the two distinct spectroscopies.

An exploratory data analysis performed by means of a PCA model based on two components (chosen by minimum error in cross validation) able to describe 99.5% of the total variance, was conducted on the fused dataset.

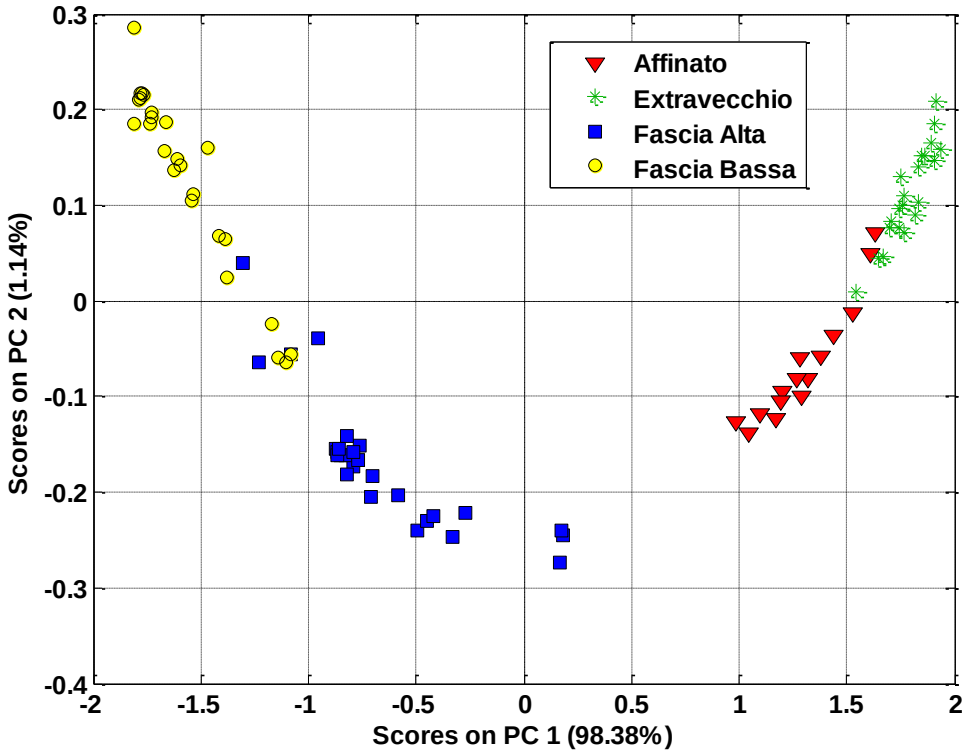

Figure 4.1.3.5 *Scores plot of the PCA model built using the fused dataset*

From the inspection the of scores plot of Figure 4.1.3.5 it emerges clearly a parabolic distribution of the samples which are also ordered, from the lowest quality Fascia Bassa

to the highest quality Extravecchio, counterclockwise starting from the second quarter to the first quarter.

The first principal component is able to distinguish mainly ABTM samples (positive scores values) from ABM samples (negative scores values), while the second one consents, with inverse trends, to separate the high quality products belonging to the two family of vinegars: in the case of ABTM the most valuable products are present at positive values for the second PC whilst at negative for ABM samples.

In order to understand which are the spectral zones responsible for the grouping of the samples, a deep investigation of loadings profiles, reported in Figure 4.1.3.6 has to be done.

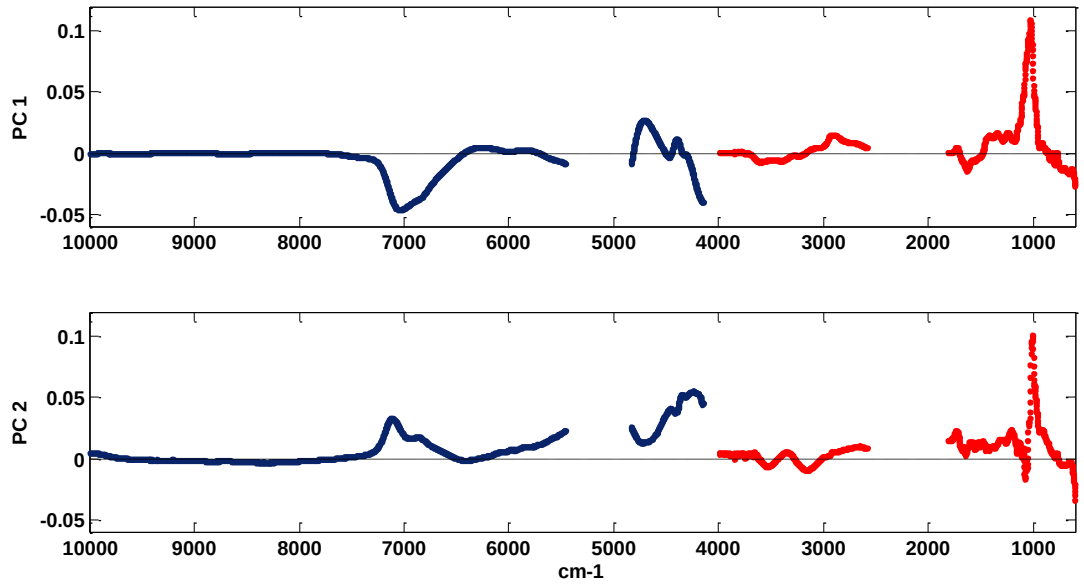

Figure 4.1.3.6 *Loadings plot of the first and second component of the PCA model built using the fused dataset, NIR region in dark blue and MIR region in red*

The first principal component, able to distinguish ABTM samples from the ABM ones, is inversely correlated with water content. In particular O-H signals such as the overtone band around 7000 cm^{-1} in the NIR region and O-H stretching above 3000 cm^{-1} in MIR region present negative contributions, while the signals attributable to dissolved compounds (sugar, poli-alcohols etc.) which are more concentrated when a smaller

amount of water is present, show positive values for example in the fingerprint region of MIR (around 1000 cm^{-1}) and combination region in NIR (around 4500 cm^{-1}).

The behavior explained by the second component, able to distinguish with opposite trends the most valuable products inside each of the two classes of vinegars, is quite difficult to interpret.

Assuming that all the information directly attributable to water content, and indirectly to dissolved compounds, is explained by the first principal component (in fact the inverse contributions present on first PCA are positively correlated in the second one), the residual part of the information held by the second PC provides a similar behavior between Extravecchio and Fascia Bassa samples and between Affinato and Fascia Alta samples. To understand the regions most of all responsible of the separations of the two sets of classes, the scatter loadings plot, reported in Figure 4.1.3.7, (quite uncommon to be used when spectral signals are analyzed) for the first two components has to be investigated.

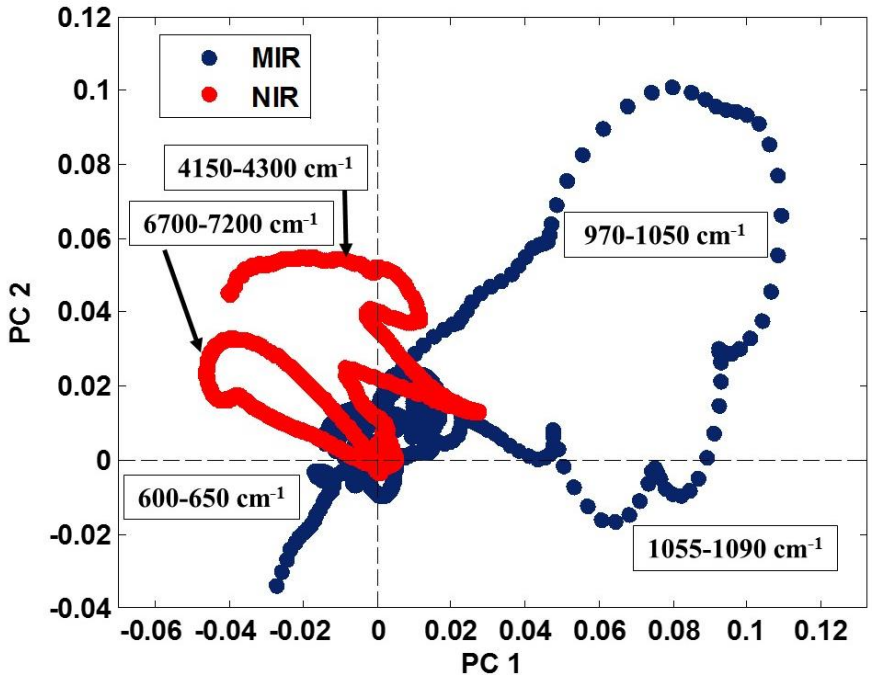

Figure 4.1.3.7 *Loadings plot PC1vsPC2 of the PCA model built using the fused dataset, NIR region in dark blue and MIR region in red*

In particular, few spectral regions seem to be mainly responsible for the separation among the different classes. The contributions able to differentiate Extravecchio and Affinato samples are referable mainly to MIR regions around 1000 cm^{-1}, and can be attributable to the presence, or not, of species having absorbance in the fingerprint region. Further investigation has to be performed in order to understand if during the ageing period a transformation of some compounds can be explained on the basis of the different absorbance in that region of MIR spectra.

On the other hand, the distinction of Fascia Alta e Fascia Bassa samples involves also the amount of water, and not only characteristic signals present in the fingerprint region of the MIR spectra, which is important for the positioning in the second quarter of the Fascia Bassa samples (the "less concentrated").

In order to demonstrate if the consideration reported above should be imputable to different period of ageing, hence, different "*quality*" of the products, six samples from a "*batteria*" were projected on the PCA model built using ABM and ABTM samples on fused dataset. The obtained scores plot is reported in Figure 4.1.3.8.

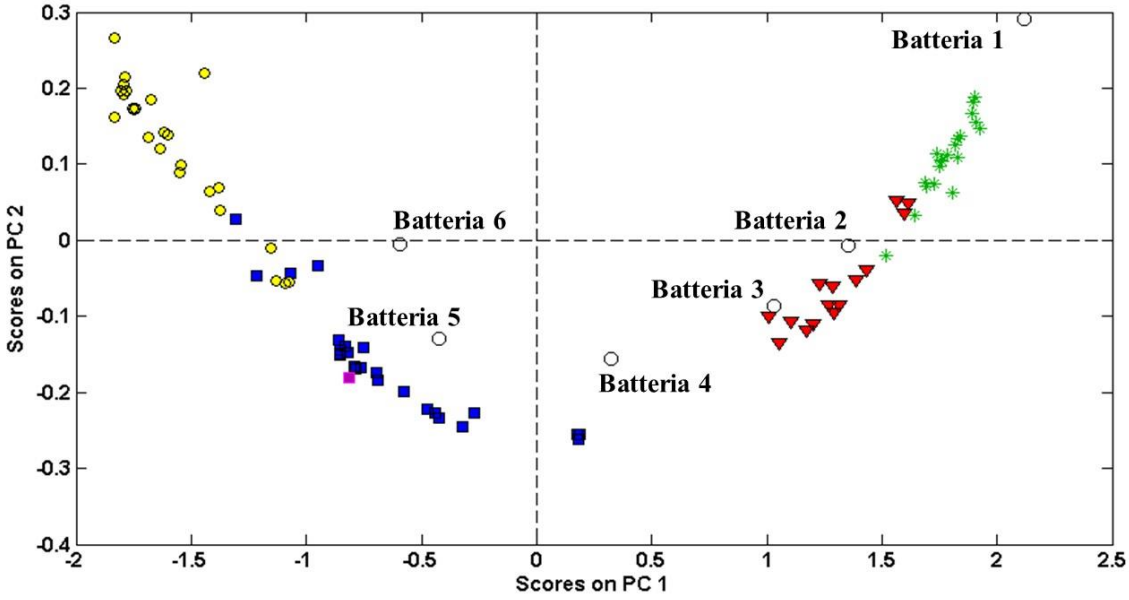

Figure 4.1.3.7 *Projection of "Batteria" samples onto the PCA model*

From Figure 4.1.3.7 it emerges that the trajectory depicted by ABM and ABTM samples is followed by "*batteria*" samples. In particular, the ageing seems to follow the parabolic shape in the scores space, in fact, Batteria 6 is the samples taken from the biggest flask while Batteria 1 from the smallest one. The ABTM samples put on the market are obtained by mixing different amounts of the products that are present in the final flasks of each batteria (Batteria 1 and 2), hence, the fact that Batteria 1 sample is placed at higher value for the second component with respect to all other ABTM samples can be well explained.

Since the labelling of Extravecchio and Affinato samples is conferred after a panel test evaluation, the possibility to develop an objective methodology able to confirm the class membership of the products put on market is an interesting aspect to face.

For this reason, the fused dataset was used to create a SIMCA [7] classification model aimed to distinguish the Extravecchio ABTM samples from the Affinato ABTM ones.

The one class SIMCA model, built using four principal components (chosen on minimum cross validation error in classification) and able to explain 99% of the total variance, was built using as training set fifteen of the twenty-one Extravecchio samples (the most added value products).

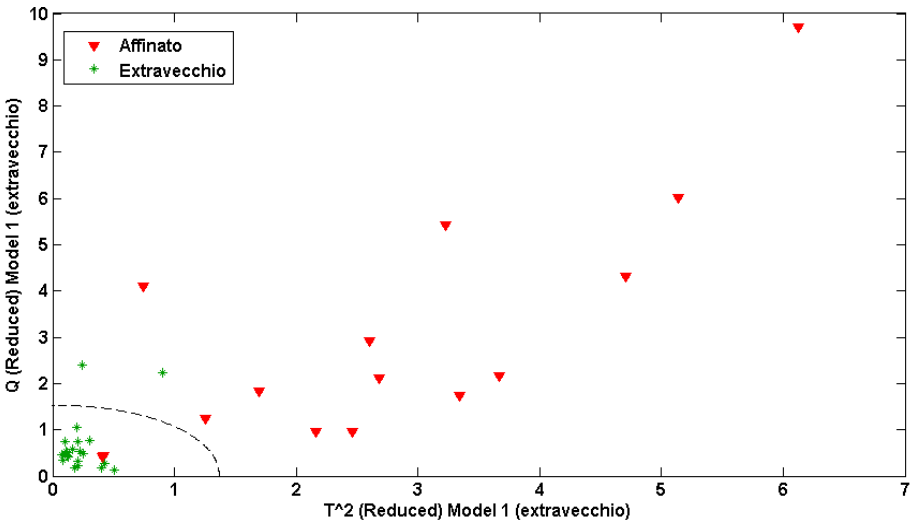

Figure 4.1.3.8 *T^2 and Q plot obtained from the SIMCA model obtained on fused dataset*

The remaining six Extravecchio samples and all Affinato ones were then predicted in order to evaluate the sensitivity and specificity of the classification model. From Figure 4.1.3.8, reporting Q and T^2 values for the obtained model it is possible to identify which are the samples correctly and incorrectly classified.

In particular, all samples beneath the confidence interval (dashed line) are accepted by the model as Extravecchio samples, all others are rejected.

All Extravecchio samples used for fitting the model and all the test set Extravecchio samples except two are correctly classify. A good specificity is also obtained, in fact, only two Affinato samples are incorrectly classified as Extravecchio.

Good results are obtained in the classification of the two distinct categories of ABTM samples (specificity 13/15 and sensitivity 4/6) using the low-level data fusion.

For a further optimization of the model, a new set of ABTM samples should be analyzed in order to integrate the existing one, trying to achieve even better results.

The SIMCA model shows in general, better classification performances if compared to the ones obtained using separate data sets (not reported). Moreover, the joint analysis of the two sets of variables allows pointing out useful information, by means of the combined interpretation of the results achieved in explorative data analysis, for the detection and explanation of the important spectral regions able to characterize the investigated classes of samples.

4.2 Mid-level data fusion application for the characterization of soil samples: a pilot study for a geographical traceability model

4.2.1 Introduction

Traceability models based on primary indicators, as described in Section 1.3, are focused on obtaining a direct link between parameters present in the soils in which a certain product is cultivated and the product itself.

The variability of primary indicators in the collected samples, the modifications that could occur on the concentration of metals in the production chain and the biological variability of plants (different uptake, selectivity to certain ions/complexes etc.) may influence dramatically the achievement of the model able to link the product to its territory of origin.

For these reasons, a pilot study aimed at obtaining information about the influence of some sources of variability, such as inter- and intra field variability for soil, seasonality, etc…, was conducted within the AGER project [8].

In particular, the geochemical variability of soils on which the cultivation of Lambrusco P.D.O wines insists was investigated with the goal to obtain useful information for the development of the second phase of the study in which an extended sampling procedure has to be performed.

Lambrusco PDO wines, typical of the District of Modena, are the most exported oenological products all over the world (750 000 hL/ year). According to their production regulation [9] four typologies of Lambrusco wines can be produced as PDO product, namely: Lambrusco Grasparossa di Castelvetro PDO, Lambrusco di Sorbara PDO, Lambrusco Salamino di Santa Croce PDO and Lambrusco di Modena PDO.
By means of a mid-level data fusion strategy, based on the analysis performed by MCR-ALS [10] on two distinct datasets containing geographical traceability indicators (metals concentrations with $^{87}Sr/^{86}Sr$ isotopic abundance ratio and X-Ray Diffraction on Powder – XRDP signals) the characterization of soils samples from four different

producers was performed with the intent to assess and understand the horizontal (within the fields and among the fields) and vertical variability (depth).
The joined approach performed with the use of data fusion allowed to point out useful correlation between the two sources of information and to well understand the geochemical variability of soils within the District of Modena.

This work was published [11], and it is attached in the last section (**Paper I**) of the Thesis, as cover article in Analytica Chimica Acta, 2013, 769, 1-9.

The novelty of the proposed approach is attributed to the use of Multivariate Curve Resolution in each step of the data fusion framework. In particular, in the lower level modelling, MCR was adopted as variable reduction tool in order to extract few chemically meaningful concentration profiles which, in the higher level modelling, will give rise to the fused dataset.

4.2.2 Samples Description

The production regulations of the three varieties of Lambrusco PDO wines define the area, belonging to the District of Modena, which are allowed for the cultivation.

Since the territory subjected to the cultivations is quite large (more than 90 km^2) and the producers numerous (more than four thousands) a pilot study conducted on four producers, selected by taking into account the geological, pedological and morphological property of the territories of the District, was performed.

Three of the four producers fields insist on in-plain region (producer A, B, D), while the last one insists on hill region (producer C). Depending on the dimension of the fields, from three to five corings were collected for each producer and then, each core was split in five aliquots of 10 cm of length, starting from 10 cm of depth to 60 cm. All depths were analyzed for the hill field and only lower and upper aliquots for the plain ones for a total of 47 samples. The collected samples are reported in Table 4.2.2.1

Table 4.2.2.1 *Analyzed samples*

Producer	***Altitude***	***Corings***	***Considered Depths***	***Samples***
A	*Plain*	3	2	6
B	*Plain*	5	2	10
C	*Hill*	5	5	25
D	*Plain*	3	2	6
		16		***47***

4.2.3 Results and Discussion

All the collected samples were analyzed by means of different analytical techniques. The description of instrumental parameters and samples pretreatments are reported in detail in Paper I and II [11-12] and in Appendix I. In particular, the dried and milled soil samples were analyzed by means of XRDP, a technique able to detect the mineralogical species present in soils, and after digestion (acid leaching assisted by microwave, via concentrated HNO_3), by means of MC-HR-ICP/MS, ICP/MS, F-AAS in order to quantify the concentrations of 34 metals (K, Na, Ca, Mg, V, Cr, Co, Ni, Cu, Zn, Ga, As, Rb, Sr, Cd, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu, Gd, Dy, Ho, Er, Tm, Yb, Lu, Tl, Pb, Th, U) and the isotopic abundance ratio of $^{87}Sr/^{86}Sr$.

A mid-level data fusion strategy was adopted in order to analyze simultaneously the different sources of information. By means of low-level data fusion, a dataset containing the information of metals concentration and the isotopic abundance ratio of $^{87}Sr/^{86}Sr$ was built; hence, the resulting data table was analyzed using MCR-ALS, able to extract four resolved profiles. On the other hand, three components were extracted using MCR-ALS on the XRDP dataset. The seven concentration profiles extracted in the first steps of data analysis were then scaled and merged together in order to obtain the "fused"

dataset, used to build a final MCR-ALS model. The global data fusion framework is schematized in Figure 4.2.3.1.

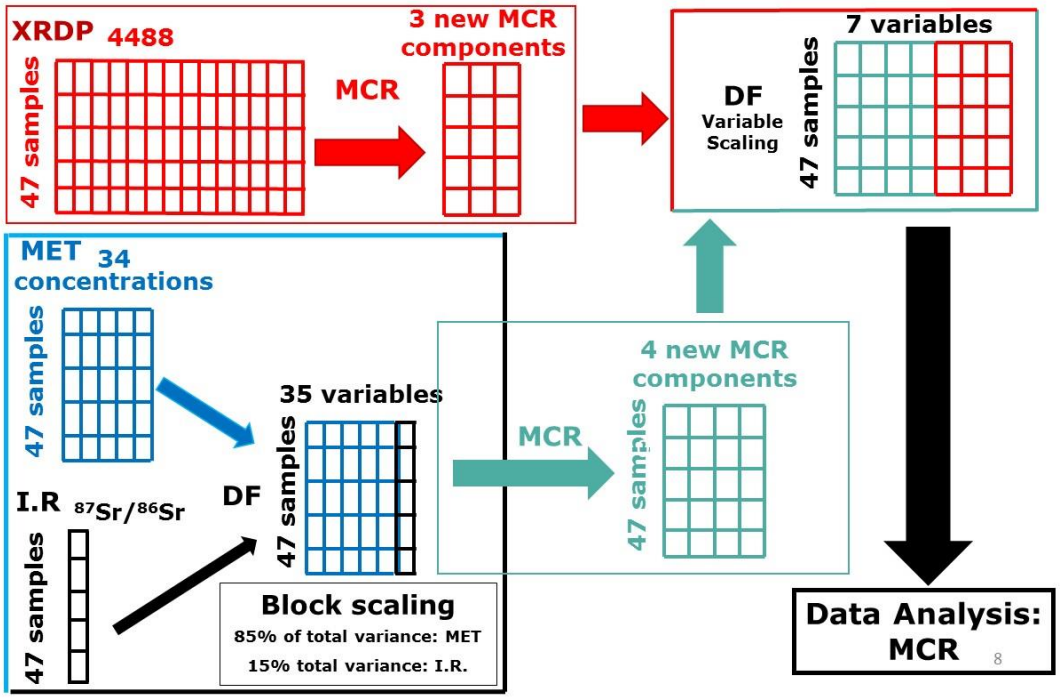

Figure 4.2.3.1 *Schematization of the mid-level fusion approach*

All the models built in each phase of the data fusion process will be treated separately, in order to avoid misinterpretations. The data from XRDP analysis will referred to "*XRDP dataset*", the ones containing the metals concentration to "*MET dataset*", the isotopic abundance ratio of $^{87}Sr/^{86}Sr$ are contained in the "*I.R. dataset*" while the table obtained in the last step of fusion is named "*DF MET-I.R.-XRDP dataset*"

Analysis of XRDP dataset

XRDP signals are characterized by having high complexity: each mineralogical phase presents in soils gives origin to a series of peaks that could cover the whole 2θ domain.

The ratio between the intensities of peaks corresponding to the same mineralogical phase are constant but are influenced by the amount of the same mineralogical phase in the investigated soils. The number of peaks present in each diffractogram depends on the complexity of the samples (number of mineralogical phases or chemical rank) and on the number of peaks of each species.

In this work, MCR-ALS was used as data reduction tool to extract few chemically meaningful variables from the original complex signals. When considering high rank data, as in the case of the "*XRDP-dataset",* the resolution of pure contributions using MCR-ALS is not an easy goal to reach. MCR-ALS was used in this work to achieve a qualitative interpretation of the hidden information present on the original data and not in the common quantitative way. The use of rank deficient models may not lead to the perfect resolution of the pure compounds, giving as a result overlapped peaks or components in which more than one mineralogical phase are present. Even if the MCR-ALS models obtained using a lower number of components with respect to the chemical rank are not able to quantitatively resolve the investigated system, it may occur that the main contributions, responsible of almost all the variability of the data, can be identified, as in in this case.

MCR-ALS used as exploratory tool has the advantage that the resolved profiles, even if not referred to "pure" compounds, bring a chemically meaningful sense, hence they may help the interpretation of the results in all the levels of the data fusion approach.

Prior to data analysis, the diffractograms were preprocessed [12] in order to remove the high values 2θ region without signals, to reduce noise and background effects and to minimize horizontal shifts by means of iCoshift algorithm [13]

Several models were evaluated for the achievement of the most suitable one on the XRDP data. A preliminary model was computed, based on three components according to SVD decomposition, and constrained for non-negativity, both in concentration and spectra profiles, using SIMPLISMA [14] to estimate spectra profiles, which were previously normalized.

The results obtained from the preliminary model (not reported) show good similarity among the resolved profiles and pure compound diffractograms such as calcite (the most stable polymorph of calcium carbonate) and quartz (silicon dioxide). In the third component spectrum, most of the bands were present in the low 2θ region and many clays related peaks could be identified in the middle region of the spectrum.

Quartz, clays (a class containing several mineralogical phases) and calcite, in order of abundance, are the main constituents, covering the majority of the compositional profile, of soils existing in the Modena district. In order to reduce the rotational ambiguity of the system, and to refine the preliminary model, selectivity constraints were applied. Selectivity constraints allow to force the resolution of the model in a way that the constrained regions (if selectivity is applied as in this case to spectra profile) result modelled only by one component. Since the implementation of selectivity constraints can affect strongly the resolution step, it was decided to force only few peaks for the calcite and quartz component. In particular, the XRPD region corresponding to 50.1 2θ was imposed to be selective for quartz, whilst the regions corresponding to 29.5 2θ, 39.4 2θ, 48.5 2θ were imposed to be selective for the "calcite" component. In all the constrained regions the presence of signals not referable to calcite or quartz is unlikely, since only rare, or at least very low in concentration, compounds can be found.

The implementation of constraints, as shown in Figure 4.2.3.2, allowed resolving peaks that were not forced to be modelled only by one components. Even not referred literally to "pure" compounds", the components resolution has benefited from the imposition of selectivity constraints, suggesting that the majority of the information present in each component is attributable to the main compounds. Hence, for convenience, aware of having chosen a rank deficient model, the terms "clays", "calcite" and "quartz" will be associated to the resolved spectra profiles.

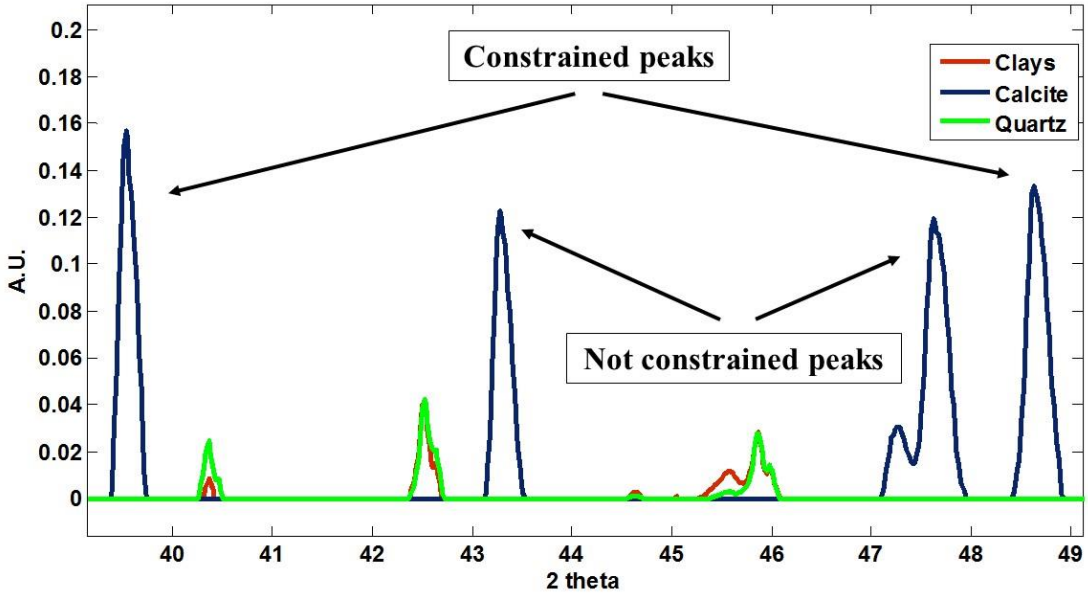

Figure 4.2.3.2 *Highlight of the 40–49 2θ region showing superimposed the constrained MCR-ALS resolved profiles for the three XRDP components*

The spectra profiles, resolved from the constrained MCR model, able to explain 92% of total variance with a lack of fit with respect to PCA of 1.1% are reported in Figure 2 of Paper I.

The information about samples, contained in the resolved concentration profiles, as described in Paper I, is here proposed in form of scatter plots instead of the more common bar graphs of the separate profiles. Even if the factors resolved by MCR-ALS are not orthogonal (while they are in the case of PCA), by means of a scatter plot representation, being the coordinates in the concentration space directly related to the concentration of the resolved species in the samples, an easy and intuitive inspection of the results can be achieved.

Similarities or dissimilarities about samples, projected in the concentration space, can be found, but it must be noticed that the concept of distance in the case of MCR is quite different from the one in PCA, in which the components are orthogonal and chosen accordingly to the maximum explained variance.

From Figure 4.2.3.3, where the concentration profiles are reported, some striking observations emerge about the distribution of the samples.

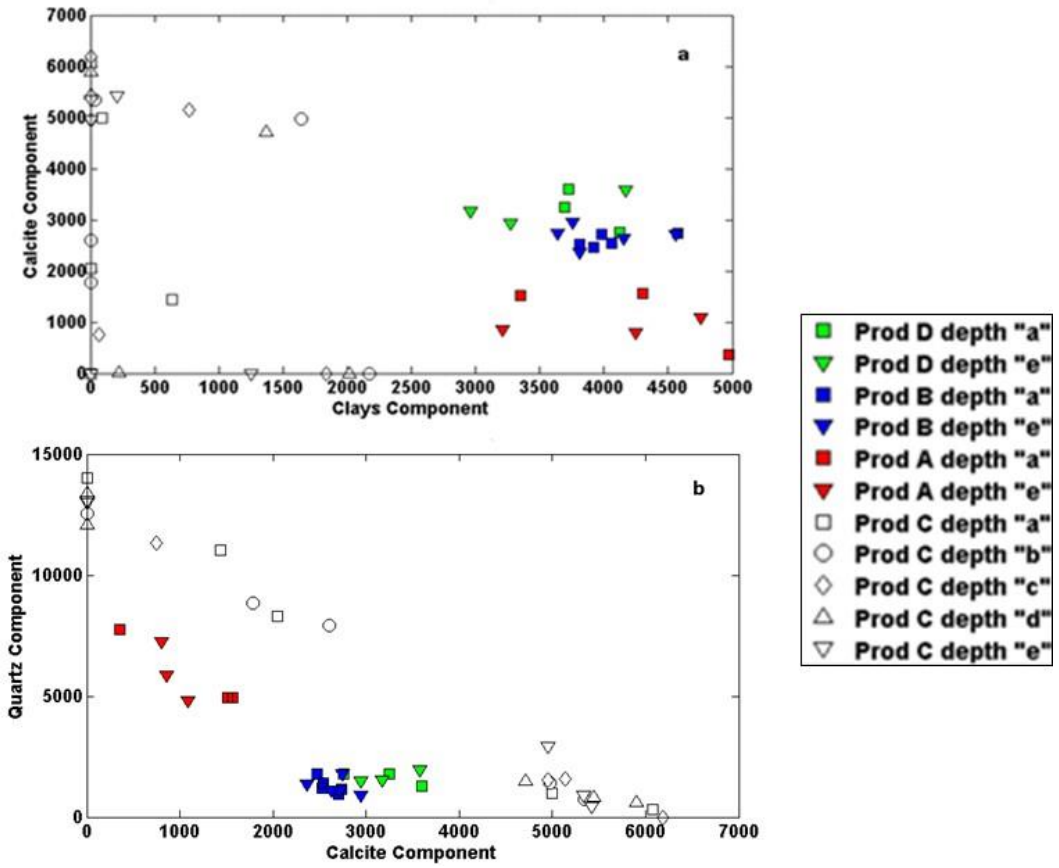

Figure 4.2.3.3 *MCR resolved profiles from XRDP dataset. a) "clays" concentration profiles versus "calcite" concentration profiles b) "calcite" concentration profiles versus quartz concentration profiles*

In particular, hill samples (Producer C) are separated from all in plain ones (Producer A, B and D) by the clays component, which indicates that the amount of clays is higher for the samples in plain with respect to the hill samples.

Figure 4.2.3.3 shows that the samples of Producer A can be distinguished by all other in plain ones for the different content in quartz, which is also responsible, together with calcite, of the separation of hill samples in two groups.

The separation in two clusters of the hill samples is in agreement with a certain degree of soil variability observed during the on field sampling of Producer C and confirmed by preliminary texture analysis of the same soils, where the whole content of sand, clay, silt and $CaCO_3$, were determined. Moreover, the intra-site heterogeneity of hill samples appear to be more pronounced with respect to the variability of all in-plain samples, reflecting the presence of soils characterized by having an abundant calcareous fraction with respect to others in which an high amount of quartz is present.

Analysis of MET and I.R. datasets

A low-level data fusion approach was used to combine the information related to metals content and the isotopic abundance ratio of $^{87}Sr/^{86}Sr$. As described in Chapter 2 the most important aspect to face when dealing with low-level data fusion is the way in which the different blocks of variables are scaled prior to the concatenation. The most common way to proceed, if "*a priori*" knowledge are not available, is to impose to each set the same variance (block scaling).

The isotopic abundance ratio of $^{87}Sr/^{86}Sr$ is an important indicator in geochemistry and geochronology, its values depend on several factor such as the age of formation of soils and the initial concentrations of ^{86}Sr, ^{87}Sr and ^{87}Rb. ^{87}Rb is not a stable isotope and decays to ^{87}Sr with half-life time ($t_{1/2}$) of more than ten to the tenth years.

Given all these properties, it was decided to enhance the importance of this variable with respect to all the others, giving to the isotopic abundance ratio of $^{87}Sr/^{86}Sr$ six times the importance of each metal concentration (85% of total variance imposed for the 34 metals concentrations and 15% to $^{87}Sr/^{86}Sr$). Prior to the weighing of the blocks, each variable was scaled to unit variance.

The resulting dataset, consisting of 47 rows (samples) and 35 (variables) was analyzed by means of MCR. It has to be noticed that when punctual variables are analyzed jointly using MCR, the resolution give as result profiles which can be attributed to common patterns of variation of the same variables and not to pure resolved profiles arising from single species as in the case of XRDP.

A four components MCR model was built using SIMPLISMA on concentration direction for the evaluation of initial estimates, imposing non-negativity constraints to both "spectra" and concentration profiles (99% of total variance explained, lack of fit with respect to PCA 0.8%).

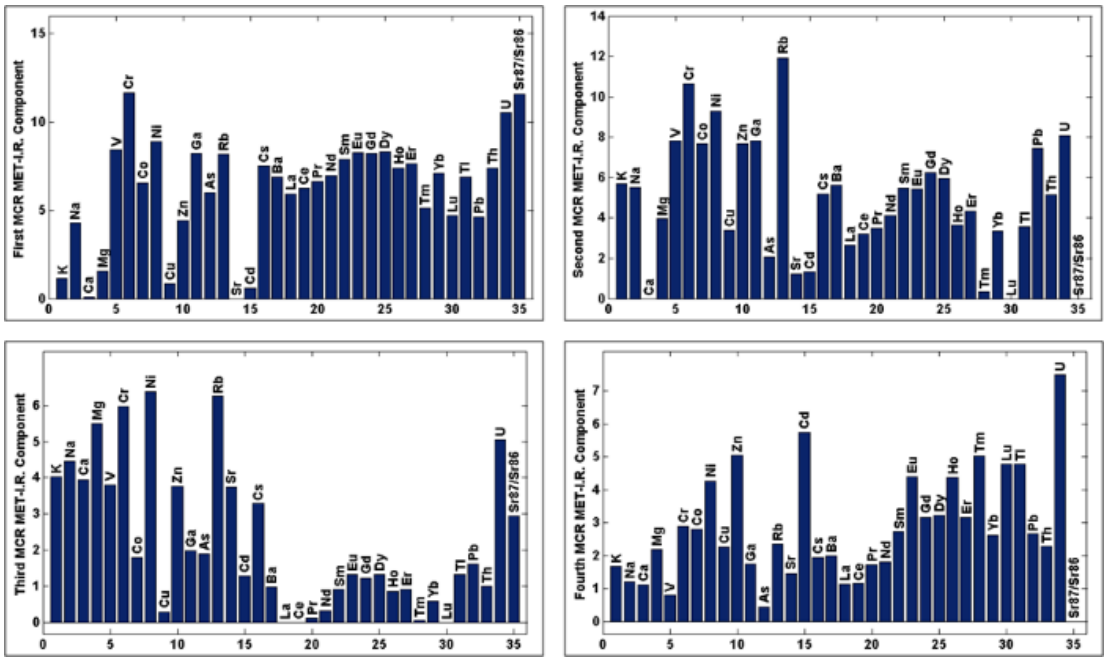

Figure 4.2.3.4 *MCR spectra resolved profiles from MET-I.R. dataset*

From the resolved "spectra" (variables) profiles, reported in Figure 4.2.3.4, it is possible to characterize the samples on the basis of the groups of variables which vary likewise.

In particular, some variables such as Cr, Ni and U present high values for almost all the components. On the other hand, elements such as calcium and magnesium present a high values for only one (the third) of the four resolved components. Sodium and potassium have and high contribution for the second and the third components but not for the other two. Rare earth elements contribution is manly observed in first, second and fourth components while the isotopic abundance ratio of $^{87}Sr/^{86}Sr$ is present with values equal to or close to zero in the components two, three and four. Considering the concentration profiles showed in Figure 4.2.3.5 a high grade of similarity can be observed with respect to the same profiles resolved in the XRDP dataset.

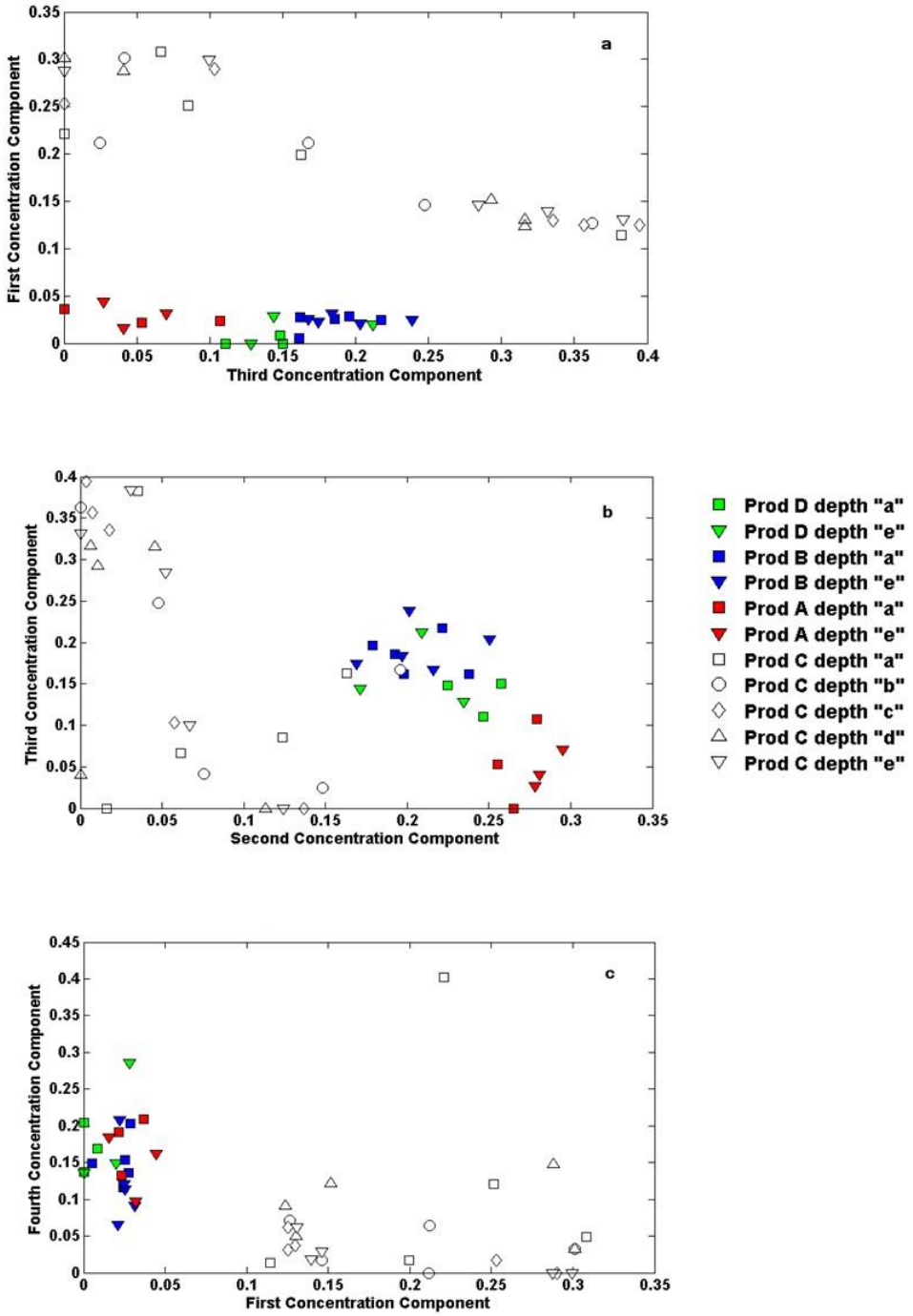

Figure 4.2.3.5 *MCR concentration resolved profiles from MET-I.R. dataset*

In particular, the second component resolved from the *MET-I.R.* dataset is able to differentiate the samples in a way very similar to the one obtained by the clays component resolved from *XRDP* dataset.

The same consideration can be extended for the third MET-I.R. component and the calcite ones. The first component, on the other hand, presents higher values for all the hill samples with respect to the in-plane ones. The fourth component is not able to highlight differences on the groups of soils, only one hill sample is present at very high values indicating a great amount for this soil of one or more of the variables with high contribution on the fourth component (e.g. cadmium).

Analysis of the "fused" dataset DF MET-I.R.-XRDP

A block scale preprocessing was used to concatenate the four resolved components from the MET-I.R. dataset and the three from XRDP dataset, in order to give to each block the same variance.

A three components MCR model (accordingly to singular value decomposition) was computed on the fused dataset, using SIMPLISMA on concentration profile for the evaluation of initial estimates and applying non-negativity constraints for the spectra and concentration profiles (95.9% of variance explained, lack of fit with respect to PCA 0.3%).

In order to understand better the correlations among the resolved factors, both bar graphs and a tridimensional scatter plot representations of the spectra profiles are reported in Figure 4.2.3.6.

The first factor is manly influenced by the "quartz" component from XRDP data set and the first one from MET-I.R.

The first MET-I.R. component is principally related to the values of the isotopic abundance ratio of $^{87}Sr/^{86}Sr$, as well as U, Cr and to the rare earths pattern. Thus, in the investigated soil samples to higher quartz content corresponds a higher isotopic abundance ratio $^{87}Sr/^{86}Sr$.

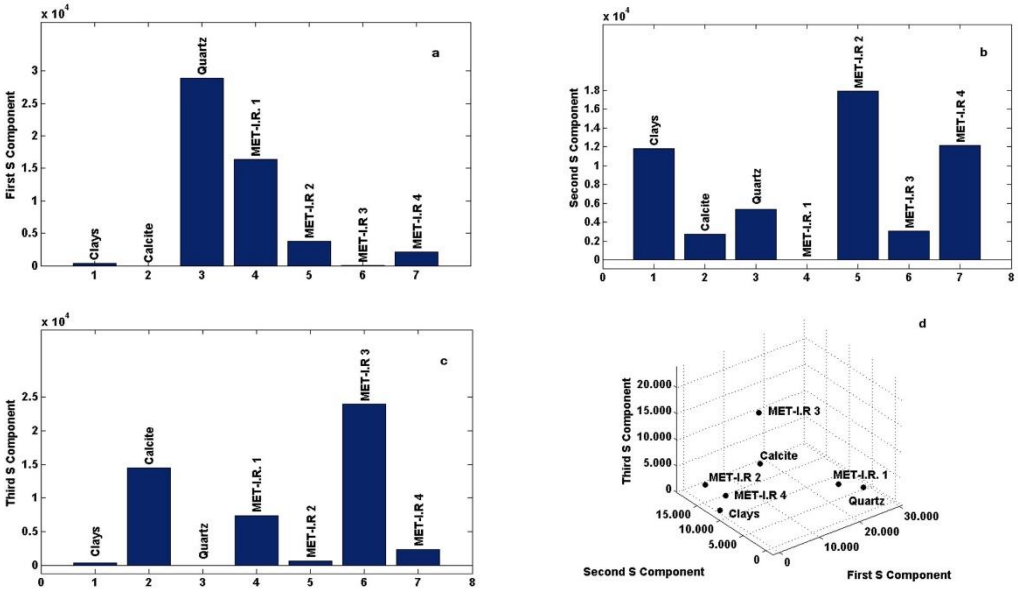

Figure 4.2.3.6 *MCR spectra resolved profiles from the fused MET-I.R.-XRPD dataset*

Concerning the second component, a conspicuous weight is attributed to the "clays" component from XRDP and the second and fourth from MET-I.R. dataset. The correlation between the XRDP clays component and two components from the metal dataset (influenced mainly by transition metals: zinc, nickel, cobalt, vanadium, cadmium, monovalent elements such as rubidium, potassium, sodium, and by the rare earths pattern) highlights the complexity of the possible relations among the different variables, that could also be imputable to the not complete resolution of the clays component, that is a combination of several crystalline structure of different clays (serpentine, muscovite, chlorite, illite, to name a few) typically presenting the inclusion of different metals in the lattice.

The third component is mainly influenced by the "calcite" factor from XRDP and the third from MET-I.R., as expected, since the latter results primarily correlated to the amount of calcium and bivalent elements, common in calcareous soils, such as magnesium, strontium, zinc, nickel and monovalent metals such as sodium, potassium and rubidium.

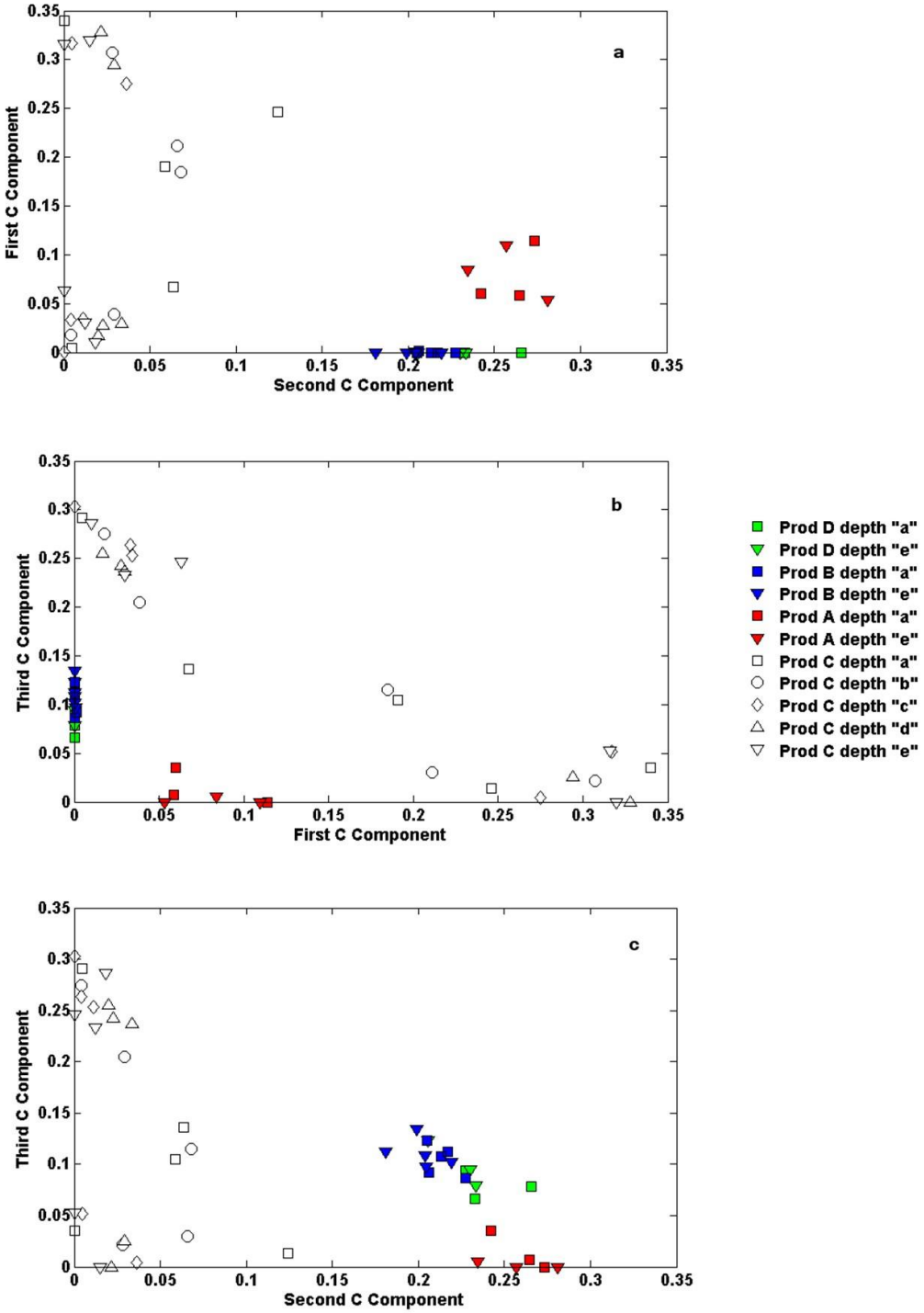

Figure 4.2.3.7 *MCR concentration resolved profiles from the fused MET-I.R.-XRPD dataset*

The resolved concentration profiles reported in Figure 4.2.3.7 show a distribution of samples in the new space of coordinates which is quite similar to the one obtained in the previous models, built on separate datasets.

The in-plain samples are placed at high values for the second component (as they showed on the second component in MET-I.R. model and on the "clays" component in the XRDP model) but are better separated by the first and second component with respect to the results obtained from separate datasets. In particular, the samples from Producer A present more positive values for the first component and lower ones for the third, indicating a great amount of quartz and strontium isotopic ratio, the most important variable for the second component of the MET-I.R. model.
The hill samples present high variability with respect to the samples from in-plain region, and can be differentiated in two clusters by means of the first and third resolved profiles. These factors are related to variables from XRDP and MET-I.R. dataset which appear to be mutually complementary. The presence of a high amount of one of them is linked to the presence of a lower one for the other component. These results are in good agreement with preliminary information regarding the textures analysis of the same soils, such as the composition of sand and silt. This variability is so pronounced that some samples from the same hole but at different depth present high values for the third component in some cases and lower for some other.

Concluding, the proposed mid-level fusion approach allowed operating with data arising from different techniques and characterized by having very different dimensionalities. The joint interpretation of the results obtained after the fusion step resulted helpful for understanding the correlation among the different blocks of variables and to strengthen the hypothesis made in the early stages.

Regarding the soils characterization, the use of different sources of information allowed to enhance the differentiation of soils, especially the ones from in-plain region. Moreover, MCR-ALS proved to be a powerful variable reduction tool integrated in the data fusion process and demonstrate the capability of this tool of achieving good results

when used on fingerprinting techniques data in a way similar to other exploratory analysis tools.

4.3 Mid and high-level data fusion application for the varietal discrimination of PDO Lambrusco wines

4.3.1 Introduction

As reported in Section 4.2.1, three P.D.O. Lambrusco wines can be obtained accordingly to their production regulations: Lambusco Grasparossa di Castelvetro P.D.O, Lambrusco di Sorbara P.D.O and Lambrusco Salamino di Santa Croce P.D.O.

The ampelographic compositions, set by the production codes, are defined as follow: at least 60% of Lambrusco di Sorbara grapes, up to 40% of Lambrusco Salamino grapes, up to a maximum 15% of other Lambrusco grapes, either of one variety or in combinations, for Lambrusco di Sorbara; for Lambrusco Salamino di Santa Croce and Lambrusco Grasparossa di Castelvetro at least 85% of grapes from vines of the same name, respectively, and the remaining 15% of other Lambrusco (Ancellotta, Fortana, and Malbo Gentile) grapes. The different grapes used for the production of Lambrusco wines belong to the same family; moreover, the production methodology is similar for the three P.D.O. products which are obtained by Charmat process, consisting of a second fermentation performed in pressurized tanks. For these reasons, the varietal classification of Lambrusco wines is a complex task to face, since the compositions and characteristics of samples are quite similar. A mid-level data-fusion approach is proposed, in order to achieve a classification model aimed to distinguish the three varieties of Lambrusco wines, using the information arising from three different analytical techniques: ^{1}H-NMR (proton nuclear magnetic resonance), HPLC-DAD (high performance liquid chromatography coupled with diode array detector) of the phenolic fraction, and EEM (excitation emission fluorescence matrix). Being the output of the analytical techniques characterized by different dimensionality and orders (matrixes and tensors), the variable extraction step was conducted with the use of several multivariate data analysis tools: MCR-ALS, PCA and PARAFAC.

For each dataset arising from the three analytical platforms, a model for the extraction of latent variables, able to describe the analyzed samples, and a classification model for the identification of the varieties of the wines were computed. For all classification

models, a classification rule was adopted that samples are assigned to the class for which the predicted y-value is higher, e.g. if the predicted vector of responses for an unknown sample, is: [-0.2 0.7 0.1] (in the case of a three classes problem), it will be assigned to class two. In the final part of the data fusion framework, the variables extracted in the lower level step were merged together. The resulting data table was used to investigate the correlations among the different sources of information and to build a super-classification model. The results were then compared with the ones obtained using a high-level data fusion approach, in which the classification indexes from the separate datasets were used to evaluate the fused classification indexes.

4.3.2 Samples description

A total of fifty-eight Lambrusco samples (twenty Lambrusco Salamino, nineteen Lambrusco Grasparossa and nineteen Lambrusco di Sorbara), were analyzed by means of the three analytical techniques (Appendix I). Due to the goal to pursue, the set of samples was split in two groups (training set forty-five samples and test set thirteen samples), accordingly to Duplex [15] algorithm performed on ^{1}H-NMR dataset. The same split was maintained for the other data sets and the fused data set, after checking, by exploratory data analysis, that each sets spanned the whole variability domain. In each step of the data fusion framework, the models were built using the training set whilst the samples belonging to the test set were projected on them in order to evaluate the classification performances of the obtained models. In Table 4.3.2.1 the investigated wine samples are summarized.

Table 4.3.2.1 *Analyzed Lambrusco samples*

	Lambrusco Grasparossa	*Lambrusco di Sorbara*	*Lambrusco Salamino*	*Total*
Training Set	14	15	16	**45**
Test Set	5	4	4	**13**
	19	19	20	***58***

4.3.3 Results and Discussion

The information arising from three distinct analytical techniques was used to build the mid and high-level data fusion framework as reported in Paper V [16] and schematized in Figure 4.3.3.1.

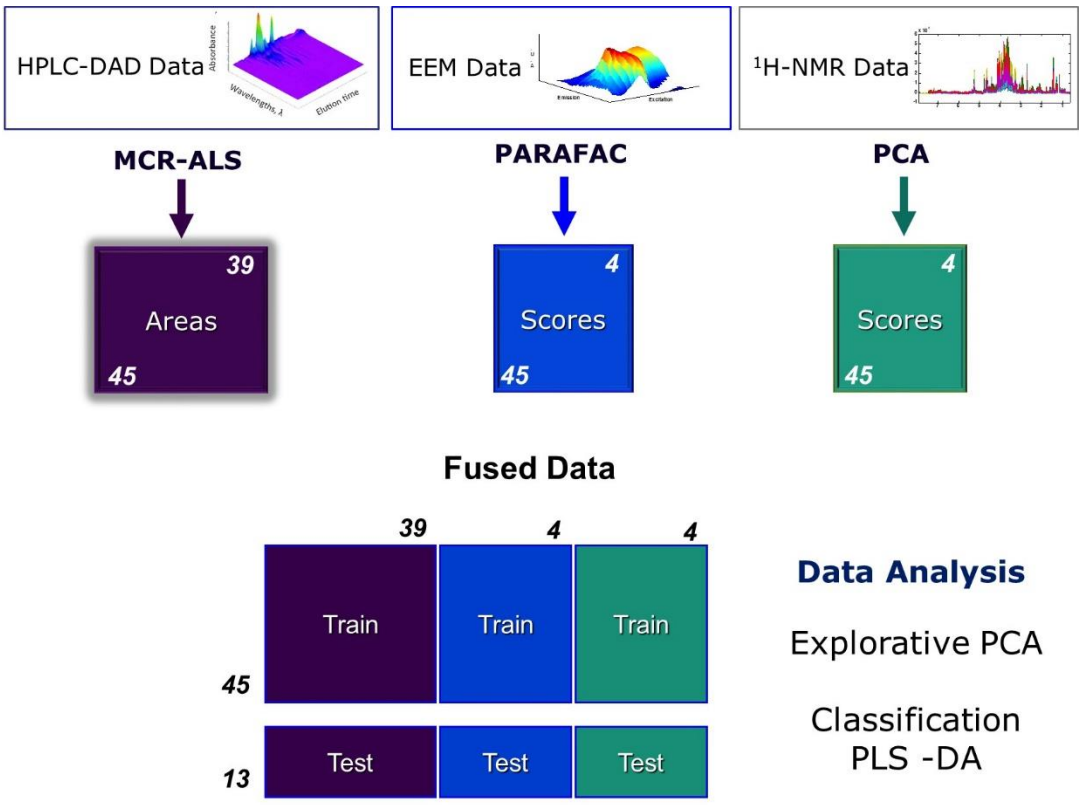

Figure 4.3.3.1 *Schematization of the data fusion framework*

In particular, from the EEM analysis a "cube" of data was obtained containing the fluorescence properties of the wines (excitation and emission profiles), related with the presence of compounds able to absorb light and emit photons with lower energy (higher wavelengths).

This kind of analysis was conducted on degased samples and gives as output a tridimensional structure of data in which the excitation and emission profiles of the

active species are present for each samples. A PARAFAC model was used to extract four factors involved in the final step of fusion.

Samples preparation and instrumental parameters for the acquisition of ^{1}H-NMR spectra are described in Paper II [17] and summarized in Appendix I. The NMR signals are characterized by having high complexity and overlapped peaks and are organized in a table consisting of a number of rows equal to the number of samples and a number of variables equal to the acquired points (chemical shifts) in the spectra. By means of PCA four components were extracted and used for the creation of the final fused dataset.

The description of analytical parameters and samples pretreatments related to the acquisition of HPLC-DAD data is reported in detail in a previous work [18], summarized in Appendix I and briefly described in next sections. A multiset MCR approach was conducted in order to resolve (starting from eight separate elution windows) the concentration of the compounds present. The thirty-nine resolved concentrations were merged together with PARAFAC (EEM dataset) first mode loadings and PCA (from ^{1}H-NMR) scores in order to obtain the "fused" dataset used for the joint analysis of all the sources of information.

In order to understand the steps involved in the data fusion framework, the results obtained from the modelling of the information arising from each techniques will be reported separately and finally the considerations about the fused data table will be furnished.

EEM data analysis and modelling

The signals were preprocessed as described in [16] in order to remove regions affected by noise and Rayleigh scattering. Due to the structure of the data (tensor), in order to extract the salient information present (variable reduction), a PARAFAC model was computed imposing non-negativity constraints for the second and third mode (excitation and emission profiles). A four factors model, built using the training set samples, was chosen as best compromise between explained variance (99%), core consistency and split half analysis results [19]. The test set samples were then projected on the model in

order to compute the first mode loadings used for the evaluation of a classification model based on NPLS-DA

The obtained four resolved profiles are in good agreement with the ones reported in literature [20]. In particular, as reported in Figure 4.3.3.2, the first resolved factor presents a broad maximum in excitation at 260 nm – 280 nm and a corresponding maximum in emission at 360 nm, very similar to the profile of vanillic acid. The second factor has fluorescent properties close to p-Coumaric acid, trans-resveratrol, trans-piceid and gentisic acid. The third factor well matches the wavelengths of catechin and epicatechin, having maximum in excitation at 280 nm and 320 nm in emission. The fourth factor cannot directly be attributable to a molecule or class of molecules and deepest investigations must be performed in order to characterize it

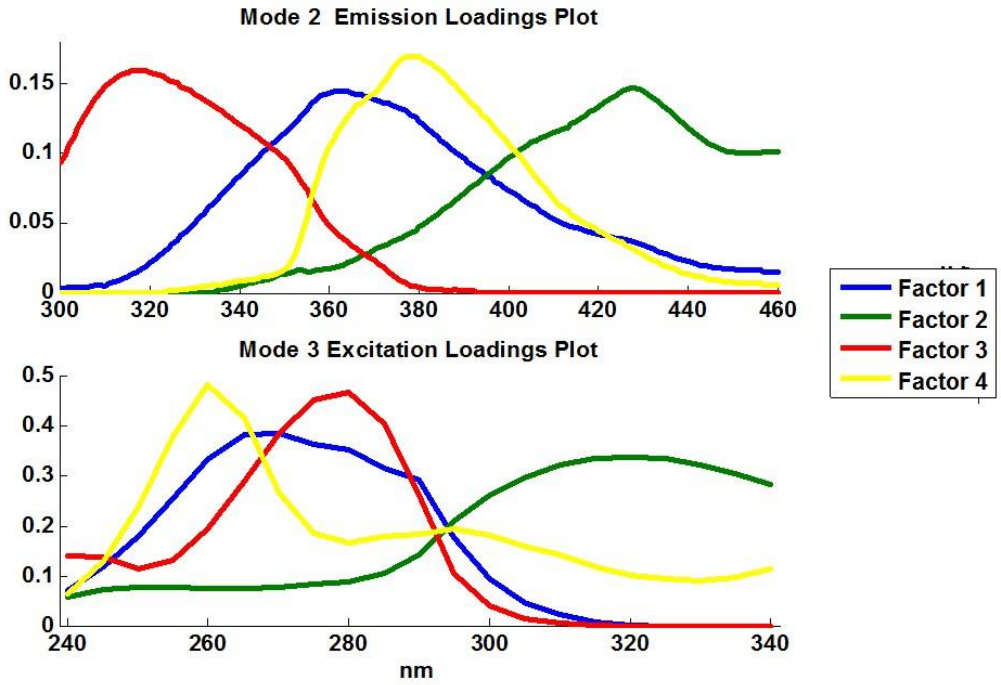

Figure 4.3.3.2 *Second and Third mode loadings obtained from the four factors PARAFAC model built on EEM data*

Regarding the first mode profiles, reported in Figure 4.3.3.3, an effective separation of wine samples on the basis of their varietal composition cannot be individuate.

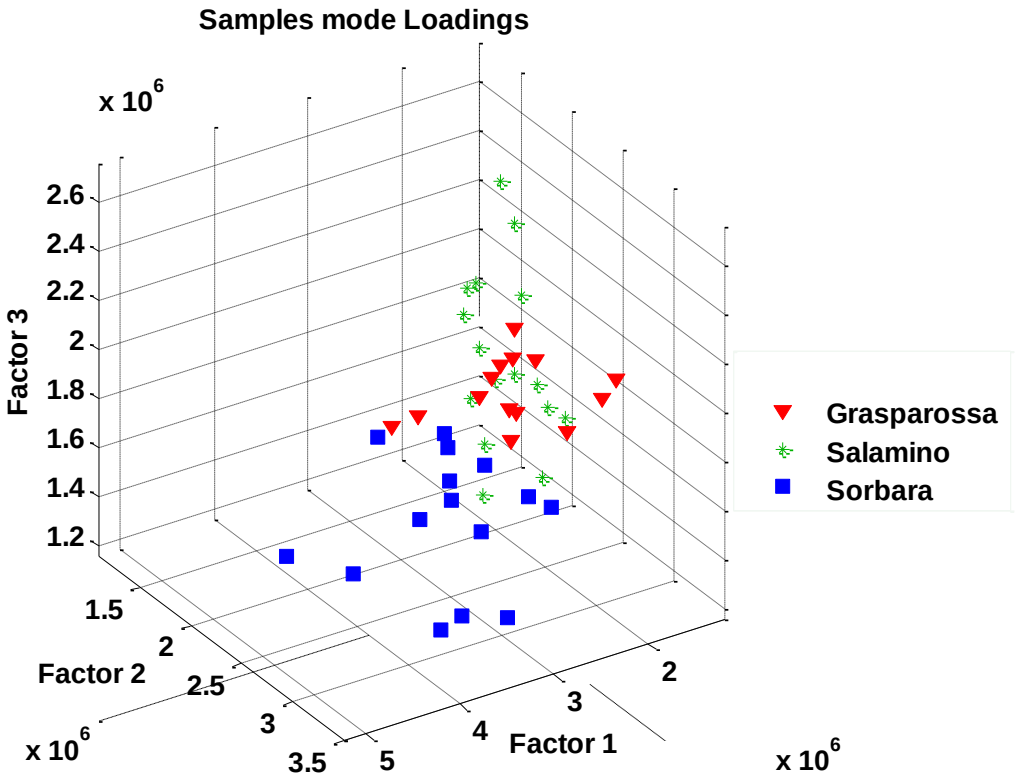

Figure 4.3.3.3 *First mode loadings obtained from PARAFAC model built on EEM data*

On average, Sorbara samples present higher values with respect to Grasparossa and Salamino samples for the first factor, indicating a greater amount of vanillic acid or compounds with the same fluorescent properties.

A nine components NPLS-DA classification model (lowest cross validation classification error rate) was computed. The classification performances, reported in Table 4.3.3.2 together with the ones obtained from NMR and HPLC-DAD dataset, highlight the fact that the information present from EEM signals is not completely adequate to differentiate the samples belonging to the three distinct classes.

HPLC-DAD data analysis and modelling

The output arising from HPLC-DAD analysis are arranged in a data cube: the first mode is related to the samples, the second to elution times and the third to the absorbances at the different wavelengths. This kind of data structure can be analyzed by means of MCR-ALS using the so called "multiset approach" which requires a preliminary step of unfolding (column-wise augmentation) in order to reduce the order of the data (from tensors to matrix) as described in Chapter 2.2. As in the case reported in Chapter 4.2, the data to be modelled using MCR are characterized by high complexity and chemical rank. For these reasons, in order to reduce the chemical rank of the system, and to make feasible the attempt to resolve the HPLC-DAD data in a quantitative way (contrary to what was done for XRDP data, modelled in a qualitative/fingerprinting way), it was decided to split the whole signals in eight separate elution windows, which were modelled independently. This approach allowed dealing with subsets of the whole data, characterized by having a lower chemical/mathematical rank, more suitable for the application of MCR-ALS. For each elution window, a MCR model was computed applying non-negativity, unimodality and local rank (selectivity) constraints when necessary [18]. The choice of the number of components for the resolution of each elution window was based on the number of compounds present and accordingly with singular value decomposition. The number of factors chosen for each elution windows and the variance explained by each separate model are reported in Table 4.3.3.1. Thirty-nine concentrations were extracted, the same used in the final step of the data fusion framework in combination with the variables extracted from NMR and EEM datasets. The elution windows will be mentioned in the text with the name of a compound, for example Gallic Acid or p-Coumaric Acid, present in that windows. Two windows, namely Windows1 and Windows2 were mathematically resolved by the MCR-ALS approach but not labeled as others since no one of the compounds present was confirmed using reference standards. In Figure 4.3.3.4 a schematization of the elution windows is reported.

Table 4.3.3.1 *Description of the models obtained from HPLC-DAD dataset*

Eluition Windows	***Resolved MCR components***	***Explained Variance %***
Gallic Acid	3	99.2
(+)-Catechin	3	98.0
Window 1	3	97.1
Caffeic and Syringic Acids	4	98.8
p-Coumaric Acid	7	99.3
Window 2	9	98.6
Myricetin	4	99.5
Quercetin	6	99.7

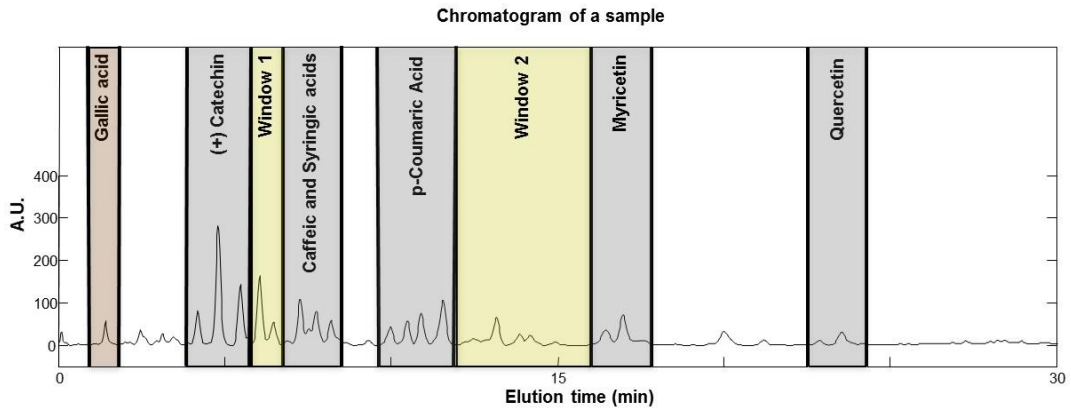

Figure 4.3.3.4 *HPLC-DAD TIC chromatogram of a sample: the subdivision in elution windows is shown*

A four components PCA model, able to explain 60% of the total variance, was built using the thirty-nine autoscaled concentrations resolved by the MCR models. The scores values obtained for the first two components, reported in Figure 4.3.3.5 highlight a rough separation of the samples belonging to the three different varieties of Lambrusco wines. In particular, all Grasparossa samples are present at negative values

for the second component, almost all Sorbara samples have negative contributes for the first component and the majority of Salamino samples has positive values for the first one; Salamino category overlap with the other two

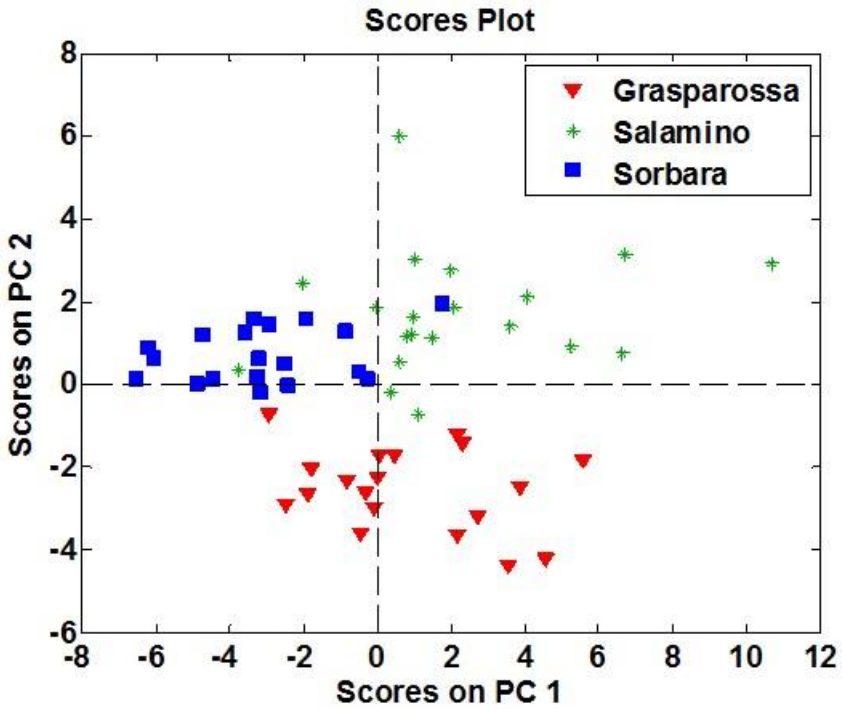

Figure 4.3.3.5 *Scores plot of the PCA model built using "HPLC-DAD dataset"*

From the loadings profiles reported in Figure 4.3.3.6 it is possible to point out which are the most important concentration profiles, the ones which mainly influence the position of the samples in the scores space.

The first component loadings profile highlights that all variables present positive contributions. Since Sorbara samples are placed at negative values for the same component, the loadings profiles suggest that the global amount of phenolic compounds in Sorbara samples is lower with respect to Grasparossa and Salamino samples.

The loadings of the second component, able to distinguish Grasparossa from Salamino samples, present a more complex profile, in which variable with high positive values, with high negative values, and with values close to zero can be found.

For example, (+)-catechin and unassigned compounds SC3 (in the window of syringic and caffeic acids) and p-C5 (in the window of p-coumaric acid), having high positive contribution, are influent in giving positive scores values to Salamino samples, on the other hand, variables such as SC4 and Quer5 are important, having high negative contribution, to the disposition in the fourth quarter of Grasparossa samples.

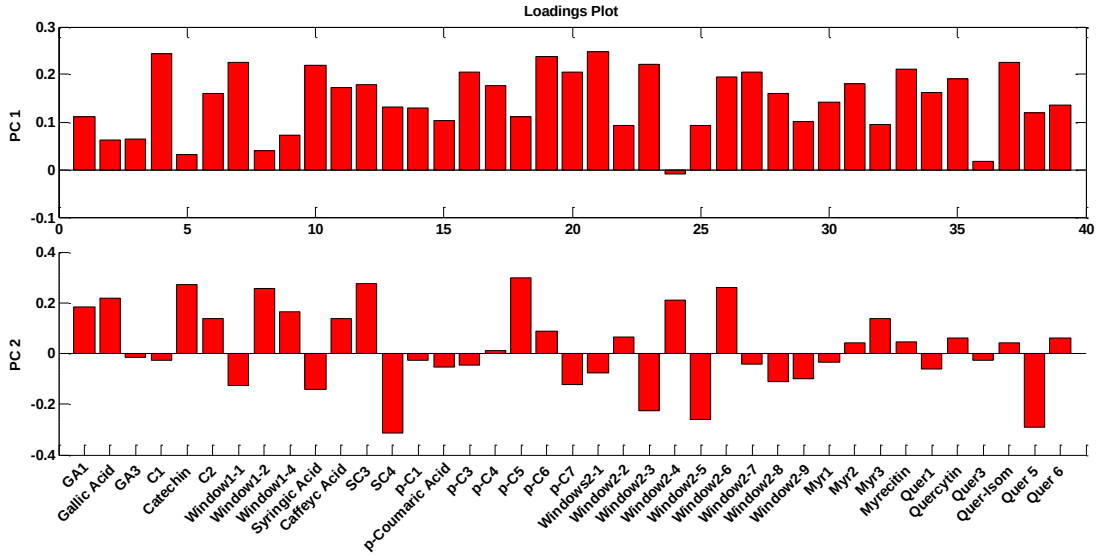

Figure 4.3.3.6 *Loadings plot of the PCA model built using "HPLC-DAD dataset"*

The concentration profiles of samples belonging to the training set, resolved from the MCR-ALS models, were also used to build a PLS-DA classification model. The test set samples class membership was predicted using the model in order to evaluate the classification performances. The results reported in Table 4.3.3.2 demonstrate the ability of phenolic compounds in discriminating wines on the basis of their varietal composition [21-22].

^{1}H-NMR data analysis and modelling

Prior to data analysis, several preprocessings were applied on ^{1}H-NMR signals in order to remove the not informative part of the spectra, the horizontal and baseline shifts and

to rescale the regions of the spectra characterized by good variability but low signals [16,17].

Since hundreds or even thousands of molecules can be detected by ^{1}H-NMR on complex matrices, and giving, each single compound, origin to an high number of peaks at different chemical shifts values, with different shapes (singlet, doublet, multiplet) and intensities, the interpretation of the hidden information veiled in the complexity of the signals of this kind of data is a difficult task to face.

In order to recover the majority of the information present in the original data, a PCA model, based on four components (chosen accordingly to the minimum error in cross validation) and able to describe the 86% of the total variance was computed on the training set samples.

As reported in Figure 4.3.3.7 it appears that several parts of the spectra present high loadings values for all the components, as in the case of the doublet at 1.4 ppm, attributable to CH_3 group of lactic acid, which is positively associated to the first, second and fourth component and negatively to the third.

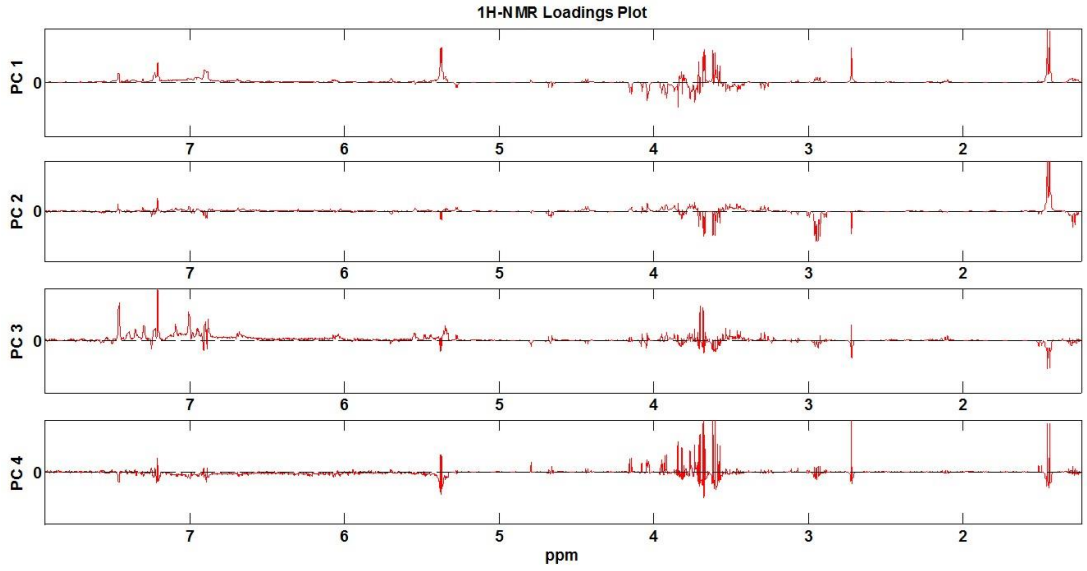

Figure 4.3.3.7 *Loadings plot of the PCA model built using ^{1}H-NMR dataset*

Other signals result overlapped and the association to a single compound is not possible. In fact, the anomeric region which spans the chemical shift range from three to five ppm comprises a numerous group of signals, as well as the aromatic one which covers the seven ppm region. In these cases, an overall contribution, related to a class of compound, could be attributed to a given component, as in the case of aromatic region, in which polyphenol signals are intensively present for the third component or the case of sugars signals present in anomeric region and positively correlated with the fourth component.

From the examination of the scores plot reported in Figure 4.3.3.8 it emerges that the separation of the samples belonging to the three varieties of wines is not as pronounced as in the same plot performed on HPLC-DAD data.

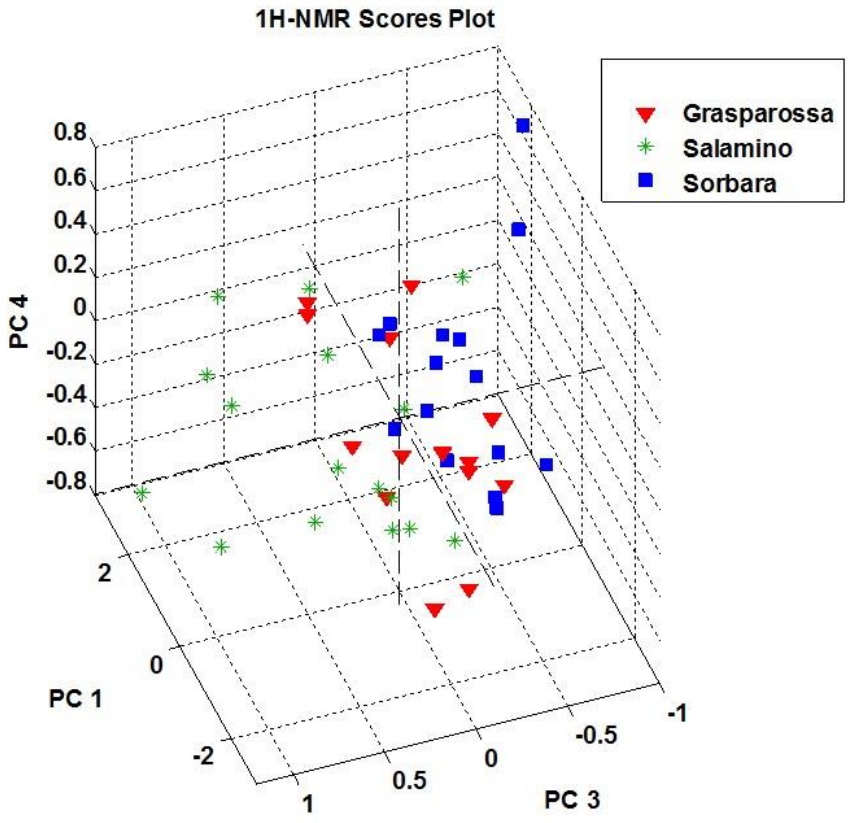

Figure 4.3.3.7 *Scores plot of first, third and fourth component extracted by means of PCA from the ^{1}H-NMR dataset*

The third component, mainly related to the aromatic part of the signal, seems to be able to distinguish, but not completely, at positive values Salamino and Grasparossa samples from Sorbara samples present at negative values. Worst results were obtained also for the PLS-DA classification model reported in Table 4.3.3.2.

Table 4.3.3.2 *Performances of the classification models obtained from the analysis of separate datasets*

Dataset	**Class**	***HPLC-DAD***	***1H-NMR***	***EEM***
Model		*PLS-DA*	*PLS-DA*	*NPLS-DA*
Latent variables		3	4	9
N° of misclassified samples in CV	***Grasparossa***	0	3	3
	Salamino	2	5	3
	Sorbara	0	2	1
% Correct classification rate in CV	***Grasparossa***	100	78	79
	Salamino	88	69	81
	Sorbara	100	8	93
N° of missclassified Test set samples	***Grasparossa***	0	0	1
	Salamino	0	1	0
	Sorbara	1	1	1
% Correct classification rate Test set	***Grasparossa***	100	100	80
	Salamino	100	75	100
	Sorbara	75	75	75
N° of missclassified Training set samples	***Grasparossa***	0	1	1
	Salamino	1	3	0
	Sorbara	0	0	0
% Correct classification rate Training set	***Grasparossa***	100	93	93
	Salamino	94	81	100
	Sorbara	100	100	100

Data analysis and modelling of the "fused" dataset

In order to perform the joint analysis of all the information related to the three analytical techniques used to characterize the wine samples and accordingly with the mid-level data fusion framework reported in Figure 4.3.3.1, a concatenation of the variables extracted from the separate data analysis models (concentrations, scores and first mode loadings) was performed. As described in Chapter 3, in mid-level data fusion methodologies, the way in which the weighing and scaling procedures are applied in order to produce the "fused" data table is one of the most crucial problem to face.

In this case, the variables involved in the concatenation are thirty-nine concentration profiles, (each concentration is referred to a single compound), four first mode loadings and four scores which are referred to groups of compounds. The variables extracted in the lower level of modelling are able to explain almost all the variance of the systems, hence, the group scaling could be a suitable way to proceed. Notwithstanding that, it was decided to follow a "trial and error" strategy in order to individuate the best scaling procedure aimed to produce the most suited "fused" dataset, evaluating several possibilities of weighing.
In particular, block scaling, block autoscaling (prior to block scaling each variable was autoscaled) and autoscaling were performed. Even if few minor differences were noticed in the PCA-based explorative analysis, with respect to other scaling procedures, the classification model obtained on fused dataset using autoscale gave better results, hence, it was decided to autoscale each variable from each block prior to merge them in the "fused" dataset, formed by forty-seven variables.

Indirectly, the adoption of autoscaling as weighing methodology gives a higher importance to the more numerous block of variables (the thirty-nine concentration from HPLC-DAD analysis). The thirty-nine concentration profiles extracted from HPLC-DAD data are the only directly related to single compounds, for this reason it was decided to accept the fact that the global weight of these variables was higher if compared to the others extracted from NMR and EEM datasets. A PCA model, based on three components (chosen according to minimum cross validation error) and able to

explain 51% of the total variance of the "fused" dataset was built for explorative purpose. In Figure 4.3.3.8 the scores plot for the first two components highlights analogies with the one obtained from the HPLC-DAD dataset (Figure 4.3.3.5).

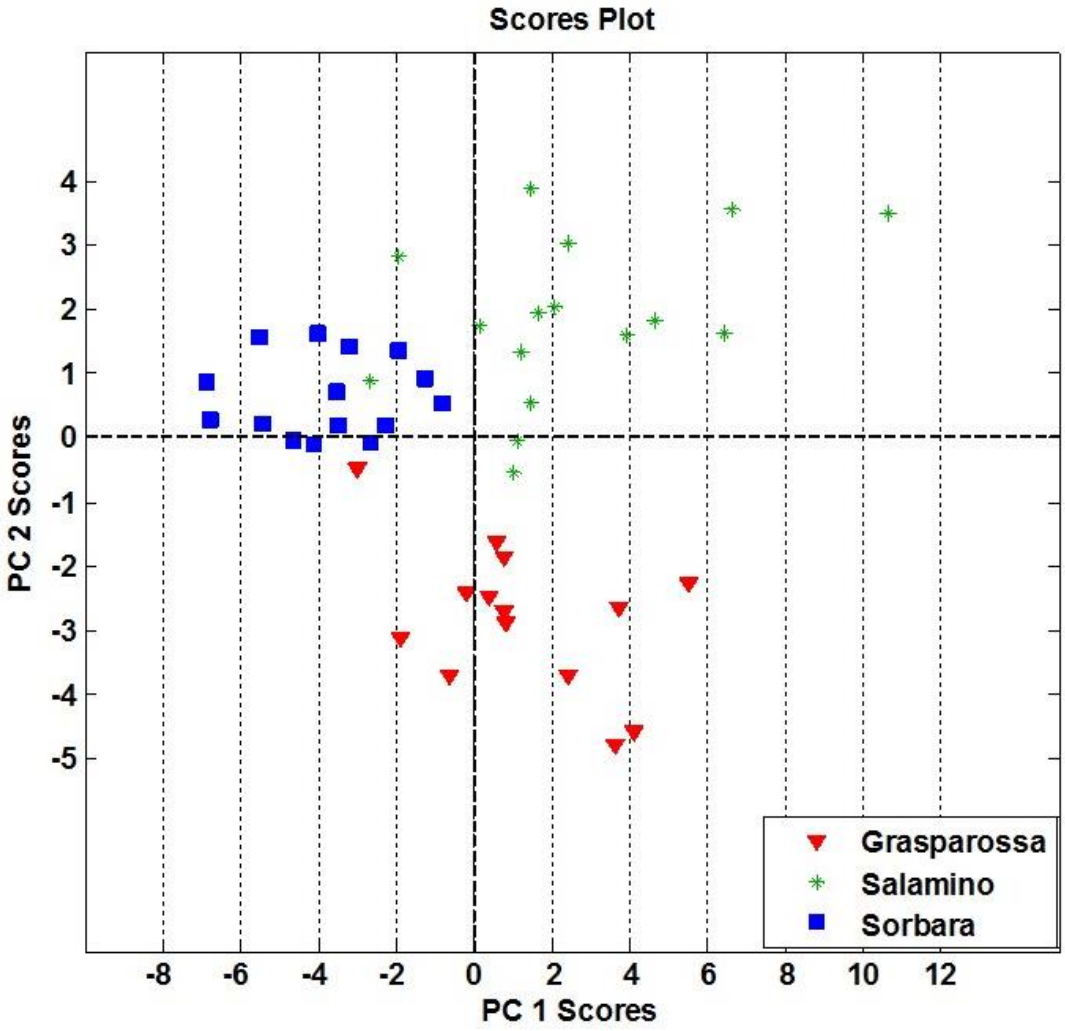

Figure 4.3.3.8 *Scores plot of first two components extracted by means of PCA from the "fused" dataset*

In particular, all Grasparossa samples are placed at negative values for the second component, almost all Salamino samples at positive values for the first one whilst all Sorbara samples at negative values for the first component. Some improvements can be noticed in the separation of samples belonging to different classes, as matter of fact, no

one of the Sorbara samples is confused within the Salamino cluster in the first quarter.

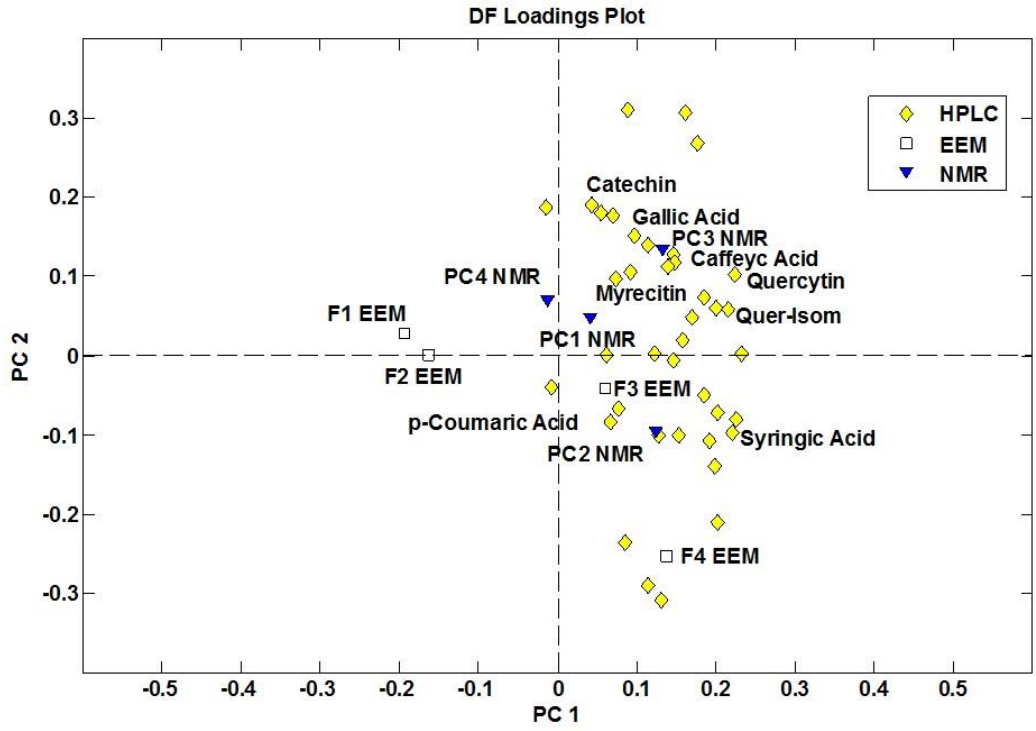

Figure 4.3.3.9 *Loadings plot of first two components extracted by means of PCA from the "fused" dataset*

Figure 4.3.3.9, in which the first two components loadings plot is reported, highlights that all variables give a contribution to the position in the scores space of the samples. The first and second factors from the EEM dataset appear to be most of all responsible of the improvement in the separation of Salamino and Sorbara samples, having high negative values for second component, while the fourth EEM factor seems characteristic for Grasparossa class. Almost all the variables arising from the HPLC-DAD dataset are located at positive values for the first component. The third component extracted from the NMR dataset show positive loadings in both PCs, contributing to separation of Salamino samples from Sorbara ones while the second component seems characteristic of Grasparossa samples.

A PLS-DA classification model (five latent variables chosen accordingly to minimum cross validation classification errors) was built using the "fused" dataset. The results of

the classification model built using the "fused" dataset via mid-level data fusion, reported in Table 4.3.3.3, highlight the improvement with respect to the models obtained starting from the separate datasets analysis.

Table 4.3.3.3 *Performances of the PLS-DA classification model obtained from the analysis of the "fused" dasets*

Dataset	**Class**	***Fused Dataset***
Model		*PLS-DA*
Latent variables		5
N° of misclassified samples in CV	***Grasparossa***	0
	Salamino	2
	Sorbara	0
% Correct classification rate in CV	***Grasparossa***	100
	Salamino	88
	Sorbara	100
N° of missclassified Test set samples	***Grasparossa***	0
	Salamino	0
	Sorbara	1
% Correct classification rate Test set	***Grasparossa***	100
	Salamino	100
	Sorbara	75
N° of missclassified Training set samples	***Grasparossa***	0
	Salamino	0
	Sorbara	0
% Correct classification rate Training set	***Grasparossa***	100
	Salamino	100
	Sorbara	100

Figure 4.3.3.10 shows the scores plot of the first three latent variables of the PLS-DA model, and highlights the good separation among the distinct typologies of Lambrusco wines. In particular, all Salamino samples present negative values for the first and second latent variable, Sorbara samples are placed mainly at positive values for the third latent variable and close to zero for the others, whilst Grasparossa samples are situated at high positive values for the second latent variable.

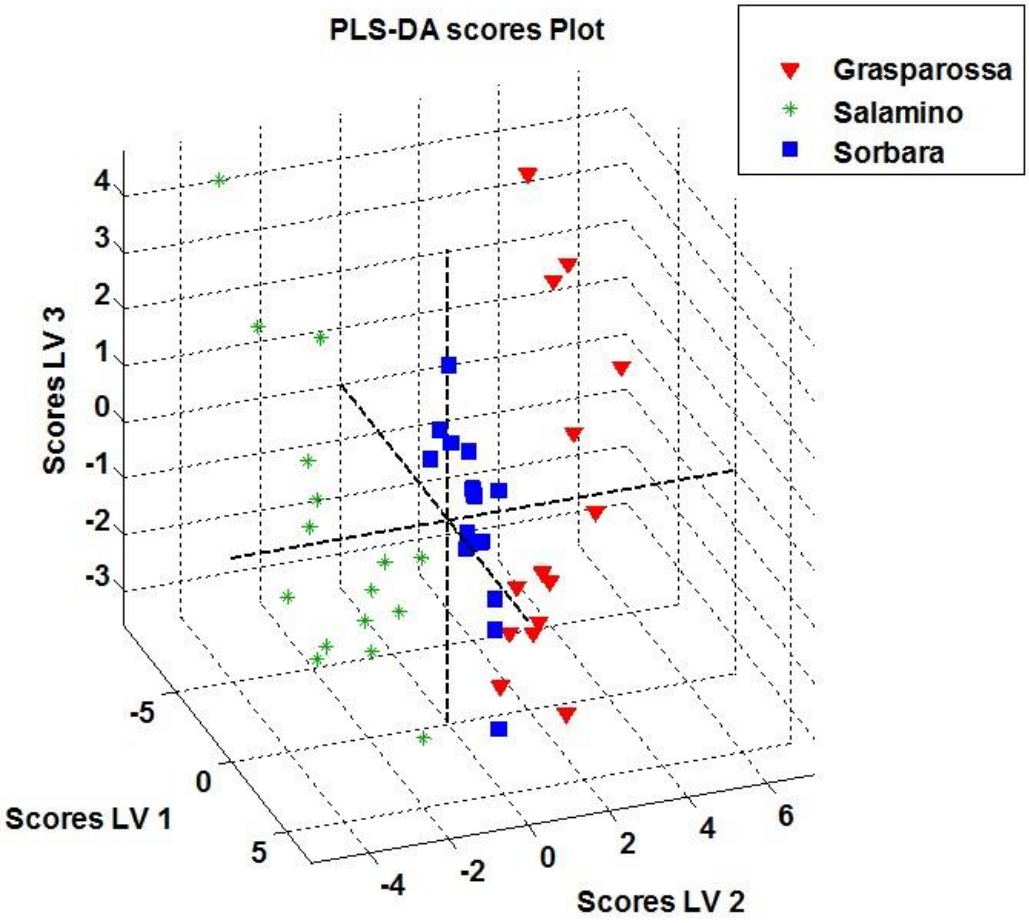

Figure 4.3.3.10 *Scores plot of the PLS-DA model built using "Fused dataset"*

From the inspection of the weights plot reported in Figure 4.3.3.11 it is possible to individuate which are the variables which most of all are responsible of the separation of the samples in the scores space, and the correlations among them.

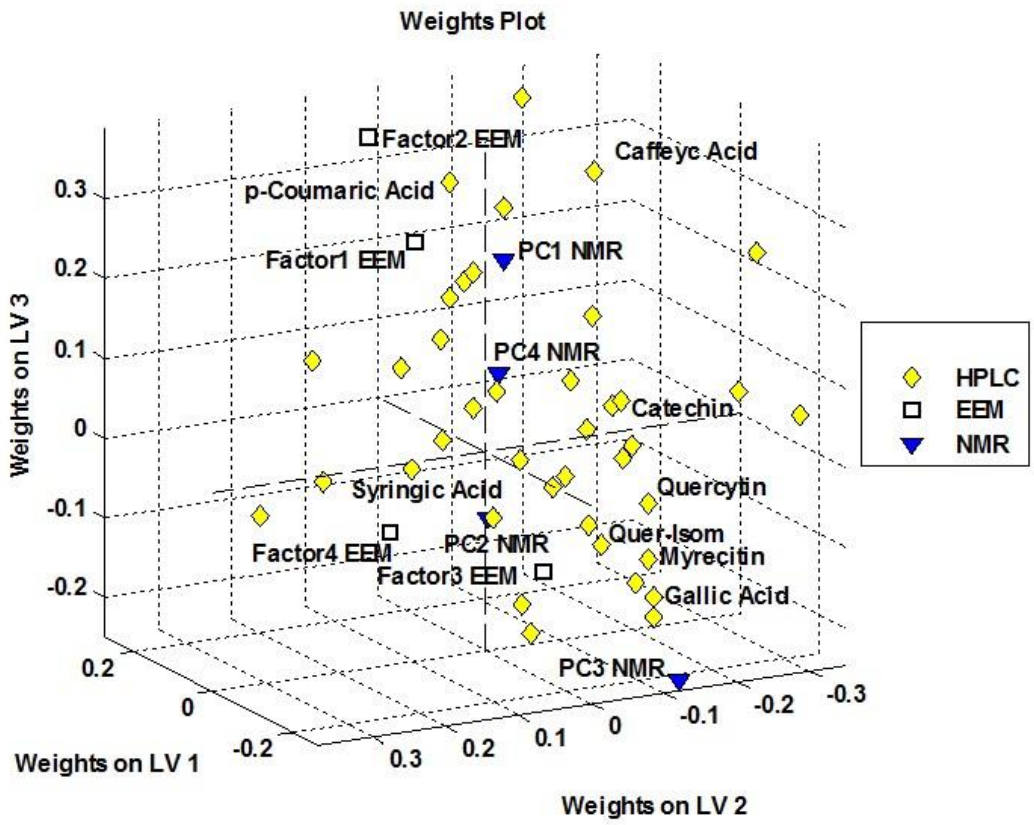

Figure 4.3.3.11 *Weights plot of the PLS-DA model built using "Fused dataset"*

In particular, the first and second factors from EEM dataset, Caffeyc acid, p-Coumaric acid and two unassigned molecules from the Window2 cluster and myrecitin elution window present high weights values for the third latent variables, the one mainly responsible of the separation of Sorbara samples. The second EEM component, as described above, presents fluorescent properties compatible with several molecules such as, trans-resveratrol, trans-piceid, gentisic acid and p-Coumaric acid, which is one of the resolved compound by MCR analysis of HPLC-DAD data that stays very close to the second EEM factor in the scores space obtained by the PLS-DA model.

The first factor resolved from EEM data is close to a group of unassigned compounds resolved from the analysis of HPLC-DAD data. Future investigation will be oriented to

the assessment of compounds compatible with the fluorescent properties described by the first EEM factor (vanillic acid).

In addition, the first component arising from the PCA analysis of ^{1}H-NMR data, mainly influenced by signals in the anomeric region, brings a relevant contribution for the discrimination of Sorbara samples.
The third component from the ^{1}H-NMR signals elaboration, mainly associated with the phenolic signal region, and a wide group of variables obtained via HPLC-DAD analysis, such as Gallic acid, myrecitin, quercytin etc., are relevant for the classification of Salamino samples, having high negative contributions for both first and second latent variables.
On the other hand, Grasparossa samples are described mostly by the fourth component arising from the EEM part, syringic acid and other molecules obtained from the analysis of HPLC-DAD data.

Since for each separate dataset a classification model based on PLS-DA, or NPLS-DA, depending on the order of the data, was built, a high-level classification can be evaluated. The response Y vectors of the classification models obtained in the lower level of modelling can be considered jointly in order to compute a super-classification model based on the predicted memberships of the samples. Three classes of samples were investigated for each dataset, hence, three matrixes containing the predicted values for each of the three class have to be arranged in order to produce the data table (the nine classification indexes are merged). In order to assess the class membership of the samples by means of the fusion of the classification indexes obtained in the lower level of modelling, several rules can be adopted as described in Chapter 3.4.

In all the classification models obtained from the separate datasets, the assignment of a sample to a given class was decided with a majority vote rule, or better, samples are assigned to the class for which the predicted y-value is higher. In order to not disregard the information related to the y-values corresponding to the not recognized classes, e.g.

considering responses values [-0.6, 0.1, 0.8] the sample is assigned to class three but y-values are also computed for the second and third class, it was decided to concatenate all the responses vectors in order to produce a data table of classification indexes used as a base for the development of the super-classification model.

The response vectors were merged together without any kind of normalization, since they are predicted in a range close to -1/+1. A PLS-DA model, built using the response vectors obtained from the forty-five samples of training set, based on two latent variable (chosen on the minimum cross validation error) was computed and the test set samples were predicted.

In Figure 4.3.3.12 the scores plot is reported, which highlights a good separation among the different classes of samples.

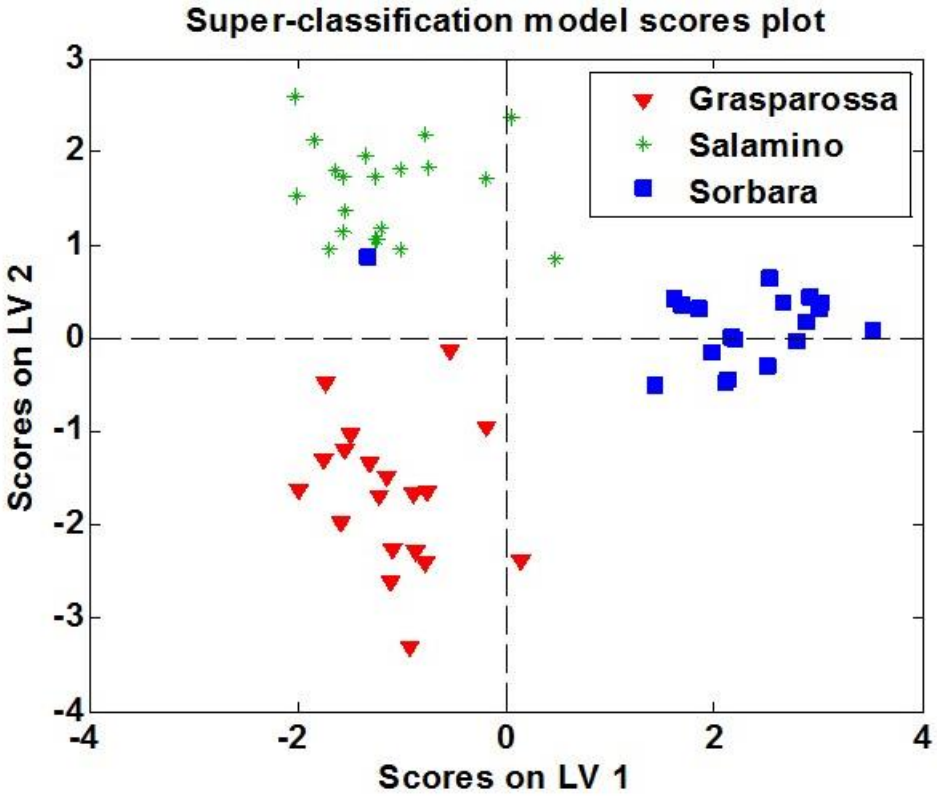

Figure 4.3.3.12 *Scores plot from the super-classification PLS-DA model*

In particular, all training set samples are correctly assigned and only one Sorbara sample from the test set is predicted as Salamino. Since the production regulation of Sorbara wines allows an ampelographic composition up to 40% of Salamino grapes, this particular samples could present an amount of the latter variety higher with respect to all others Sorbara samples. Moreover, the same error in classification was found in all other classification models. In general, the classification performances obtained with the use of the high-level data fusion approach are the same obtained from the mid-level one, reported in Table 4.3.3.3. In both cases, very good classification performances were achieved and no improvements were found by adopting an high-level data fusion approach. Concluding, the possibility to discriminate the varietal origin of Lambrusco wines was demonstrated with the adoption of multiple data fusion frameworks. These methodologies proved to be suitable to handle complex analytical data, characterized by having different orders (matrixes and tensors) and systematic variations unrelated to the responses to obtain. Good classification performances were achieved with the use of data fusion based classification models, obtaining at the same time important information about the correlations among variables arising from different sources.

4.4 Coupled matrix tensor factorization (CMTF) application in food characterization

4.4.1 Introduction

As described in Chapter 3.5, coupled data fusion methodologies are aimed to the joint analysis of different sources of information attempting to extract simultaneously shared components/features from each of them. In particular, the CMTF algorithm [23] and its derivation ACMTF-OPT [24] allow dealing with data of different orders (matrixes and tensors) in a way that both shared and unshared components can be extracted. In this chapter, a preliminary application of ACTM-OPT is proposed in order to point out its potentiality and fields of application. In order to confirm the hypothesis proposed in the previous section about the correlations among the resolved profiles from the EEM dataset and some compounds resolved by MCR-ALS on the HPLC-DAD dataset, the coupled tensor factorization was applied to two tensors: the EEM 3-way array and the p-Coumaric elution windows arranged as tensor (samples x elution time x wavelength), as schematized in Figure 4.4.1.

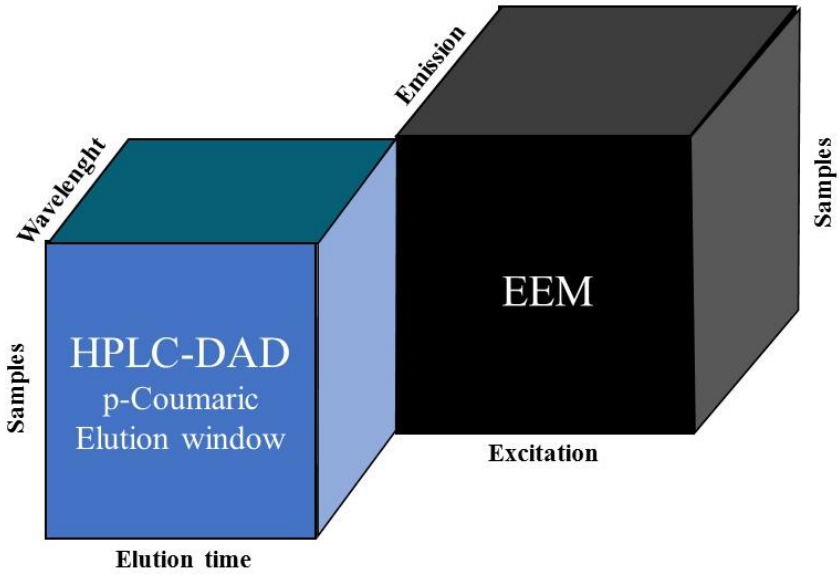

Figure 4.4.1.1 *Schematization of the data modelled by ACMTF-OPT. The two tensors share the samples direction*

4.4.2 Results and discussion

The EEM and HPLC-DAD tensors were jointly factorized by extracting the same factor matrix from the sample mode using a six components ACMT-OPT model. A schematization of the joint factorization of the couple of tensors is shown in Figure 4.4.2.1.

$$\min_{A,B,C,D,E,\Lambda_x,\Lambda_y} \|X - [\![\Lambda_x; A, B, C]\!]\|^2 + \left\|Y - [\![\Lambda_y; A, D, E]\!]\right\|^2 + \beta\|\Lambda_x\|_1 + \beta\left\|\Lambda_y\right\|_1$$

$$\text{s.t. } \|a_r\|_2 = \|b_r\|_2 = \|c_r\|_2 = \|d_r\|_2 = \|e_r\|_2 = 1, \quad \text{for } r = 1 \ldots R$$

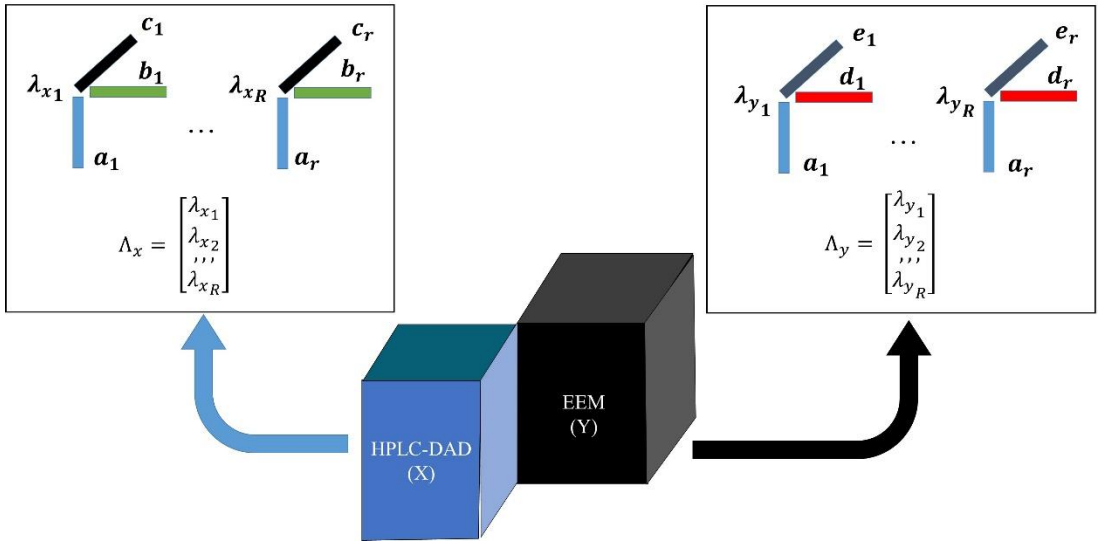

Figure 4.4.2.1 *Schematization of the joint factorization of the couple of tensors: X (HPLC-DAD p-Coumaric window) and Y (EEM dataset)*

The joint factorization of the two tensors can be interpreted in a similar way to how a PARAFAC-based decomposition is discussed. In this case, the decomposition extracts a first mode-loadings matrix referred to the shared mode (in this case the sample direction) which is weighted for the corresponding λ values for each factor and tensor. The weights indicate what is common between the two blocks or what is not. If for examples, the λ values obtained for the two blocks were λx=[1 1 1 0 1] and λy=[1 1 1 1 0] , the first three components are shared among the two blocks, the fourth is only attributable to the Y block and the fifth to the X block. Hence, the loadings profiles of

the second and third mode for the first three factors of the two blocks should contain the information related to the same compounds (groups of compounds), which share the same first mode (related to the amount/concentration) factors. When dealing with complex data, as the case here described, the quantitative resolution of the investigated system is difficult to achieve and values of weights not close to one can be found [23-24]. In general, only when the weights are equal to zero a factor can be considered not common. Hence, in order to understand if a factor is shared or not, the interpretation of the loadings profiles related to that factor, in addition to the weight values, have to be considered. The resolution by means of MCR of the p-Coumaric elution window [18] was performed using seven components imposing constraints in order to achieve a quantitative resolution, whilst the one on the EEM array, described in the previous sections, by means of a four factors PARAFAC model. Several ACMTF models were tried with different number of components, but the best results, were obtained using a six components model, which indicates that some factors could contain information related to more than a compounds.

The λ weight coefficients resolved from the six factor model highlight that three out of six components are in common among the HPLC-DAD and EEM array whilst the last three are only referred to the HPLC-DAD dataset.

$$\boldsymbol{\lambda}_{\mathbf{EEM}} = [0.920 \;\; 0.325 \;\; 0.098 \;\; 0.000 \;\; 0.000 \;\; 0.000]$$

$$\boldsymbol{\lambda}_{\mathbf{HPLC}} = [\, 0.123 \;\; 0.191 \;\; 0.799 \;\; 0.282 \;\; 0.265 \;\; 0.0967]$$

In the currently available implementation of ACMTF-OPT it is not possible to impose constraints (such as nonnegativity or unimodality); in addition, the best resolved model is rank deficient with respect to the one resolved by means of MCR for the HPLC-DAD part. Although the obtained model cannot represent a perfect resolution for all the contributions of the EEM and HPLC-DAD parts, it is still possible and interesting to interpret the results. It is important to remember that the discussion should not be carried

out in terms of "pure" profiles, which cannot be obtained (for example UV spectra could have negative values), but by considering an overall interpretation of peculiar features (for example at which wavelength a profile presents a local maximum or minimum). This could help to unveil the hidden correlation among the different sources of information. Moreover, since the factorization is obtained simultaneously for the two blocks, the information related to only one block could help to solve the other one, giving results that cannot be obtained by the separate analysis of the sources of information.

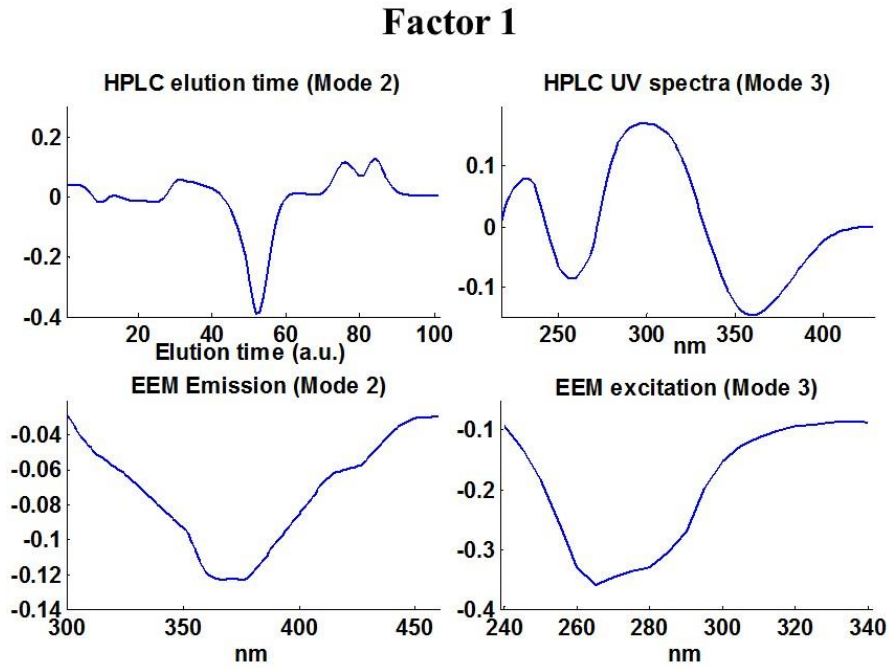

Figure 4.4.2.2 *Loadings profiles for the first common factor resolved by means of the joint factorization of EEM and HPLC-DAD (p-Coumaric elution window) tensors*

In Figure 4.4.2.2 the loadings profiles of mode two and three of the first common factor (first mode loadings are the same for EEM and HPLC-DAD 3- way arrays) are reported.

Even if a quantitative resolution of the system was not achieved, from the second mode loadings profile obtained on the HPLC-DAD tensor it is possible to individuate a peak (in the central region of the plot), which seems to dominate the elution pattern. The corresponding absorption spectra (HPLC-DAD Mode 3 loadings) is compatible with

the one of vanillic acid, the same for the excitation (EEM Mode 3 loadings) and emission (EEM Mode 2 loadings) profiles resolved from the EEM dataset [20]. In all figures related to Factor 1 it is important to remember that sign ambiguity is not resolved, thus having profiles which can be compared to "real" ones by considering their sign changed. The second common factor, on the other hand, presents very complex structures for all the resolved profiles, as shown in Figure 4.4.2.3. Due to the rank deficiency of the model, the contributions of many compounds are present, but in general, the spectral minima and maxima (from mode 3 of HPLC-DAD and mode 2 and 3 of the EEM) are compatible with the ones of p-Coumaric acid [20]. The second mode loadings from the HPLC-DAD highlights the complexity of the investigated system, in particular the final part of the elution domain presents a negative and a positive peak.

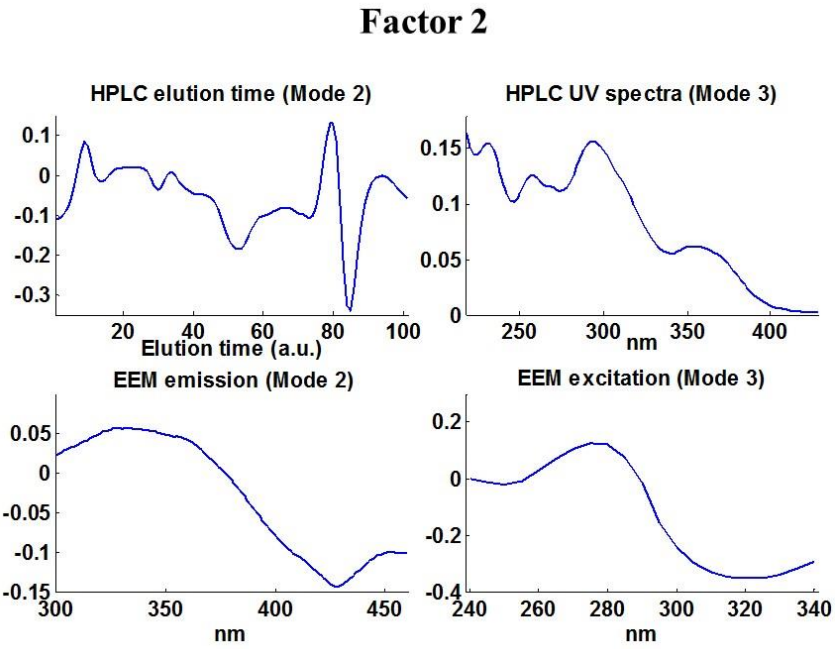

Figure 4.4.2.3 *Loadings profiles for the second common factor resolved by means of the joint factorization of EEM and HPLC-DAD (p-Coumaric elution window) tensors*

The same elution domain resolved from the separate analysis of the p-Coumaric elution window by means of MCR-ALS [18] demonstrated the presence of more compounds, which in the ACMTF-OPT approach was not possible to quantitatively resolve.

When considering the factors not in common, related only to the resolution of the HPLC-DAD tensor, promising results were obtained. In particular, the fifth factor points out the presence of a compound (also in this case a non-quantitative resolution was achieved), which is eluted in the first region of the elution window as, reported in Figure 4.4.2.4. Further investigation of the spectroscopic properties (the third mode loading profile) should be performed in order to assess the chemical structure of the unknown compound.

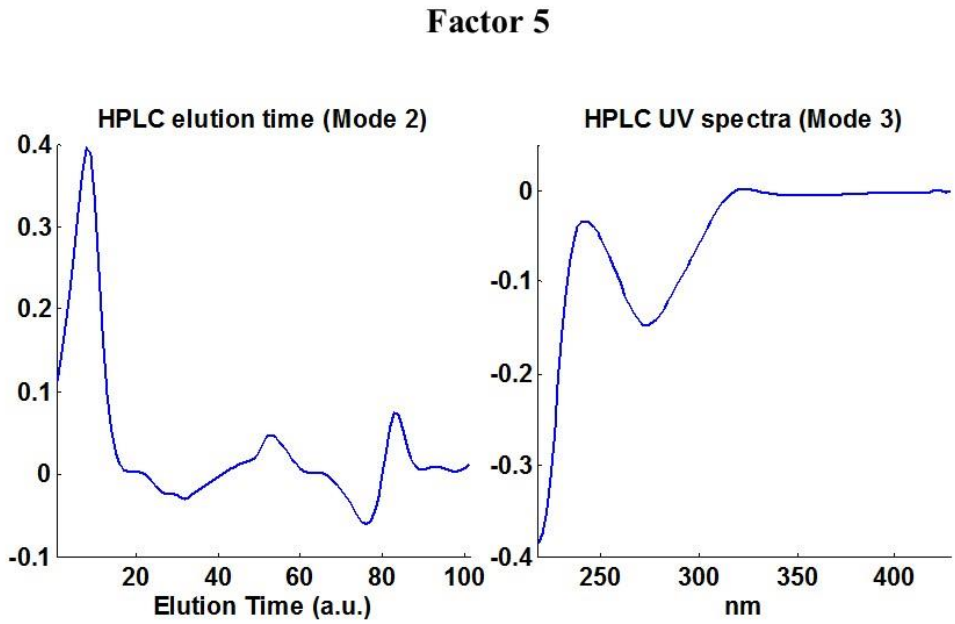

Figure 4.4.2.4 *Loadings profiles for the fifth factor (unshared) resolved by means of the joint factorization of EEM and HPLC-DAD (p-Coumaric elution window) tensors*

As a conclusion, this preliminary study allows strengthening the hypothesis made about the mid-level data fusion approach described in the previous section. Moreover, further information were obtained, and the resolution, even if partial, of a compound presents in the first part of the elution window of the p-Coumaric acid was obtained.

Further work is being carried out; in particular, attention is given to the use of a version of the algorithm, still in a phase of development, which allows imposing some

constraints: this could represent a benefit in the interpretation of the common and unshared component profiles.

4.5 References

[1] Council Regulation (EC) No 813/2000 of 17/04/2000

[2] Gazzetta Ufficiale of the Italian Republic, No 124 of 30/052000

[3] Council Regulation (EC) No 583/2009 of 03/07/2009

[4] Gazzetta Ufficiale of the Italian Repubblic, No 235 of 09/10/2009

[5] M.B. Wise, J.M. Shaver, N.B. Gallagher, W. Windig, R. Bro, R.S. Koch, "*Chemometrics Tutorial for PLS_Toolbox and Solo*", Eigenvector Research Inc., Wenatchee, USA, 2006

[6] H. F.M. Boelens, R.J. Dijkstra, P.H.C. Eilers, F. Fitzpatrick, J. A. Westerhuis, "*New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection",* Journal of Chromatography A, 2004, 1057(1), 21-30.

[7] S. Wold, "*Pattern recognition by means of disjoint principal component models.*",Pattern Recognition, 1976, 8, 127-139.

[8] AGER, Agroalimentare e Ricerca. *"New analytical methodologies for varietal and geographical traceability of oenological products"*, contract n. 2011 – 0285

[9] DM 27/07/2009, Gazzetta Ufficiale of the Italian Republic no. 184-187-188, 13/08/2009

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

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

[12] L. Bertacchini, C. Durante, A. Marchetti, S. Sighinolfi, M. Silvestri, M. Cocchi, "*Use of X-ray diffraction technique and chemometrics to aid soil sampling strategies in traceability studies",* Talanta, 2012, 98, 178–184

[13] 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,

[14] W. Windig, "*Spectral data files for self-modeling curve resolution with examples using the Simplisma approach*", Chemometrics and Intelligent Laboratory Systems, 1997, 36(1), 3-16

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

[16] M. Silvestri, A. Elia, G. Papotti, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, "*A Mid-Level Data Fusion strategy for the varietal classification of Lambrusco P.D.O. Wines",* Chemometrics and Intelligent Laboratory Systems, Submitted Dec 2013.

[17] G. Papotti, D. Bertelli, R. Graziosi, M. Silvestri, L. Bertacchini, C. Durante, M. Plessi, "*Application of one-and two-dimensional NMR spectroscopy for the characterization of Protected Designation of Origin Lambrusco wines of Modena",* Journal of agricultural and food chemistry, 2012, 61 (8), 1741-1746

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

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

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

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

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

[23] E. Acar, T. G. Kolda, and D. M. Dunlavy, "*All-at-once Optimization for Coupled Matrix and Tensor Factorizations*", KDD Workshop on Mining and Learning with Graphs, 2011

[24] E. Acar, A. J. Lawaetz, M. A. Rasmussen, and R. Bro, "*Structure-Revealing Data Fusion Model with Applications in Metabolomics*", Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC'13), 2013, 6023-6026

CHAPTER 5

Final Remarks

The three-year activities reported in this Thesis dealt with data fusion techniques, their potentialities, *pros* and *cons*, and some applications in the context of food analysis.

When dealing with the analysis of complex matrices, as foodstuffs are, characterized by having strong interferences, additional uncontrolled variability due to biological, chemical and physical transformations and redundancy, the reductionistic approach, which implies the evaluation of the results obtained by means of a single analytical technique, could not be sufficient to describe the investigated system.

The holistic approach, followed by the adoption of a data fusion framework, could result mandatory in all that cases in which the reductionistic one cannot bring the expected results. The joint evaluation of different sources of information, each of them able to point out some features related to the investigated samples, could help to unveil the hidden information able to characterize the samples, moreover, offers the possibility to highlight the correlation among the variables arising from the different sources.

Depending on the typology of data to be used for the formulation of the data fusion framework, several kind of data fusion strategies could be chosen: low level data fusion (or data level fusion), mid level data fusion (or feature level data fusion), high level data fusion (decision level data fusion) or coupled data fusion.

Each data fusion strategy was discussed in the early chapters of the thesis by a theoretical point of view, highlighting potentialities and limitations. In the last chapter some applications are proposed in order to give to the reader the basis to face practically the data fusion issues.

In particular by means of a low-level data fusion strategy, a classification model aimed to distinguish different classes of Aceto Balsamico Tradizionale di Modena, coupling spectroscopic signals (mid infrared MIR and near infrared NIR) was built. The choice of the best weighting procedure for the concatenation of the two blocks of variables was illustrated and good classification performances were obtained. Infrared spectroscopies

proved to be suitable techniques for the characterization of these kind of oenological products when used in synergy, by means of a data fusion strategy.

A mid-level data fusion framework was proposed within a geographical traceability study aimed to preliminarily characterize soils sample of the District of Modena by means of primary and secondary geographical traceability indicators (elemental composition, isotopic abundance ratios and X-ray diffractograms). A lower level modelling was conducted by means of MCR on the separate blocks of variables in order to extract few informative and chemical meaningful variables which were concatenated to produce a fused data set, subjected in the higher level modelling to multivariate data analysis. Results able to well characterize the investigated samples and to confirm the correlations among variables arising from different block were obtained. The comprehension of the variability of primary indicators with respect to the position and altitude of soils sample within the Modena District has helped the formulation of the extensive sampling.

The varietal characterization of Lambrusco P.D.O. wines was faced with the development of different data fusion frameworks. A mid and high level data fusion strategy was built in order to jointly analyse several sources of information (NMR, HPLC-DAD and EEM) with the purpose to distinguish, by means of a super-classification model obtained on the fused features arising from the lower level of modelling, the three typologies of Lambrusco wines. A preliminary study conducted by means of a coupled data fusion strategy, *via* ACMTF-OPT, was used to confirm the hypothesis proposed as results of the mid-level data fusion.

With the hope that the potentialities of data fusion techniques, described both from a theoretical and practical point of view, were clearly described and readable in this Thesis, I would like to encourage the reader to consider this holistic approach when the results obtained by means of separate data analysis appear meaningless if not contradictory. Or, as frequently happens, at least to me, when the data seem to be kidding you instead of speaking to you.

APPENDICES

APPENDIX I

Analytical Apparatuses

Infrared Spectroscopies

The acquisition of Near-Infrared signals was performed using a FTIR spectrophotometer Bruker Optics Vector 22N coupled with a specific probe for liquid samples (Hellma) connected with the use of optic fiber.

The acquisition of Mid-Infrared signals was performed using a FTIR spectrophotometer Bruker Optics, Vertex70 equipped with an single reflection diamond ATR measure cell Golden Gate (Specac Limited)

Table 1 *NIR and MIR instrumental parameters*

	FT-MIR coupled with ATR	**FT-NIR with optic fiber probe**
Spectral Range	600-4000 cm^{-1}	4150-10000 cm^{-1}
Number of Scans	32	32
Total Number of Points	1763	3035
Acquisition Mode	Absorbance	Absorbance
Resolution	4cm^{-1}	4 cm^{-1}
Optical Path	ATR	1mm

Determination of metals concentrations and $^{87}Sr/^{86}Sr$ isotope ratio

The determinations of the concentrations of metal ions in soils was perfomed using Flame Atomic Absorbtion Spectroscopy (FAAS) by means of a spectrometer Varian SpectrAA 220FS coupled with SIPS unit, and by means of Inductively Coupled Plasma Mass Spectrometry (ICP/MS) using a ThermoFisher Scientific XSeriesII mass-spectrometer. In the next tables the instrumental parameters used are listed for each analyzed element

Table 1 *FAAS instrumental parameters*

Element	Linearity Interval (ppm)	Flame	Wavelenght(nm)	Lamp Current (mA)	Ionization suppressor	Width (nm)
Na	0-1	Air-acetylene	589.0	10	CsCl	0.5
K	0-2	Air-acetylene	766.5	10	CsCl	1
Ca		N_2O-acetylene	422.7	10	CsCl	0.5
Mg		N_2O-acetylene	285.2	4	CsCl	0.5

The simultaneous determination of ^{51}V, ^{52}Cr, ^{59}Co, ^{60}Ni, ^{63}Cu, ^{66}Zn, ^{71}Ga, ^{75}As, ^{85}Rb, ^{88}Sr, ^{114}Cd, ^{133}Cs, ^{137}Ba, ^{139}La, ^{140}Ce, ^{141}Pr, ^{146}Nd, ^{149}Sm, ^{151}Eu, ^{158}Gd, ^{163}Dy, ^{165}Ho, ^{167}Er, ^{169}Tm, $^{17}2Yb$, ^{175}Lu, ^{205}Tl, ^{208}Pb, ^{232}Th, ^{238}U was performed using the XSeriesII modality CCT-KED with He/H_2 flow set to 5 mL min^{-1}.

Table 2 *ICP-MS instrumental parameters*

Modality	He (mL/min)	Delay (sec)	Extraction (V)	L1 (V)	L2 (V)	Focus (V)	D1 (V)	D2 (V)	Pole Bias (V)
CCT-KED 5.0	5	0	-114	-1160	-80	-10	-43.8	-138	-15

Modality	Hex. (V)	Neb (L/min)	L3 (V)	Forward (W)	Hor	Vert	DA (v)	Cool (L/min)	Aux. (L/min)
CCT-KED 5.0	-20	0.9	-195.3	1400	106	281	-29.8	13	0.61

Strontium isotope ratio measurements were performed by means of an MC-HR-ICP/MS spectrometer (Neptune, ThermoFinnigan, Bremen, Germany). Data acquisition, in low resolution mode, was simultaneous for all the measured ion masses, m/z: 82 (L4), 83 (L3), 84 (L2), 85 (L1) 86 (C), 87 (H1), 88 (H2). The instrumental parameters are summarized in Table 3.

Table 3 *MC-HR-ICP-MS instrumental parameters*

Parameter	**Value/Type**
Rf Power	1245 W
Auxiliary gas flow rate	1.20 L min^{-1}
Cooling gas flow rate	16 L min-^{1}
Sample/Skimmer cone	Ni
Spray chamber	Cyclonic + Scott type
Nebulizer	PFA micro Flow Self aspirating
Number of block	1
Number of cycles	100
Integration time	8.839 s
Measure time for each sample	14 min
Uptake time	300 s
Wash time	100 s
Idle time	10 s
Mass analyzer pressure	< 10^{-8} mbar
Background/baseline determination	HNO_3 (4% w/w)
Control cup for peak centering	C

X-Ray Diffraction of Powders

X-ray diffraction of powder (XRDP) was carried out on soil samples by a θ/θ PANalytical X'Pert powder diffractometer equipped with a Real Time Multiple Strip (RTMS) detector (PANalytical X'Celerator). A 0.51 divergence slit and a 0.51 anti-scatter slit as well as a soller slit (0.04 rad) and a 10mm mask were mounted along the incident beam pathway. The diffracted beam pathway included a Ni filter, a soller slit

(0.04 rad) and an anti-scatter slit (5 mm). The XRDP data were collected from 5 to 120 2θ with steps of 0.01671 2θ; the counting time was of 1.905s per step.

The samples were loaded on aluminum sample holders by using a side-loading technique. A reference silicon tablet was measured at the beginning of each measurement session, in order to test the analytical settings and to monitor the instrumental drift.

High Performance Liquid Chromatography

Separation of the phenolic compounds by means of solid phase extraction, Supelco DSC-18 cartridges with 6 mL tubes, was performed on wine samples prior to the acquisition of the chromatograms. Samples were analyzed by reversed phase liquid chromatography by a Beckman System Gold coupled with a diode array detector. The column used was a reversed-phase Atlantis dC18 Waters-Milford-MA. The mobile phase used was formed by two solvents: solvent A: water (0.1% TFA); solvent B: 80 % acetonitrile and 20% water (0.1% TFA). An elution linear gradient was used. The wavelength range in the diode array detector was from 220 to 430 nm with a resolution of $\Delta\lambda$=2 nm.

^{1}H Nuclear Magnetic Resonance

^{1}H-NMR spectra were acquired with a Bruker FT-NMR Avance 400 spectrometer (Bruker Biospin GmbH Rheinstetten, Karlsruhe, Germany) operating at 400.13 MHz All of the experiments were performed at 300 K and nonspinning. ^{1}H NMR data were acquired using the Bruker spin−echo sequence "cpmgpr.fb" (Carr−Purcell−Meiboom−Gill, Bruker Library) with water presaturation, applied to suppress broad resonance signals

Emission Excitation Matrix Fluorescence

The acquisition of EEM signals was performed on degased wine samples using a FLS920 fluorescence spectrometer (Edinburgh Photonics) equipped with a variable-

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

APPENDIX II

List of Publications

PAPER I

L. Bertacchini, C. Durante, A. Marchetti, S. Sighinolfi, M. Silvestri, M. Cocchi

Use of X-ray diffraction technique and chemometrics to aid soil sampling strategies in traceability studies

Talanta, 2012, 98, 178–184

PAPER II

G. Papotti, D. Bertelli, R. Graziosi, M. Silvestri, L. Bertacchini, C. Durante, M. Plessi

Application of one-and two-dimensional NMR spectroscopy for the characterization of Protected Designation of Origin Lambrusco wines of Modena

Journal of agricultural and food chemistry, 2012, 61 (8), 1741-1746

PAPER III

M. Silvestri, L. Bertacchini, C. Durante, A. Marchetti, E. Salvatore, M. Cocchi

Application of data fusion techniques to direct geographical traceability indicators

Analytica Chimica Acta, 2013, 769, ***Cover Page Article****, 1-9*

PAPER IV

C. Durante, C. Baschieri, L. Bertacchini, M. Cocchi, S. Sighinolfi, M. Silvestri, A. Marchetti

Geographical traceability based on 87Sr/86Sr indicator: a first approach for PDO Lambrusco wines from Modena

Food Chemistry, 2013, 141(3), 2779–2787

PAPER V

M. Silvestri, A. Elia, G. Papotti, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi

A Mid-Level Data Fusion strategy for the varietal classification of Lambrusco P.D.O. Wines

Chemometrics and Intelligent Laboratory Systems, 2013, In Press

OTHER PUBLICATION

L. Bertacchini, M. Cocchi, M. Li Vigni, A. Marchetti, E. Salvatore, S. Sighinolfi, M. Silvestri, C. Durante

The Impact of Chemometrics on Food Traceability

Chapter 10, Data Handling in Science and Technology,2013, 28, Chemometrics in Food Chemistry, Edited by Federico Marini