Università degli Studi di Milano
Facoltà di Agraria
Dipartimento di Scienze e Tecnologie Alimentari e Microbiologiche

Corso di Dottorato di Ricerca in Biotecnologie degli Alimenti
Ciclo XIX

tesi di Dottorato di Ricerca

Chemometric characterisation of physical-chemical fingerprints of food products

AGR15

Davide Ballabio

Tutor: Prof. Alberto Schiraldi
Co-tutor: Prof. Saverio Mannino
Coordinatore del Dottorato: Prof. Luciano Piergiovanni

Anno Accademico, 2005/2006

Cover illustration: "how cool this fingerprint is?" by db
PhD thesis:
Chemometric characterisation of physical-chemical fingerprints of food products
2006, Davide Ballabio

Special thanks to Roberto for teaching me chemometrics and for shearing both ideas and his office for coffee and cigarettes.

I would like to acknowledge prof. Saverio Mannino and prof. Alberto Schiraldi for their support and contribution to the development of what is now collected in this thesis.

Many thanks to Rasmus Bro for both his help, hospitality, suggestions and again for shearing his balcony for coffee and cigarettes.

Finally, thanks to all the people I met in the lab these years, in particular: Andrea, Alberto, Manuela, Viviana, Matteo, Stella, Susanna, Simona, Jibril and Thomas.

Ai miei
genitori

Contents

part I

Theory

---

chapter 1

Introduction

1.1 Chemometrics

Chemometrics has been defined in broad terms as the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods according to the International Chemometrics Society, 2002.

However, the definition of the word chemometrics has been a subject of discussion and no exact consensus is available, despite of the fact that two international scientific journals and numerous of international and national scientific societies are dedicated to chemometrics and use the word in their titles. It is known that Svante Wold invented the word chemometrics in 1972 to describe the discipline of extracting chemically relevant information from chemical experiments [Wold (1990)]. He tried to re-define the word as how to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into the data [Wold (1972)]. A more precise definition can be found in a textbook by Massart et al. [Massart *et al.* (1997)], stating that chemometrics is the chemical discipline

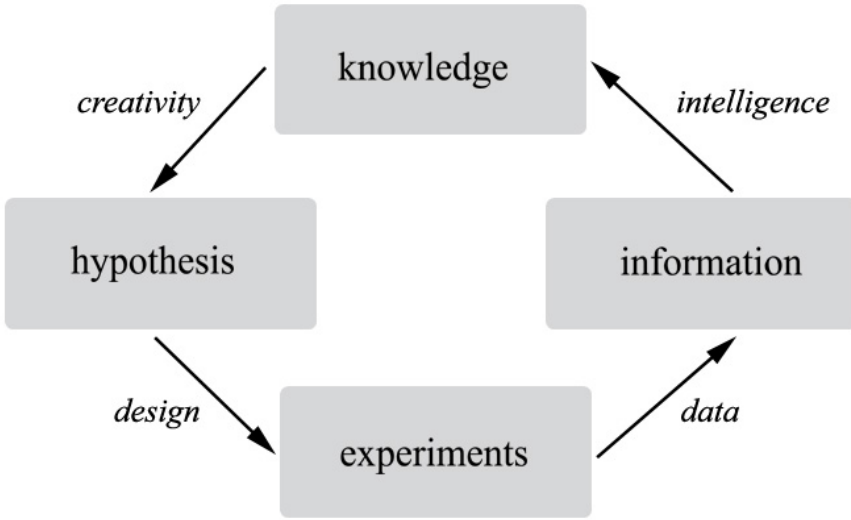

Figure 1.1: Chemometric rule in the knowledge circle

that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analysing chemical data; and (c) to obtain knowledge about chemical systems (Figure 1.1.). This definition is very close to the formulation used by Svante Wold and Bruce Kowalski when founding the first Chemometrics Society in 1974.

The use of chemometrics also implies the use of multivariate data analysis, in which several related samples are analysed simultaneously. A multivariate approach when handling and exploring complex chemical data and designing experiments is certainly part of the foundation of chemometrics. Multivariate data analysis as opposed to using only one or a few variables in the data analysis is based on the fact that complex problems - by nature - need multiple variables to be described. Thus, by using and combining more variables, more information about the chemical system can be retrieved. In standard multivariate data analysis, data are arranged in a two-way structure, a table or a matrix. An example is a table

in which each row corresponds to a sample and each column to a variable describing the complex system. This is the typical input for multivariate techniques: when these matrices are analysed by means of chemometrics, all the variables are considered at the same time and consequently the extracted information represent a global overview of the system.

Since chemometrics proved to be able to handle large amounts of data and to extract useful information, it has been successfully applied in different fields. During the last years it has so increased in uses and applications that now modern analytical techniques are usually combined with chemometric methods.

1.2 Chemometrics and food science

Quality control of production systems and authenticity testing of products are increasing in importance in food industry, since they represent the new required issues to compete in the present-day market. Both production systems and food products can be described as complex systems, where several factors can interact and play a fundamental rule: consequently all these factors should be monitored and their synergic effects controlled.

There is also a need in the food industry to rationalise and improve quality and process controls. The modern production systems require fast and automatic on-line monitoring, which should be able to extract the maximum amount of available information, in order to assure the optimal system functioning. On the other hand, food products acquire an higher value when their authenticity is protected, controlled and assured: in fact, consumers are more oriented towards purchasing food products of a certified origin. Consequently, during recent years there has been increasing interest in the origin authentication of food products, since authenticity can be often associated with food quality. Following this aim, the protected denomination of origin (PDO) for agricultural products has been introduced with official European regulations (2081/1992). Given these premises, it's clear how the development of reliable methods for assuring authenticity is becoming very important and several efforts have been

Table 1.1: Some references on authentication, classification and characterisation of wine and oil by means of chemometrics.

wine	
Arvanitoyannis *et al.* (1999)	Forina and Drava (1997)
Ortiz *et al.* (1996)	Kallithraka *et al.* (2001)
Buratti *et al.* (2004)	Kosir *et al.* (2001)
Cozzolino *et al.* (2005)	Ortiz *et al.* (1995)
Frias *et al.* (2003)	Gonzalez and Pena-Mendez (2000)
oil	
Armanino *et al.* (1989a)	Guadarrama *et al.* (2000)
Alves *et al.* (2005)	Garcia-Gonzalez and Aparicio (2004)
Aparicio *et al.* (1997)	Gonzalez Martin *et al.* (1999)
Boggia *et al.* (2002)	Guimet *et al.* (2004)
Brodnjak-Voncina *et al.* (2005)	Lanteri *et al.* (2002)
Christy *et al.* (2004)	Pinheiro and Esteves da Silva (2005)
Cerrato Oliveros *et al.* (2002)	Tsimidou *et al.* (1987)
Eddib and Nickless (1987)	Zupan *et al.* (1994)
Cosio *et al.* (2006)	Guadarrama *et al.* (2001)

made to authenticate the origin of food products, with different chemical and physical parameters and on several food matrices (Table 1.1 and Table 1.2).

The use of traditional analytical techniques does not always match with these requirements, since they can be time-consuming and expensive, while rapid and cheap methods are essential, in order to assure a continuous monitoring. Consequently, modern analytical techniques have been used for these issues, since they enable more rapid and non-invasive characterisation of foods: nuclear magnetic resonance (NMR), near infrared spectroscopy (NIR), electronic sensors and image analysis are only a few of the involved analytical methods. A common characteristic of these techniques is also the production of a huge amount of spectra, so that large data sets are usually obtained and must be interpreted.

Summarising, quality and authenticity control faces with complex systems, described by a large amount of data: as a consequence, specific tools should be used in order to assure the correct interpretation.

Multivariate Statistics can provide these specific tools: in the last years

Table 1.2: Some references on authentication, classification and characterisation of different food matrices by means of chemometrics.

honey	
Lopez *et al.* (1996)	Marini *et al.* (2004b)
Ampuero *et al.* (2004)	Nozal Nalda *et al.* (2005)
Benedetti *et al.* (2004)	Kelly *et al.* (2004)
spirits and beer	
Legin *et al.* (2005)	Kokkinofta and Theocharis (2005)
Cardoso *et al.* (2004)	Camean *et al.* (2001)
Lachenmeier *et al.* (2005)	Alcazar *et al.* (2002)
coffee	
Charlton *et al.* (2002)	Martin *et al.* (1999)
Maeztu *et al.* (2001)	
meat and cheese	
Sawyer *et al.* (2003)	Raatikainen *et al.* (2005)
Carpino *et al.* (2002)	Karoui *et al.* (2003)
Moller *et al.* (2003)	
rice, bread and potatoes	
Vinaixa *et al.* (2004)	Casanas *et al.* (2002)
Cocchi *et al.* (2005)	Marini *et al.* (2004a)
Vinaixa *et al.* (2005)	
pesto, juice and fruit	
Reid *et al.* (2004)	Antonelli *et al.* (2004)
Kim *et al.* (2000)	Llobet *et al.* (1999)
Christenses *et al.* (2005)	
saffron, sugar and vinegar	
Zalacain *et al.* (2005)	Ovejero-Lopez *et al.* (2005)
Bro (1999)	Benito *et al.* (1999)

multivariate statistics proved to be able to handle huge amount of data, to process them, and to give useful results that can be interpreted by the operators. Another fundamental characteristic of these statistical methods is the simplicity of their output responses: this means that the mathematical models used to interpret the data can be complex and structured, but the given responses must be clear, in order to face with their application purposes. In this way, for example, a binary response (yes/not, acceptable/not acceptable) can be provided to the on-line monitoring or to the authentication control systems and this make multivariate statistics

completely reliable for the proposed issues.

1.3 Thesis structure

This thesis is focused on the application of chemometrics on food data and mostly on the authentication and characterisation of food product fingerprints by means of multivariate analysis.

Consequently, classification method have been deepened, since classification is one of the fundamental methodologies in multivariate analysis and consists of finding a mathematical model able to recognise the membership of each object to its proper class, i.e. to identify specific class fingerprints. Once a classification model has been obtained, the membership of new objects to one of the defined classes can be predicted. Hence, statistical classification methods are central for building models that can assign samples to e.g. acceptable versus not-acceptable. While the theory for multivariate quantitative calibration is extensive and has been developed over the last thirty years [Martens and Naes (1989)], multivariate classification can be deeply explored.

Great attention has been also given to variable (or feature) selection methods, since classification models can be improved by selecting the variables which contain significant information for the specific task e.g. in this case the food chemical fingerprints. Variable selection have been applied both on classical and on multidimensional data sets: selection methods able to handle multidimensional data sets have not been explored yet and therefore developing new variable selection techniques for classification routines on these data can assume great interest in chemometrics.

Both classical and new proposed methods have been used to these purposes. Summarising, the structure of the thesis can be outlined with the following points:

- in the first part of the thesis a brief introduction to different data structures related to food products is presented in chapter 2, while the classical chemometrics methods used during the applications are

collected in chapter 3. Afterwards, three new proposed chemometric techniques are extensively described: the Classification And Influence Matrix Analysis (CAIMAN) deals with classification based on leverage-scaled functions (chapter 4); in chapter 5 a novel similarity measure based on Hasse matrices for the characterisation of sequential data is introduced, while in chapter 6 a new strategy for the compression of dimension of sequential data is described.

- in the second part of the thesis, applications of new and classical chemometric methods on several different food data are presented. In chapter 7 a brief resume of all the applications is showed; in chapters 8 and 9 the application of CAIMAN and Hasse similarity measures, respectively; in chapter 10 all the applications dealing with electronic sensors; in chapter 11 the applications based on acoustical and mechanical signals; in chapter 12 the applications on multidimensional (multiway) data sets of sensory data; finally, in chapter 13 the classification of GCMS data of wine based on data compression and variable selection.

In order to apply the new algorithms, some routines have been developed during the thesis by using the software MatLab 6.5 (MathWorks). The code sources of these routines and the web sites where it is possible to download them are presented in the appendix.

chapter 2

Analytical multivariate data

Analytical chemical systems can be described by means of tables (matrices) in which each row corresponds to a sample and each column to a variable describing the system. This is the typical input for chemometric methods, which consider all the variables at the same time and extract information in a global way.

2.1 Classical data structures

Traditionally, in chemometrics, **X** denotes the data matrix, while the number of rows (samples) and columns (variables) is usually indicated by n and p respectively. Each entry of this matrix, x_{ij}, represents the value of the p-th variable for the i-th sample. Other qualitative information regarding the samples can be added to the data matrix, in order to make the results more readable, but only the data matrix **X** is considered during the algorithms.

Depending on the applied chemometric method, some other information can be needed in order to develop a multivariate model: when classification or regression techniques are used, a response vector (or matrix) **Y**

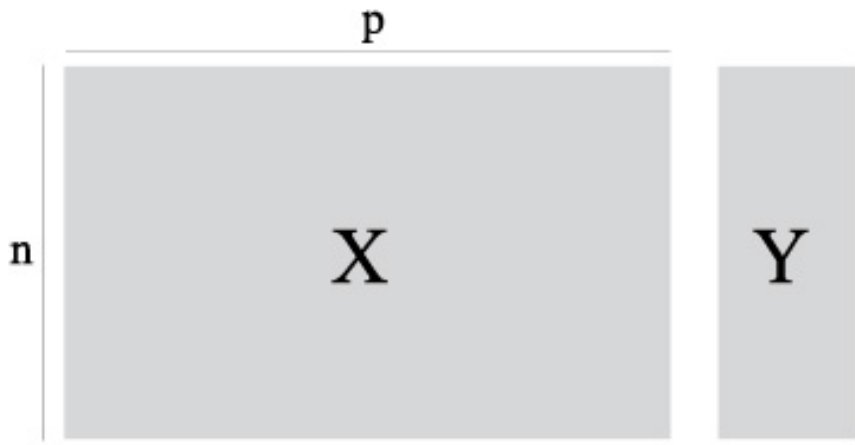

Figure 2.1: Typical representation of analytical chemical data

is used during the calculations. This vector (matrix) contains the qualitative or quantitative responses to be modelled and has usually dimensions n times 1, i.e. each entry y_i of the vector represents the value of the response for the i-th sample. If more responses are considered in the same model, $\mathbf{Y}$ has dimensions n times r, where r is the number of considered responses. In Figure 2.1 a schematic representation of a typical analytical chemical data structure is shown.

In most of the applications of this thesis, each column of $\mathbf{X}$ represent a single chemical or physical parameter, describing the system, while in some applications the whole analytical profiles have been considered: in the following paragraphs brief introductions to the data structure of this second kind of data are presented.

2.2 Electronic nose sensor signals

Electronic noses are basically made by non-selective gas sensors able to simulate human sensing: the main principle involved in electronic noses is the transfer of the total headspace of a sample to a sensor array, where

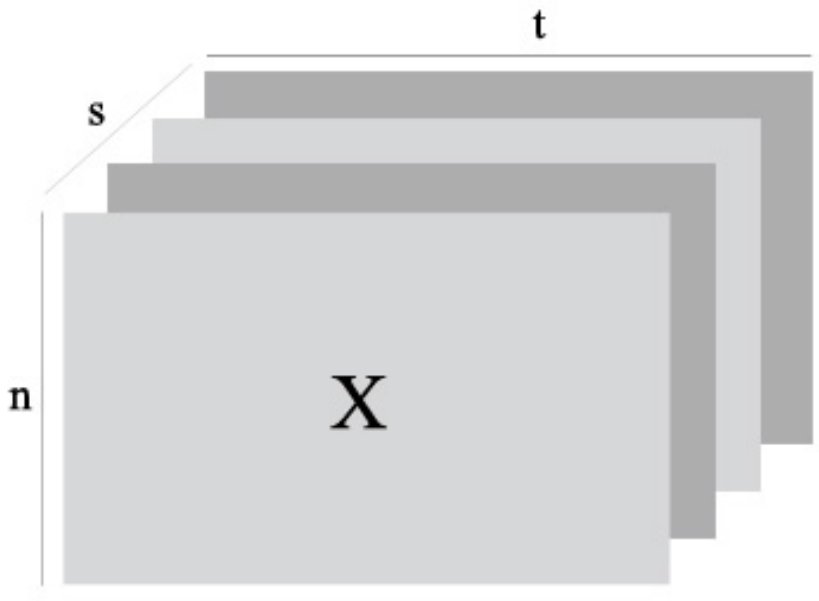

Figure 2.2: Typical representation of three-way electronic nose data

each sensor has partial specificity to a wide range of aroma molecules.

Each sensor gives an a-specific fingerprint of the analysed food product by means of a signal collected during the time. Parameters are usually extracted from the electronic nose signals. in this case a data matrix $\mathbf{X}$ with dimensions n times s is obtained, where s is the number of considered sensors. As a result of this parameterisation, relevant information from the raw data could be lost, while by preserving the time information, in some cases more information can be extracted. In this case, the data matrix can be arranged as a cube and subsequently analysed by means of appropriate chemometric methods: consequently $\mathbf{X}$ will have dimensions n (samples) times s (sensors) times t, where t is the number of considered points in the electronic nose signal (Figure 2.2). Each entry x_{ijk} of this three-way array represents the value of the signal analysed by the j-th sensor for the i-th sample at the k-th time.

Otherwise the three-way array $\mathbf{X}$ can be unfolded and a data matrix

with dimensions n times $s*t$ can be derived: the so-called unfolding simply consists in the alignment of all the slices of the three-way array in a simple two-way matrix, that can be subsequently analysed by means of the majority of the classical chemometric methods.

2.3 Mechanical and acoustic signals

Mechanical measurements of crispness of food products are performed on instruments originally developed for material science, providing physical parameters with fundamental significance in terms of rheological properties, i.e. the materials' response to the applied force. The texture analysers typically have a crosshead containing a load cell, which is driven vertically at a range of constant speed. The load is recorded relative to time or to deformation distance, and recorded as force-deformation curves.

Furthermore, when a force is applied to a crisp item, its structure is stressed until a critical point is reached: the action of external force causes the rupture of the brittle walls of the cellular structure, which start to vibrate. The vibration is transmitted trough the air as acoustic waves, which generates the sound. The amplitude of displacement of molecules from the equilibrium is recorded by acoustic recording during the mechanical test and is recorded as amplitude - time curves.

Several parameters can be extracted from mechanical and acoustic signals, collected in a data matrix and subsequently analysed. In this case the data matrix $\mathbf{X}$ has dimensions n times p, where p is the number of parameters extracted from the original mechanical and acoustic profiles. Depending on the food matrix in analysis, the whole signal could also be considered: in this case, as before, the data matrix has dimensions n (samples) times t, where t is the number of considered points in the mechanical or acoustic profile.

2.4 Gas Chromatography profiles

Gas chromatography (GC) is a type of chromatography in which the mobile phase is a carrier gas and the stationary phase is a microscopic layer of liquid on an inert solid support, inside glass or metal tubing, called a column.

Gas chromatography is used to separate chemicals in a complex sample. A gas chromatograph uses a flow-through narrow tube known as the column, through which different chemical constituents of a sample pass in a gas stream (carrier gas, mobile phase) at different rates depending on their various chemical and physical properties and their interaction with a specific column filling, called the stationary phase. As the chemicals exit the end of the column, they are detected and identified electronically. The function of the stationary phase in the column is to separate different components, causing each one to exit the column at a different time (retention time). Other parameters that can be used to alter the order or time of retention are the carrier gas flow rate, and the temperature.

The rate at which the molecules progress along the column depends on the strength of adsorption, which in turn depends on the type of molecule and on the stationary phase materials. Since each type of molecule has a different rate of progression, the various components of the analyte mixture are separated as they progress along the column and reach the end of the column at different times (retention time). A detector is used to monitor the outlet stream from the column; thus, the time at which each component reaches the outlet and the amount of that component can be determined. Generally, substances are identified by the order in which they emerge from the column and by the retention time of the analyte in the column. In this case, a data matrix containing the concentration of each detected compound in each sample can be obtained.

On the other hand, Gas Chromatography profiles can be also used as a chemical fingerprint and consequently considered all together in the data matrix. In this second case, the final data matrix will have a number of columns p equal to the number of considered points of the Gas

Chromatography profiles.

2.5 Time intensity signals

Time Intensity (TI) is a dynamic sensory method, in which sensory attributes can be evaluated as it changes over time. The time-intensity curves are based on the responses of the assessors and are made by an increasing part, a plateau and a final decreasing part. In order to analyse TI data with the classical multivariate techniques, a weighted average of the individual curves is usually calculated or a parameter-extractions is applied. As usual, in this way, a part of information could be lost. TI data can be also seen as a three-way matrix, where the first dimension (also called mode) is represented by samples, the second mode by time and the third mode by assessors.

Consequently the data matrix $\mathbf{X}$ will have dimensions n (samples) times t times a, where a is the number of assessors and t is the number of time units considered for the signal acquisition. Each entry x_{ijk} of $\mathbf{X}$ represents the value of the intensity assessed by the k-th assessor for the i-th sample at the k-th time. The graphical representation of the structure of this kind of data is similar to the one represented in Figure 2.2.

chapter 3

Chemometric methods

The chemometric methods used during the applications of this thesis are collected and briefly explained in the present chapter.

3.1 Data structure analysis

Principal Component Analysis

Principal Component Analysis (PCA) is the most common method used to display the structure of the multivariate data [Wold *et al.* (1987), Kvalheim (1987)]. PCA is a well-known chemometric technique, which projects the data in a reduced hyperspace, defined by the principal components. These are linear combinations of the original variables, with the first principal component having the largest variance, the second principal component having the second-largest variance, and so on. In this way it is possible to retain a number of components lower than the number of original variables, i.e. it is possible to reduce the data dimension: the number of components to be retained can be chosen on the basis of different parameters, linked to the variance explained by each principal component.

Multidimensional Scaling

One of the most important goals in visualising data is to find a sense of how near or far points are from each other. Sometime, this can be faced with a simple scatter plot, while when several variables describe the data, appropriate techniques for the dimension reduction are needed. Multidimensional scaling (MDS) reconstructs dissimilarities between pairs of samples by distances in a low-dimensional space [Winsberg and Carroll (1989), Kruskal (1964)]: MDS allows the visualisation of the similarities/dissimilarities between the samples and produces a data representation in a small number of dimensions. In general, MDS attempts to arrange the samples in a space with a particular number of dimensions so as to reproduce the observed distances. Thus, a scatter plot of the samples produced by MDS provides a visual representation of the original distances and can be used to easily analyse the relationships between samples.

Besides the classical data projection techniques, such as Principal Component Analysis, the selection and retention of a correct number of dimensions must be considered. When multidimensional scaling is applied, the number of dimensions to be taken into account can be selected on the basis of the residuals between the original distances and the distances represented in the low-dimensional space: lower is the residual, better is the data approximation obtained in the reduced space. Therefore, the optimal number of dimensions can be selected by optimising these residuals.

3.2 Classification methods

Discriminant Analysis

Discriminant Analysis (DA) is one of the most popular classifiers [McLachlan (1992)]. The method is a probabilistic parametric classification technique: it maximises the variance between categories and minimises the variance within categories, by means of a data projection from a high dimensional space to a low dimensional space. In this way, a number of

orthogonal linear discriminant functions equal to the number of categories minus one is obtained.

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are used in turn depending on the linear/non-linear class separability and on the reliability of the class covariance matrices. In fact, for LDA only the pooled covariance matrix is calculated, while for QDA the covariance matrix is calculated for each class separately. In order to estimate the class covariance matrix, the number of class objects must be greater than the number of variables, while LDA can be applied only if the total number of samples is greater than the number of variables. Linear and Quadratic Discriminant Analysis are both based on the Bayes rule and require a multinormality assumption.

Partial Least Square Discriminant Analysis

Partial Least Square Discriminant Analysis (PLS-DA) is essentially the inverse-least-squares approach to LDA [Barker and Rayens (2003)], with the noise reduction and variable selection advantages of PLS. In PLS-DA the interpretation can be performed with respect to the original high-dimensional data space. Basically, PLS-DA performs a dimension reduction on highly collinear data, just as PLS does [Stahle and Wold (1987)], but the predictions in a PLS-DA model are used to classify unknown samples, i.e. to predict the class membership of each sample.

The distribution of calibration-sample predictions obtained from a PLS-DA model built for two or more logical classes can be used to automatically determine a threshold value, which will best split those classes with the least probability of false classifications for future predictions. In the present work, PLS-DA models have been investigated both considering all the available classes at the same time and considering each class at a time, i.e. building a classification model for each class versus all the others.

Extended Canonical Variates Analysis

Extended Canonical Variates Analysis (ECVA) has been recently proposed as a new classifier [Norgaard *et al.* (2006)]. It can be defined as a modification of the standard Canonical Variates Analysis (CVA) method, able to cope with collinear high dimensional data. The method uses Partial Least Squares regression as an engine for solving an eigenvector problem involving singular covariance matrices: the proposed method calculates canonical variates directly in the original high-dimensional space making it possible to interpret the model in relation to this space.

The modified CVA method forces the discriminative information into the first canonical variates and the weight vectors found in the ECVA method hold the same properties as weight vectors of the standard CVA method. The combination of the suggested method with e.g. Linear Discriminant Analysis as a classifier gives an efficient operational tool for classification and discrimination of collinear data.

Kohonen and Counter Propagation Artificial Neural Networks

Counterpropagation Artificial Neural Networks (CP-ANN) consist of two layers, a Kohonen layer, which can be used also for the data structure analysis, and an output layer [Zupan (1994), Zupan *et al.* (1997), Zupan and Gasteiger (1999)]. This technique is based on the search for sample similarities and allows projecting the samples into a topological space where similar samples are close to each other and dissimilar ones are far apart.

The projection space is the Kohonen Layer and is usually characterised by being a square toroidal space; this means that each edge of the Kohonen map has to be seen as connected with the opposite one. This layer is built with a certain number of neurons: the weights of each neuron are updated on the basis of the input vectors, for a certain number of times (called training epochs). The user must choose the number of neurons and training epochs. In each training step, for each input vector, a winning

neuron in the Kohonen layer is selected; the winning neuron is the neuron most similar to the input vector. Then, the weights of the r-th neuron ($\mathbf{w}_r$) change on the basis of the difference between their old values and the values of the input vector ($\mathbf{x}_i$); this correction is scaled according to the topological distance from the winning neuron (d_r):

$$\Delta \mathbf{w}_r = \eta \cdot \left(1 - \frac{d_r}{1 + d_{max}}\right) \cdot \left(\mathbf{x}_i - \mathbf{w}_r^{old}\right) \tag{3.1}$$

where η is the learning rate. The topological distance d_r is defined as the number of neurons between the r-th neuron and the winning neuron. The size of the considered neighbourhood d_{max} decreases during the training phase. The learning rate η is also changing during the training phase, as follows:

$$\eta = (\eta_{start} - \eta_{final}) \cdot \left(1 - \frac{t}{t_{tot}}\right) + \eta_{final} \tag{3.2}$$

where t is the number of the current training epoch, t_{tot} is the total number of training epochs, η_{start} and η_{final} are respectively the learning rates at the beginning and at the end of the training step.

In addition to the Kohonen layer, an output layer is defined. The weights of the output layer $\mathbf{y}_r$ are also updated, in a supervised manner, considering the response component of each i-th sample ($\mathbf{c}_i$):

$$\Delta \mathbf{y}_r = \eta \cdot \left(1 - \frac{d_r}{1 + d_{max}}\right) \cdot \left(\mathbf{c}_i - \mathbf{y}_r^{old}\right) \tag{3.3}$$

Counterpropagation Artificial Neural Networks can be considered a class modelling method, since it is able to recognise samples belonging to none of the class spaces. This happens when samples are assigned to neurons where the weights of the output layer are similar, i.e. the neuron can not be assigned to a specific class. The original variables are usually pretreated by scaling in the 0 to 1 range in order to make them comparable with the networks weights.

3.3 Variable selection techniques

Variable selection techniques can be used in order to improve chemometric models: these techniques are in fact able to retain and preserve only the variables, which contain significant information for a specific task. Moreover, the increase of dimensions and complexity of datasets and the decrease in time-consumption in algorithms support approaches based on variable selection techniques.

All subset selection

The all subset selection method is the most simple variable selection method: this technique searches all the possible models by using all the available combinations of variables. Usually, an exhaustive search of all the possible solutions is not feasible: in fact, if there is a total number of p variables, the number N of all the possible combinations is:

$$N = \frac{p!}{(p-c)!c!} \tag{3.4}$$

where c is the number of considered variables for each combination. This means that considering 50 variables ($p = 50$) and selecting just 5 variables ($c = 5$), the total number of combinations is $N = 50!/((50\text{-}5)!5!) = 2118760$, i.e. a huge number of models should be computed. Consequently, this selection method is applicable only with a low number of variables.

Forward selection

Forward Variable Selection is a simple selection technique, which starts with no variables and adds one variable at a time to the chemometric model: the inclusion of a variable is based on the optimisation of a chosen parameter [Jennrich (1977)] that depends on the selection task, e.g. a classification quality parameter, such as the number of errors, or a regression parameter, such as the response residuals.

Forward Selection can depend on the first selected variables, since all

the others are added to the model when these variables are still present and consequently the new variables can just contribute to solve marginal modelling fittings. On the other hand this method is usually faster and less time-consuming than other classification techniques, such as Genetic Algorithms, which explore in a more complete way the available information and searches for the best solution with an higher number of possibilities, but, as a consequence, are more time-consuming.

Genetic Algorithms

Genetic Algorithms (GAs) select subsets of variables that maximise the predictive power of multivariate models and perform this selection by considering populations of models generated with an evolution process and optimised according to an objective function [Goldberg (1989), Leardi (2001), Leardi *et al.* (1992)].

A population is made of a series of chromosomes. Each chromosome is a binary vector, where each position (a gene) corresponds to a variable (i.e. a chromosome represents a model made up of a subset of selected variables). The evolution process is based on three main steps: initially the model population is randomly built. The value of the objective function of each model is calculated and the models are then ordered with respect to this objective function.

After that, the reproduction step selects pairs of models (parents) and from each pair of models a new model (son) is generated preserving the common characteristics of the parents (i.e. variables excluded in both models remain excluded; variables included in both models remain included) and mixing the opposite characteristics. If the generated son coincides with one of the individuals already present in the actual population, it is rejected; otherwise, it is evaluated. If the objective function value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise, it is no longer considered. This procedure is repeated for several pairs.

The mutation step instead changes every gene of each chromosome into

its opposite according to a defined probability. If the objective function of each mutated model is better then the worst value in the population, the model is included in the population. Reproduction and mutation steps are alternatively repeated until a stop condition is occurred or the evolution process is ended arbitrarily.

3.4 Multiway Analysis

The described bilinear models are used in analysis of data matrices in two dimensions, structured as samples times variables. Multi-way data analysis refers to multivariate data analysis performed of data arrays in higher dimensions than two. The multi-way approach allows information about each dimension to be obtained at the same time: it has the ability to show the global, and at the same time, particular information in an easier way. In fact, in this way all the interrelations in-between dimensions can also be analysed and considered, while with a 2-way approach a part of this information is necessarily lost. The most common technique for multi-way data structure analysis is the Parallel Factor Analysis.

Parallel Factor Analysis

Parallel Factor Analysis (PARAFAC) can be basically considered as an evolution of the bilinear PCA analysis [Smilde *et al.* (2004), Bro (1997, 1998)]: it performs a decomposition of the original data matrix with multiple variables to a set of scores and loadings, but with loadings in more than one direction. The principle behind the PARAFAC decomposition is to minimise the sum of squares of the residual, e_{ijk}, as indicated in 3.5 for a three-way PARAFAC model.

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{3.5}$$

where x_{ijk} represents the data for i.th sample, for the j-th and k-th variables of the two different dimensions. The three way data array are thus

decomposed into a set of sample scores a_{if}, loadings in the first variable directions, b_{jf} and loadings in the second variable direction, c_{kf}.

The described PARAFAC model is a trilinear decomposition that requires trilinearity of the data. This means that each phenomenon in data which is to be modelled, can be explained by the product of a score value and a set of loading vectors, 3.5, implying that the structure in the two variable directions are independent.

An evolution of the PARAFAC model is the so called PARAFAC2 model [Bro *et al.* (1999), Smilde *et al.* (2004)]. If one of the dimensions of the multi-way data describes a time-depending value, the PARAFAC2 model has the advantage of automatically overcoming the presence of shifts in the time mode. The structure of the PARAFAC2 model can be written as:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k \qquad k = 1, ..., K \tag{3.6}$$

where $\mathbf{X}_k$ is the I x J sub matrix with elements x_{ijk} and k fixed and where $\mathbf{B}_k$ is fixed such that:

$$\mathbf{B}_k^{\mathrm{T}}\mathbf{B}_k = \mathbf{B}_p^{\mathrm{T}}\mathbf{B}_p \tag{3.7}$$

for any k and p. Therefore, the PARAFAC2 model allows that every variable can have its distinct set of time loadings. Hence, there is not only one loading matrix for the time profiles as there would be in a PARAFAC model.

3.5 Model validation

Since the main aim of chemometric models (both regression and classification models) is the application of the models to unknown samples, great attention has been focused on their predictive capabilities.

In fact, if we consider a simple regression model, it is demonstrated that the fitting performances of the model always increase if new variable are

added, even if these new variables are random variables or do not contain useful information. On the other hand, the predictive performances of the model increase only when informative variables are added to the model, otherwise they decrease. This simple case clarifies why the prediction capabilities of a model have to be always tested. All the models carried out in this thesis have been validated using different procedures, depending on the case in analysis.

The leave-one-out (LOO) procedure is one of the most used validation techniques: it removes each sample from the data set, one at a time, then the model is rebuilt and the response of the removed sample is predicted by using the obtained model. All the samples are sequentially removed and predicted. Finally the mean of the predicted responses obtained on all the samples is calculated.

Since LOO can provide over optimistic results [Golbraikh and Tropsha (2002)], also a more robust validation technique (leave-more-out, LMO) has been applied [Burden *et al.* (1997), Baumann and Stiefl (2004)]. In the LMO procedure, a percentage s of samples is randomly removed from the data set; then, the model is rebuilt without these objects and the responses of the removed sample are predicted by the obtained model. This procedure is repeated r times, always with a random selection of s samples. Finally the mean of the predicted responses is calculated.

As explained, the LMO procedure is more robust then LOO, but also more time-consuming. Furthermore, LOO gives always the same result, i.e. it is perfectly reproducible, while LMO is based on a initial random selection of the samples to be predicted and can provide different results each time it is applied. The advantages (robustness) and disadvantages (time-consuming, not perfectly reproducible) of LMO can be avoided by means of another validation technique: the samples are divided in different groups (cross-validation groups) and one group at a time is removed from the training set, the model is rebuilt without the left out objects and the responses of the removed sample are predicted. This procedure is repeated for each cross-validation group, and finally the mean of the predicted responses is calculated.

chapter 4

Classification based on leverages

4.1 Introduction

Classification is one of the fundamental methodologies in chemometrics and consists in finding a mathematical model able to recognise the membership of each object to its proper class. Once a (acceptable/good) classification model has been obtained, the membership of new objects to one (or none) of the defined classes can be predicted. Information provided from a number of measurements is used to establish whether samples originate from one or more classes. Based on different mathematical approaches, several multivariate classification techniques (classifiers) have been proposed since several years [Frank and Friedman (1989), Vandeginste *et al.* (1998), James (1985), Hand (1997)] and used in a huge of fields, such as food chemistry, archaeometry, quantitative structure-activity relationships, pharmaceutical chemistry, environmental sciences, social and economic sciences.

Classifiers which always assign an object to one of the predefined classes are called discriminant classifiers, while class modelling classifiers are characterised by class space models, allowing, besides the object as-

signment to a class, also to detect objects belonging to none of the class spaces and/or detect objects potentially belonging to more than one class space.

The Nearest Mean Classifier (NMC), based on the Euclidean distance, is the most intuitive and simple classification method. Object classification is based on the minimum distance from the class centroid. Among the most popular classifiers there is the Discriminant Analysis (DA), which is based on the Bayes rule and multinormality assumptions [McLachlan (1992)]; Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Regularized Discriminant Analysis (RDA) [Friedman (1989)] are used in turn depending on the linear/nonlinear class separabilities and on the reliability of the class covariance matrices. Classification And Regression Trees (CART) is a simple important complementary approach based on the construction of a classification tree where each knot split is based on the best discriminant single variable [Breiman *et al.* (1984)]; the final leaves correspond to the classes. K-Nearest Neighbours (KNN) is a method where the classification of the objects is based on local information determined by K neighbours [Kowalski and Bender (1972)]; its classification performance has been demonstrated to be always acceptable. The method UNEQ is a class modelling variant of QDA where the class models are defined by assuming the multinormal distribution of the data [Derde and Massart (1986)]. In the present paper the classical UNEQ method based on the Mahalanobis distance has been used instead of the probability density approach. Soft Independent Model of Class Analogy (SIMCA) is also among the class modelling approaches [Wold (1976)], where the class space description is based on principal components and the classes are separately modelled. Other classifiers such as Partial Least Squares-Discriminant Analysis [Stahle and Wold (1987)] and Stepwise Discriminant Analysis [Jennrich (1977)] are regularly used, while other classifiers such as potential classifiers [Coomans *et al.* (1981), Forina *et al.* (1991)], ALLOC [Van der Voet and Coenegracht (1988)], DASCO [Frank (1988)], PRIMA [Juriaeskay and Veress (1985)] are actually not often used.

A new classification method is here proposed [Todeschini *et al.* (2005)], based on the influence matrix or leverage matrix [Cook and Weisberg (1982)]. The diagonal elements of this matrix, widely used in regression analysis, provide information on the influence of each sample within the model, being related to the distance of the sample from the center of the model. Starting from this matrix and analysing the leverage properties, the new proposed method, named Classification And Influence Matrix Analysis (CAIMAN), has been developed with the aim of modelling each class by using the class scatter matrix and calculating the leverage value of each object from each class scatter matrix. Depending on the purposes of the classification analysis, the CAIMAN method can be used in three different outlines: the Discriminant-CAIMAN (D-CAIMAN) which is a discriminant classification method, the Modelling-CAIMAN (M-CAIMAN), a class modelling method able to identify outliers and confused objects, and the Asymmetric-CAIMAN (A-CAIMAN) which restricts the classification to only one reference class to be modelled [Dunn III and Wold (1980)].

4.2 CAIMAN theory

Unlike traditional analytical methods, electronic nose sensor responses do not provide information on the nature of the compounds under investigation, but only give a digital fingerprint of the food product, which can be subsequently investigated by means of chemometric methods. In fact, multivariate analysis proved to be especially able to handle the large amounts of data produced by modern analytical techniques and has been successfully applied on electronic nose data.

4.2.1 The leverage matrix

Also called hat matrix or influence matrix, the leverage matrix $\mathbf{H}$ is defined as:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \tag{4.1}$$

where $\mathbf{X}$ is the data matrix, collecting n objects described by p variables: consequently the leverage matrix has dimensions n times n.

This matrix is commonly used in regression analysis [Cook and Weisberg (1982)], where it gives information on the relationship between the experimental $\mathbf{Y}$ response and the $\hat{\mathbf{Y}}$ response calculated by a regression model:

$$\hat{\mathbf{Y}} = \mathbf{HY} \tag{4.2}$$

The most important terms of the $\mathbf{H}$ matrix are the diagonal elements h_{ii}, namely the leverages, because they have some interesting mathematical properties. Each leverage h_{ii} measures the influence of the i-th object projection onto the model descriptor space. As well as in Ordinary Least Squares regression (OLS), also in PCA, PCR and PLS, low values of the leverages are associated to objects near the center of the descriptor space, while high values are associated to objects lying in the boundary. The leverage is bounded as the following:

$$c_0 \leq h_{ii} \leq \frac{1}{c_i} \tag{4.3}$$

where the constant c_0 is equal to 1/n if the descriptor matrix $\mathbf{X}$ contains a constant (e.g. the model intercept), otherwise c_0 is equal to zero. The quantity c_i represents the number of times that the i-th row of $\mathbf{X}$, i.e. the i-th object, is replicated. If the i-th object occurs only once in the training set and the non-diagonal terms h_{ij} of the influence matrix are all equal to zero, then the maximum value of h_{ii} will be 1.

The leave-one-out estimate of the leverage h^*, which in the following will be referred to as the predicted leverage, can be directly derived from the leverage h_{ii} obtained when the object belongs to the training set:

$$h^*_{ii} = \frac{h_{ii}}{1 - h_{ii}} \tag{4.4}$$

Unlike the leverages of the training objects, the predicted leverages have

only a lower bound:

$$c_0 \leq h_{ii}^* \leq \infty \tag{4.5}$$

However, an upper control limit (e.g. 1 or 2) is usually taken to decide whether an object, which has not been used in the model space definition, may belong or not to the model space.

4.2.2 CAIMAN approach

Given a data matrix $\mathbf{X}$, collecting n objects, each described by p variables and belonging to one of G defined classes, the sub-matrix $\mathbf{X_g}$ is defined by collecting the n_g objects assigned to the g-th class; these objects are centered on the g-th class centroid. Data in $\mathbf{X_g}$ are mean centered so that the $(\mathbf{X_g^T X_g})$ class dispersion matrix contains information on size and shape of the class descriptor space, being related to the class covariance matrix.

The basic idea of the new method CAIMAN is to represent each class space by the corresponding dispersion matrix estimated by the training objects belonging to the class. Obviously, as it is common to all methods based on class covariance matrices, to obtain a good estimate of the class dispersion matrix, the number of class objects should be significantly greater than the number of variables; an ratio between objects and variables greater than 2 or 3 is usually suggested. Then, each object is characterised by a G-dimensional leverage vector:

$$\mathbf{h}_i \equiv \{h_{i1} ... h_{ig}^* ... h_{iG}\} \tag{4.6}$$

which gives information on the position of the object with respect to all the class spaces. For each g-th class, the leverage of the i-th object is calculated as:

$$h_{ig} = (\mathbf{x}_i - \bar{\mathbf{x}}_g)^{\mathrm{T}} \left(\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g\right)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_g) \tag{4.7}$$

where $\bar{\mathbf{x}}_g$ is the g-th class centroid. If the i-th object does not belong to the g-th class the calculated leverage is already a predicted leverage, as no information about the object has been used in defining the class model space; otherwise, the calculated leverage is corrected according to the following:

$$h_{ig}^{*} = \frac{h_{ig}}{1 - h_{ig}} \tag{4.8}$$

in order to obtain a leave-one-out estimate. For the meaning of the leverage, it should be expected that typical and characteristic objects of the class have low leverages, while objects far from the class have high leverages. In an ideal case with some classes, the leverage vector of each object should have one low value, the one related to the membership class, and all the other ones significantly greater.

Moreover, assuming a constant number of variables, by increasing the number of objects of a class, the mean value of the object leverage decreases, thus obtaining a more significant characterisation of the class by means of its objects; this behaviour can be related to the role of the a priori class probability defined in terms of object frequency. A mathematical relationship between the leverage and the Mahalanobis distance (MD), used for instance in the discriminant analysis, is known:

$$h_{ig} = \frac{1}{n_g} + \frac{MD_{ig}^{2}}{n_g - 1} \tag{4.9}$$

by which the leverage value can be interpreted as a distance from the center of the class model space, weighted by the class dispersion matrix. However, it must be noted that the leverage does not fulfil the required properties for a distance measure. Leverage and Mahalanobis distances have different scales, as it can be noticed from the two relationships:

$$\frac{1}{n_g} \leq h_{ig} \leq 1 \tag{4.10}$$

$$0 \leq MD_{ig}^2 \leq \frac{(n_g - 1)^2}{n_g} \leq n_g \tag{4.11}$$

Obviously, the advantage of the leverage scale is that the maximum value is independent of the considered class and its size, allowing simple and objective rules to establish the object membership to model spaces. In effect, leverage values lower than 1 indicate apparent class membership. The most simple leverage-based classifier may be the minimum leverage classifier:

$$i \rightarrow g \qquad \text{if} \qquad \min_g\{h_{ig}\} \tag{4.12}$$

according to which the object is assigned to the class for which the leverage is minimum. This classifier exploits the information given by a single leverage, i.e. the minimum leverage over the classes, which, being calculated independently of the other classes, does not encode information on relationships among classes. As it will be seen in the following, this limit of the minimum leverage makes it unsuitable to solve classification problems characterised by more complex class structure.

A simple example is here introduced in order to show how the leverage can be interpreted and exploited to solve simple classification problems. A small data set (Simul1), constituted by 40 objects described by 2 variables and distributed into 3 classes (n_A=14, n_B=14, n_C=12) has been created (Table 4.1 and Table 4.2); in columns 4-6 of Table 4.1 and Table 4.2 the leverage values are collected. A scatterplot of the objects is shown in Figure 4.1, where a perfect separability of class A and a quite good separability between classes B and C can be observed. Figure 4.2 shows the leverage plots that, similar to the Coomans plots, are scatterplots obtained by projecting the objects in the space defined by the leverages referring to two chosen classes. Each axis of the leverage plot represents a class and, therefore, farther the objects are from the axis origin, greater their distance is from the corresponding class space. Specifically, objects falling in the left bottom corner are close to both the class spaces; objects falling in the left top corner are very close to the class represented by

Table 4.1: Simul1 data set (objects 1-20): X1 and X2 are the two descriptive variables; h-A, h-B and h-C are the leverages related to classes A, B, and C, respectively.

ID	X1	X2	h-A	h-B	h-C	Class
1	3.75	1.45	0.07	6.78	10.24	A
2	3.25	1.50	0.11	4.50	8.68	A
3	3.90	1.30	0.30	8.08	11.45	A
4	3.70	1.50	0.04	6.37	9.85	A
5	3.60	1.00	0.76	7.59	11.86	A
6	3.30	1.70	0.09	4.12	7.95	A
7	3.40	1.90	0.22	3.95	7.38	A
8	3.40	1.60	0.02	4.77	8.59	A
9	3.50	1.70	0.03	4.88	8.43	A
10	3.10	1.60	0.27	3.74	7.98	A
11	3.80	1.90	0.44	5.63	8.51	A
12	3.80	1.60	0.11	6.54	9.75	A
13	3.20	1.40	0.22	4.62	9.01	A
14	3.65	1.55	0.02	5.98	9.48	A
15	2.30	4.00	9.49	0.31	1.23	B
16	2.80	4.50	12.54	0.11	0.42	B
17	3.15	4.25	10.42	0.28	0.89	B
18	2.90	4.70	14.21	0.23	0.27	B
19	2.90	3.60	6.13	0.23	1.74	B
20	2.85	4.15	9.76	0.02	0.82	B

the horizontal axis and far from the other one; on the contrary, objects falling in the right bottom corner are very close to the class represented by the vertical axis and far from the other one; finally, objects falling in the right top corner are far from both the classes. Considering the class A of the Simul1 data set, it can be seen both in Table 4.1 and Figure 4.2 that the leverages for A of all the objects effectively belonging to A (ID 1-14) are very small, indicating an apparent membership to this class, while leverages for B and C of the same objects are quite large, indicating a large distance of these objects from both B and C class spaces. For example, object 9, whose leverage vector is: 0.033, 4.879, 8.425, is typical of class A, having a very low leverage value for class A with respect to the other two classes. In this case, a minimum leverage classifier will correctly

Table 4.2: Simul1 data set (objects 21-40): X1 and X2 are the two descriptive variables; h-A, h-B and h-C are the leverages related to classes A, B, and C, respectively.

ID	X1	X2	h-A	h-B	h-C	Class
21	2.60	3.50	5.93	0.10	1.87	B
22	2.40	3.70	7.40	0.12	1.59	B
23	2.50	4.00	9.11	0.07	1.07	B
24	2.60	4.40	11.92	0.15	0.55	B
25	2.30	3.30	5.51	0.32	2.42	B
26	2.70	3.25	4.53	0.35	2.39	B
27	3.00	4.20	10.06	0.10	0.83	B
28	2.90	4.30	10.86	0.06	0.65	B
29	2.70	5.10	18.12	0.61	0.08	C
30	2.90	5.60	23.43	1.07	0.02	C
31	2.50	5.20	19.32	0.97	0.21	C
32	3.20	5.30	20.15	0.67	0.17	C
33	3.20	5.70	24.70	1.10	0.18	C
34	2.80	4.80	15.17	0.27	0.23	C
35	2.80	6.10	29.58	2.02	0.39	C
36	2.60	5.60	23.57	1.41	0.15	C
37	3.00	6.10	29.60	1.80	0.34	C
38	3.30	5.10	18.11	0.58	0.45	C
39	2.50	5.00	17.33	0.72	0.28	C
40	3.00	5.20	19.03	0.55	0.05	C

classify all the class A objects. The same consideration also holds for the other two classes, since the leverage for the membership class is smaller than the other ones for all the objects.

Moreover, it can be also highlighted that object 5 belonging to the class A has a relatively high leverage value (0.761) with respect to the other objects of the same class. This is due to the fact that object 5 is relatively far along the X2 axis, thus falling near the class boundary.

4.2.3 Discriminant and Modelling CAIMAN

A classification rule based on the minimum leverage is useful in several cases, but it is not able to properly deal with asymmetric cases or nonlinear class separability. Therefore, the CAIMAN approach has been further de-

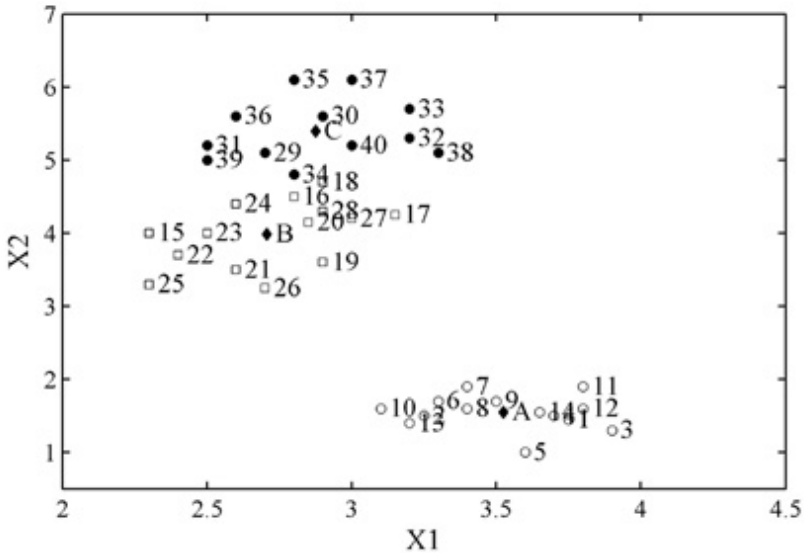

Figure 4.1: Simul1 data set: scatterplot of the two original variables, X1 and X2. Class centroids are shown with the corresponding class labels, A, B and C.

veloped defining a new mathematical concept called hyper-leverage. While the leverages extract information from the **X** descriptor space, the hyper-leverages are proposed to extract information from the space defined by the leverages themselves where relationships among classes are encoded. This involves a switch from the space of the p original variables to a space where each dimension is associated to one of the G classes.

Therefore, as the **X** matrix collects n objects each described by p variables, the matrix **K**, with dimensions n rows times G columns, is defined collecting the n objects each described by G leverages, previously calculated from the **X** matrix. Then, the sub-matrix $\mathbf{K}_g$ is derived by extracting from **K** the subset of n_g objects belonging to the g-th class. Data of the $\mathbf{K}_g$ matrix are mean centered.

Finally, for each i-th object, the hyper-leverage hh_{ig} for each g-th class

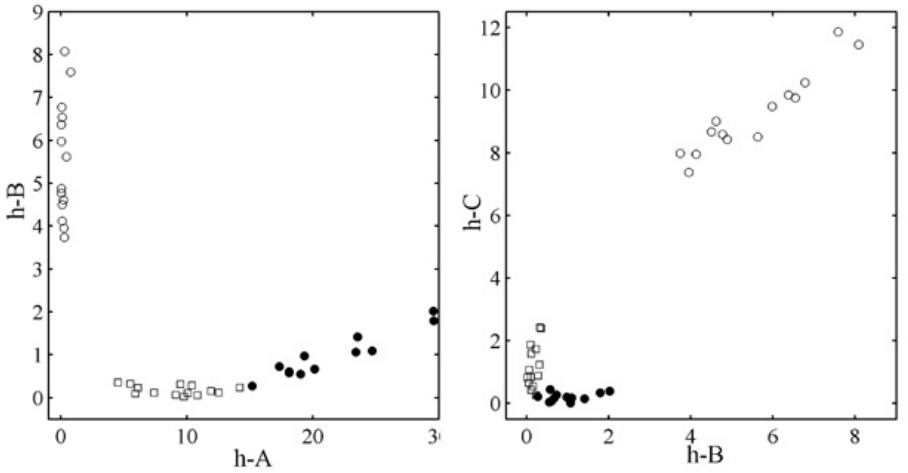

Figure 4.2: Simul1 data set: scatterplots of the leverages (leverage plots). Left: Simul1 objects in the space of the leverages derived from class A (h-A) and class B (h-B); right: scatterplot of Simul1 objects in the space of the leverages derived from class B (h-B) and class C (h-C).

is calculated as:

$$hh_{ig} = \left(\mathbf{h}_i - \bar{\mathbf{h}}_g\right)^{\mathrm{T}} \left(\mathbf{K}_g^{\mathrm{T}}\mathbf{K}_g\right)^{-1} \left(\mathbf{h}_i - \bar{\mathbf{h}}_g\right) \tag{4.13}$$

where$\bar{\mathbf{h}}_g$ is the class centroid in the leverage space. For the objects belonging to the g-th class, the leave-one-out estimate of the hyper-leverage is calculated as:

$$hh^*_{ig} = \frac{hh_{ig}}{1 - hh_{ig}} \tag{4.14}$$

In the X-space the leverage with respect to a class is calculated independently of the other classes. By projecting the objects in the H-space, the relationships among the classes become much more apparent since for each object all the leverages are simultaneously taken into account, which means simultaneously accounting for the relationships of the object with all the classes. The hyperleverage hh_{ig} is a simple tool to encode this information in a number. The hyperleverage is very useful for classify-

ing objects especially when there is a nonlinear class separability in the X-space, which becomes linear in the H-space.

Finally, combining information from leverages and hyperleverages, a final leverage score w is calculated for each object as:

$$w_{ig} = (1 - \alpha)\, h_{ig} + \alpha h h_{ig} \qquad 0 \leq \alpha \leq 1 \tag{4.15}$$

where the α trade-off parameter is optimised in order to maximise the Non-Error Rate: α values near zero indicate that a good class discrimination is directly obtained from the X-space; on the other hand, α values near one result from more complex class structures.

Then, a discriminant method, called Discriminant CAIMAN (D-CAIMAN), is proposed: it assigns each object to the class for which the leverage score is minimum:

$$i \rightarrow g \qquad \text{if} \qquad \min_g\{w_{ig}\} \tag{4.16}$$

To better explain the concept of hyper-leverage and its usefulness, an example is here reported. The Simul4 data set has been created. It consists of 32 objects, described by two variables (X1, X2) and equally distributed into two classes (n_A=16, n_B=16). As it can be observed in Figure 4.3, objects of class A are placed in such a way to form a squared frame around the class B objects, which constitute a small square whose center is a little displaced from the center of the space. This is a typical case of a class totally enclosed in another class. Looking at the leverage plot of Figure 4.4, it is evident that the minimum leverage classifier would fail in this case, since all the objects would be assigned to class A, the leverages derived from this class being smaller than those derived from class B for all the objects. In effect, class B objects are placed near the center of the class A space and, thus, are those objects with the smallest leverages derived from class A. In other words, class B objects are seen as the typical objects of class A. However, if one considers the leverages derived from class B, the two classes are clearly recognised, the leverages of class B objects being

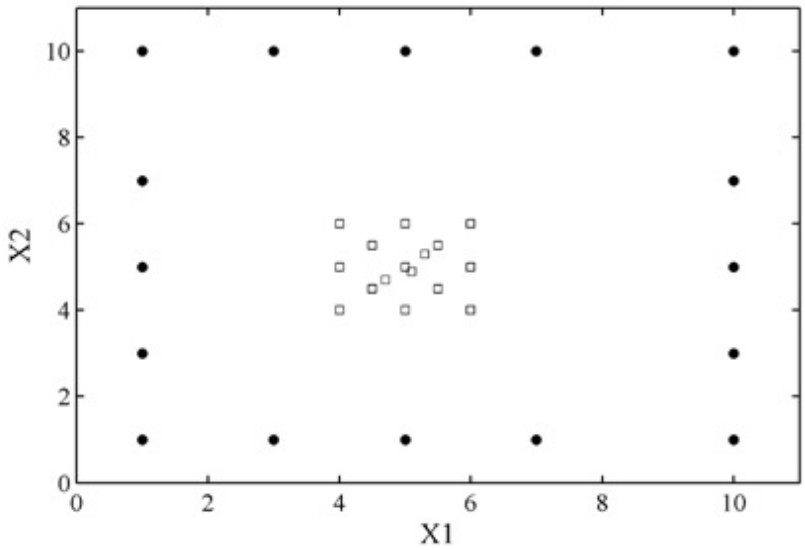

Figure 4.3: Simul4 data set: scatterplot of the two original variables, X1 and X2.

significantly smaller than those of class A objects. This means that from the class B point of view, its typical objects are recognised, while class A objects are correctly seen as outliers. Therefore, to solve this classification problem the information given by the leverages from class B needs to be accounted for along with the information provided by the leverages from class A. This is accomplished by the hyper-leverage calculation. Looking at the hyper-leverage plot of Figure 4.4, it can be seen that the hyper-leverage of the membership class is the smallest one for all the objects. This implies a correct classification.

It is also noteworthy that the nonlinear separability surface between classes in the original space has become linear in the space of the leverages. By the D-CAIMAN classifier, all the objects are always assigned to one of the defined classes. In order to face also on class modelling problems, in addition to the leverage/hyperleverage concepts, the set-up of two other parameters is required to define a class membership function.

Since the leverage score wig can be interpreted as a distance of the i-th

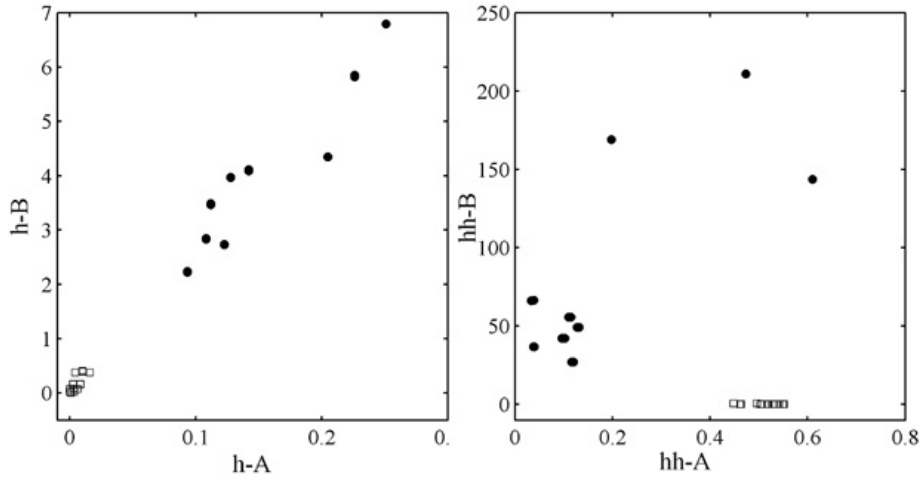

Figure 4.4: Simul1 data set: scatterplots of leverages (left) and hyper-leverages (right).

object from the g-th class, it is transformed into a similarity value by the following:

$$s_{ig} = \frac{1}{1 + w_{ig}} \qquad 0 \leq s_{ig} \leq 1 \tag{4.17}$$

where the similarity reaches the maximum value when the object is in the center of the class and decreases when the leverage score increases. Thus, for each object the similarities with respect to all classes are calculated and the first two highest values are taken into account (only two values are available when there are two classes). This means that only the two classes, which the object more likely belongs to, are considered. Based on their comparison, the object can be assigned to the class with the highest similarity or to none of the classes.

This is accomplished by a membership function M_i, defined as:

$$M_i = \tanh\left(\frac{s_{ig_1}}{s_{ig_2}} - 1\right) \qquad 0 \leq M_i \leq 1 \tag{4.18}$$

where s_{ig_1} is the greatest similarity and s_{ig_2} is the second one; the sub-

scripts g_1 and g_2 refer to the classes which the i-th object has the two highest similarities to. If the membership value is greater than a predefined threshold t_C (confusion threshold), the object is assigned to the class g1, otherwise it is not assigned at all because it is confused. If the object is assigned to the g1 class but it really does not belong the g1 class, the membership value is taken as negative (-M_i). Obviously, negative membership values may only be assigned to the training set objects.

Moreover, directly exploiting the leverage scale, an object is not assigned at all and rejected because it is an outlier with respect to all the classes, if the minimum value within its leverage vector is greater than a predefined threshold (rejection threshold):

$$i \rightarrow \text{out} \qquad \text{if} \qquad \min_g\{h_{ig}\} > t_R \tag{4.19}$$

Being M-CAIMAN a class modelling method, Non-Error Rate and Error Rate are no longer complementary quantities, but they sum up to 100% together with confused and not assigned objects. Thus, the best α tradeoff parameter is searched for maximising the product between the Non-Error Rate and 1 minus the Error Rate.

Therefore, the new proposed class modelling approach, called Modelling CAIMAN (M-CAIMAN), is based on a few rules, which simply allow to measure the degree of membership of an object to a class as well as to identify outliers and confused objects. Note that in this approach, each class model is not independent of the other classes (as, on the contrary, it happens in SIMCA) because the hyper-leverages used in the membership function encode information on the relationships among classes; therefore, M-CAIMAN is not an independent modelling technique.

4.2.4 The asymmetric case

The CAIMAN approach can be also used to face on the classification problems involving asymmetric cases. Asymmetric case means that there is only one class to be modelled and the aim is to discriminate this reference class from all the other objects not belonging to this class. This kind of

Table 4.3: Frequency table of an asymmetric classification case.

		assigned class	
		1	0
true	1	a	b
class	0	c	d

classification problem is frequent, for example, in food science to model food tipicity or in medical science to characterise a control set.

The Asymmetric CAIMAN (A-CAIMAN) classifier has been proposed to this end. It consists of calculating only the leverages with respect to the reference class. This means that there is a unique class model based on the objects belonging to the reference class (e.g. class A) and the position of all the objects with respect to this model space is evaluated by means of the leverages. Being modelled only the reference class, for the asymmetric case the hyperleverages are not calculated. Then, a threshold value for this leverage is taken so that it is possible to decide which objects belong or not to the reference class. Therefore, the A-CAIMAN results, as well as all the 2-class classification results, can be represented by a frequency table (Table 4.3). In Table 4.3 a are the true positive, i.e. the objects correctly included in class A; b are the false negative, i.e. the class A objects not included in the class; c are the false positive, i.e. the not-A objects included in class A; d are the true negative, i.e. the not-A objects not included in class A. Then, the class sensitivity (Sn) and specificity (Sp), their geometric mean (G_{mean}), the Jaccard's coefficient (Jc) and the Pearson's Φ coefficient are calculated as the following:

$$Sn = \frac{a}{a+b} \tag{4.20}$$

$$Sp = \frac{a}{a+c} \tag{4.21}$$

$$G_{mean} = \sqrt{Sn.\,Sp} \tag{4.22}$$

$$Jc = \frac{a}{a+b+c} \tag{4.23}$$

$$\Phi = \frac{a.\,d - d.\,b}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}} \tag{4.24}$$

The Jaccard's coefficient is one of the most known asymmetrical binary coefficients (i.e. where the comparison excludes the term d) likewise the Pearson's coefficient is one of the most known symmetrical binary coefficients. Depending on the actual classification problem, the optimal leverage threshold can be obtained by maximising one of the described parameters.

4.2.5 Software

Calculations of CAIMAN, LDA, QDA, UNEQ and NMC were performed with MATLAB 6.5 (Mathworks); the package SCAN (Minitab) was used to run KNN, CART and SIMCA. The in-house MATLAB modules for CAIMAN application are free and available at www.disat.unimib.it/chm. The code sources of these routines are described in the appendix.

4.3 Comparison with other classifiers

In order to check the performance of the proposed CAIMAN strategy and make extended comparisons with other classification methods (LDA, QDA, UNEQ, KNN, CART, SIMCA), 25 data sets have been considered and all the predictive performances of the listed classification methods have been tested on these data.

4.3.1 Data sets

The main characteristics of these data sets are collected in Table 4.4, while their bibliographic references are collected in Table 4.5. Two data sets have been used twice: the data set Sulfa has been used both with original variables and with principal components; the data sets Vinagres has been

used by considering both 4 and 2 principal components. The original data set VegOil consists of seven classes; however, it has been modified here because of the small number of class objects for some classes. Specifically, only 83 objects instead of 132 have been retained, by excluded the three smallest classes. For the Fish data set, the class with 3 objects has been excluded from the class model as represented by too few objects. For the data set Tobacco, the original 15 variables have been arbitrarily reduced to 6: petroleumether, nicotine, total sugar, total nitrogen, protein and ash. For the data set Ruspini, the three objects 47, 50 and 53, near to the second class but not so far from the third class, have been used as test objects. The data set Simul4 has been created in order to stress a nonlinear class separability problem (Figure 4.3), already present in the data set Perpot. As explained above, Simul4 is constituted by 16 data in a very thin squared frame and 16 data in a small square near to the center of the frame. For some data sets, the principal components are used instead of the original variables to avoid too bad ratios between class objects and variables. When principal components are used instead of the original variables, the number of PCs is reported together with the number of variables (e.g. 5/1 means that 1 PC has been used instead of the 5 original variables). Column 7 (worst n/p) represents the worst ratio of the objects of the smallest class over the number of variables. Several authors suggest to model only classes with a ratio greater than 3; however, some cases with lower ratio have been also tried in this work.

4.3.2 Results of comparison

The D-CAIMAN and M-CAIMAN classifiers have been evaluated by using the 27 data sets previously quoted; both rejection and confusion thresholds have been empirically determined. A value of 2.3 has been used for t_R and a value of 0.015 has been used for t_C.

In order to check the CAIMAN predictive performances, the leave-more-out (LMO) technique has been used, minimising the error rate. Specifically, for each α value (between 0 and 1, with step 0.1), 300 it-

Table 4.4: Data sets used for comparison.

ID	Data	Variables	Objects	Classes	Worst n/p
1	Iris	4	150	3	12.5
2	Wines	13	178	3	3.7
3	Camel-1PC	5/1	20	2	9
4	Sunflowers-5PC	21/5	70	2	5.2
5	Sulfa	7	50	2	2
6	Sulfa-3PC	7/3	50	2	4.7
7	Hirsute	7	133	2	3.7
8	Metacycline	4	22	2	2.5
9	Digits	7	477	10	5.4
10	Olitos-5PC	21/5	120	4	2.2
11	Cheese-4PC	4	134	4	4.8
12	Perpot	2	100	2	25
13	CrudeOil-3PC	5/3	56	3	2.3
14	Hemophilia	2	75	2	15
15	School	2	85	3	13
16	Membrane	2	36+11	3	6
17	Thiophene	3	24	3	2.7
18	Vinagres-4PC	20/4	66	3	2
19	Vinagres-2PC	20/2	66	3	4
20	Sand	2	81	2	17
21	VegOil4-2PC	7/2	83	4	5
22	Bank	4	46	2	5.3
23	Fish-3PC	10/3	27+3	2	4.3
24	Coffee-2PC	13/2	43	2	3.5
25	Tobacco	6	26	2	2.2
26	Ruspini	2	72+3	4	7.5
27	Simul4	2	32	2	8

erations excluding 20% of objects for each class have been performed. The leave-more-out validation has been repeated 20 times for all the data sets, both for D-CAIMAN and M-CAIMAN.

For each data set, the optimal α value has been chosen in correspondence of the minimum average error rate over the 20 repetitions; this parameter gives information about the trade-off between the leverage and hyper-leverage role in the final classification rule. A zero value indicates that only leverage is used for classification, while a value of one indicates that only hyper-leverage is used. Moreover, in order to compare the results

Table 4.5: References to the data sets used for comparison.

ID	Data	References
1	Iris	Fisher (1936)
2	Wines	Forina *et al.* (1986)
3	Camel-1PC	Mattiello *et al.* (1993)
4	Sunflowers-5PC	Saviozzi *et al.* (1986)
5	Sulfa	Miyashita *et al.* (1986)
6	Sulfa-3PC	Miyashita *et al.* (1986)
7	Hirsute	Armanino *et al.* (1989b)
8	Metacycline	A.P. Worth and Cronin (1999)
9	Digits	Todeschini (1990)
10	Olitos-5PC	Armanino *et al.* (1989a)
11	Cheese-4PC	Resmini *et al.* (1986)
12	Perpot	Forina (1990))
13	CrudeOil-3PC	Gerrild and Lantz (1990)
14	Hemophilia	CompStat (1974)
15	School	Johnson and Wichern (1992)
16	Membrane	Mager (1991)
17	Thiophene	Mager (1991)
18	Vinagres-4PC	Benito *et al.* (1999)
19	Vinagres-2PC	Benito *et al.* (1999)
20	Sand	Hamilton (2001)
21	VegOil4-2PC	Brodnjak-Voncina *et al.* (2005)
22	Bank	Johnson and Wichern (1992)
23	Fish-3PC	Forina *et al.* (1982)
24	Coffee-2PC	Streuli (1987)
25	Tobacco	Forina (1990)
26	Ruspini	Kaufman and Rousseau (1990)
27	Simul4	Artificial dataset

with other classification methods, all the calculations have been further performed with the leave-one-out validation technique, using the a values obtained by the leave-more-out technique.

In Table 4.6, Table 4.7, α values and Error Rates obtained in leave-more-out and leave-one-out are collected both for D-CAIMAN and M-CAIMAN, respectively; moreover, percentages of confused and rejected objects are reported for M-CAIMAN (Table 4.8). In the last row of these tables the average values of the parameters are given. As it can be seen, the error rate differences between LMO and LOO validation techniques

Table 4.6: Results obtained by the D-CAIMAN approach. LMO and LOO indicate the leave-more-out and leave-one-out validation techniques, respectively.

ID	Data	α	ER(LMO)	ER(LOO)
1	Iris	0.9	2.50	2.00
2	Wines	1	1.90	1.10
3	Camel-1PC	0	16.00	15.00
4	Sunflowers-5PC	0.9	9.20	11.40
5	Sulfa	0	30.00	28.00
6	Sulfa-3PC	1	26.60	20.00
7	Hirsute	0	19.20	19.60
8	Metacycline	0	20.60	18.20
9	Digits	0	30.00	29.00
10	Olitos-5PC	1	14.30	12.50
11	Cheese-4PC	0	19.00	15.70
12	Perpot	1	4.10	3.00
13	CrudeOil-3PC	0.7	11.00	7.10
14	Hemophilia	0.2	14.10	13.30
15	School	0.5	6.90	5.90
16	Membrane	0	11.30	5.60
17	Thiophene	0	24.90	20.80
18	Vinagres-4PC	0.9	3.00	0.00
19	Vinagres-2PC	0.2	5.10	4.60
20	Sand	0.6	6.30	6.20
21	VegOil4-2PC	0	8.70	7.20
22	Bank	0.6	11.80	10.90
23	Fish-3PC	0	9.60	3.70
24	Coffee-2PC	0.1	3.20	2.30
25	Tobacco	1	19.70	11.50
26	Ruspini	0	0.00	0.00
27	Simul4	1	6.30	6.30
	Average		12.40	10.40

are 2% and 0.9% for D-CAIMAN and M-CAIMAN, respectively; these results show the good predictive ability of CAIMAN approach.

The percentages of confused objects is almost constant for both validation techniques, while the percentages of rejected objects differ significantly for the data sets Metacycline (no.8), Thiophene (no.17), Vinagres-4PC (no.18) and Tobacco (no.25). For all these data sets the worst object/variable ratios Table 4.4 are lower than 3; these results can be prob-

Table 4.7: Results obtained by the M-CAIMAN approach. LMO and LOO indicate the leave-more-out and leave-one-out validation techniques, respectively.

ID	Data	α	ER(LMO)	ER(LOO)
1	Iris	0.7	2.10	2.00
2	Wines	1	1.30	1.10
3	Camel-1PC	0	15.80	15.00
4	Sunflowers-5PC	0.9	7.50	8.60
5	Sulfa	1	22.00	20.00
6	Sulfa-3PC	1	20.00	14.00
7	Hirsute	1	9.70	7.50
8	Metacycline	1	16.80	22.70
9	Digits	0	28.20	27.40
10	Olitos-5PC	1	12.70	10.00
11	Cheese-4PC	0.2	15.80	14.90
12	Perpot	1	3.70	3.00
13	CrudeOil-3PC	0.7	10.60	7.10
14	Hemophilia	0.1	11.50	9.30
15	School	0.5	6.30	5.90
16	Membrane	0	8.30	5.60
17	Thiophene	0	21.50	20.80
18	Vinagres-4PC	0.5	1.30	0.00
19	Vinagres-2PC	0	4.50	3.00
20	Sand	1	6.00	6.20
21	VegOil4-2PC	0	7.80	6.00
22	Bank	0	10.90	10.90
23	Fish-3PC	0	8.60	3.70
24	Coffee-2PC	0.1	2.60	2.30
25	Tobacco	1	8.40	11.50
26	Ruspini	0	0.00	0.00
27	Simul4	1	5.80	6.30
	Average		10.00	9.10

ably due to the fact that, when more then one object is left out from the training set, some classes are poorly represented.

In the case of M-CAIMAN, three independent quantities can be evaluated for the optimisation, i.e. error rate, confused and rejected objects. For this reason, in some cases, alternative solutions (α values) could be also considered. For example, evaluating the LMO results, for the data set Hirsute (No. 7) an α value equal to 0.7 instead of 1 gives ER%=19.2,

Table 4.8: Results obtained by the M-CAIMAN approach. C and R are the percentages of confused and rejected objects, respectively. LMO and LOO indicate the leave-more-out and leave-one-out validation techniques, respectively.

ID	Data	C(LMO)	C(LOO)	R(LMO)	R(LOO)
1	Iris	1.10	2.00	0.00	0.00
2	Wines	1.50	1.70	0.00	0.00
3	Camel-1PC	0.10	0.00	0.00	0.00
4	Sunflowers-5PC	3.50	2.90	0.00	0.00
5	Sulfa	13.80	18.00	2.70	2.00
6	Sulfa-3PC	14.50	14.00	0.00	0.00
7	Hirsute	15.30	18.10	0.00	0.00
8	Metacycline	3.10	0.00	12.50	4.60
9	Digits	2.90	3.00	0.00	0.00
10	Olitos-5PC	2.90	3.30	0.00	0.00
11	Cheese-4PC	5.10	6.70	1.10	0.00
12	Perpot	0.80	0.00	0.00	0.00
13	CrudeOil-3PC	0.10	1.80	0.00	0.00
14	Hemophilia	4.80	6.70	0.00	0.00
15	School	1.30	2.40	0.00	0.00
16	Membrane	0.90	2.80	8.20	0.00
17	Thiophene	1.30	0.00	11.40	0.00
18	Vinagres-4PC	0.10	0.00	6.30	1.50
19	Vinagres-2PC	1.30	0.00	0.30	0.00
20	Sand	0.50	0.00	0.00	0.00
21	VegOil4-2PC	2.00	3.60	0.00	0.00
22	Bank	2.30	0.00	2.80	0.00
23	Fish-3PC	1.30	0.00	0.90	0.00
24	Coffee-2PC	0.20	0.00	0.60	0.00
25	Tobacco	0.80	0.00	32.50	7.70
26	Ruspini	0.00	0.00	0.00	0.00
27	Simul4	0.50	0.00	0.00	0.00
	Average	3.00	3.20	2.90	0.60

C%=0 and R%=0; for the data set Metacycline (no.8) an α value equal to 0 instead of 1 gives ER%=18.5, C%=0.3 and R%=12.9; for the data set Bank (no.22) an α value equal to 1 instead of 0.3 gives ER%=7.7, C%=25.3 and R%=2.8.

Further considerations can be made by considering the CAIMAN results obtained by the leave-one-out technique. The two CAIMAN ap-

proaches give the same results if no objects are rejected or considered as confused. Only for 4 data sets some objects have been found as outliers (i.e. rejected) and for 5 data sets more than 5% of confused objects have been found. For these data sets, in most of the cases, M-CAIMAN gives the best ER values, being able to not classify objects lying in fuzzy regions, which are often badly classified. In particular, significant percentages (more than 10%) of confused objects have been found for the data sets Sulfa (no.5), Sulfa-3PC (no.6) and Hirsute (no.7) (18%, 14.0% and 18.1%, respectively) for which the difficulties of classification are well known [Kaufman and Rousseau (1990)]. A significant number of outliers have been found only for Tobacco data set (7.7%).

Extended comparisons have been performed using the 27 data sets modelled by the following classification methods: D-CAIMAN, M-CAIMAN, LDA, QDA, UNEQ, KNN, CART, and SIMCA. For the UNEQ method, a confidence region of 95% has been chosen; both the optimal K value for KNN method (K=1,..,10) and the optimal tree in CART method have been searched for by validation. In Table 4.9, values of the leave-one-out error rates (ER%) are collected for all the other classification methods. In the last row of the table, the average values of ER% have been also calculated for each method. By considering the average ER values obtained with the classical methods and the ones obtained with CAIMAN (Table 4.6 and Table 4.7), M-CAIMAN and D-CAIMAN give again the best results on the whole set of data; good performances are also obtained by QDA and KNN. Therefore, the new proposed approach to classification problems seems to show several advantages. First of all, it shows-on an average basis-good performance when compared to the other popular methods. The asymmetric CAIMAN is able to solve in a very simple and intuitive way classification problems related to tipicity, pathologies, and single class characterisation. Like QDA, UNEQ, KNN, SIMCA and CART, the CAIMAN method seems not to suffer of nonlinear class separability. The a priori probability based on the object frequency is implicitly assumed in the leverage algorithm and a good representation of the classes is necessary, i.e. high enough objects/variables ratios (greater than 2-3).

Table 4.9: Comparisons of leave-one-out Error Rates (ER%) obtained for the 27 data sets by the compared classification methods.

ID	Data	LDA	QDA	UNEQ	KNN	CART	SIMCA
1	Iris	2.00	2.70	2.00	3.30	4.70	10.70
2	Wines	1.10	0.60	1.70	23.00	11.20	5.60
3	Camel-1PC	15.00	15.00	15.00	15.00	40.00	20.00
4	Sunflowers-5PC	5.70	8.60	7.10	7.10	4.30	14.30
5	Sulfa	36.00	22.00	28.00	22.00	12.00	28.00
6	Sulfa-3PC	36.00	24.00	22.00	10.00	14.00	24.00
7	Hirsute	22.60	13.50	15.80	16.50	15.80	14.30
8	Metacycline	50.00	18.20	22.70	22.70	13.60	18.20
9	Digits	26.20	31.20	24.20	28.20	28.60	27.60
10	Olitos-5PC	7.50	8.30	5.80	9.20	15.80	35.00
11	Cheese-4PC	22.40	18.70	22.40	19.40	23.90	17.90
12	Perpot	15.00	7.00	11.00	10.00	3.00	15.00
13	CrudeOil-3PC	10.70	14.30	19.60	10.70	23.90	23.20
14	Hemophilia	14.70	16.00	14.70	16.00	18.70	14.70
15	School	9.40	4.70	5.90	34.10	8.20	8.20
16	Membrane	11.10	5.60	8.30	2.80	8.30	25.00
17	Thiophene	20.80	29.20	25.00	16.70	29.20	25.00
18	Vinagres-4PC	1.50	0.00	0.00	0.00	6.10	1.50
19	Vinagres-2PC	6.10	4.50	6.10	6.10	6.10	7.60
20	Sand	6.20	6.20	6.20	11.10	12.30	42.00
21	VegOil4-2PC	9.60	9.60	8.40	8.40	14.50	7.20
22	Bank	13.00	10.90	10.90	10.90	17.40	28.30
23	Fish-3PC	11.10	3.70	7.40	7.40	14.80	3.70
24	Coffee-2PC	0.00	2.30	2.30	0.00	6.70	2.30
25	Tobacco	19.20	19.20	15.40	7.70	3.90	23.10
26	Ruspini	0.00	0.00	0.00	0.00	0.00	11.10
27	Simul4	71.90	0.00	46.90	0.00	6.30	43.80
	Averages	16.50	11.00	13.10	11.80	13.50	18.40

An extended application of these methodologies to food data is presented in chapter 8.

chapter 5

Similarity measure for sequential data

5.1 Introduction

The concept of similarity and its dual concept of diversity play a fundamental role in several chemometric strategies, as well as in relatively new fields such as genomics and proteomics. Several distance measures both for quantitative and for binary variables have been defined, such as, for example, Euclidean, Manhattan, Minlowski, and Camberra distances for quantitative variables and Hamming, Tanimoto, and Jaccard distances for binary variables. Distances are the quantitative measure of diversity between a pair of objects; thus, large distances indicate large diversity (small similarity), and small distances indicate small diversity (large similarity).

A new method dealing with the fingerprint analysis of sequential data is proposed. This method involves the Hasse partial ordering approach [Halfon and Reggiani (1986), Bruggemann and Bartel (1999), Pavan and Todeschini (2004)]: the Hasse matrix can be associated to each sequence and then the similarity between two sequences can be evaluated with the

definition of a distance between the corresponding Hasse matrices. These distances have some useful properties and seem to show a high sensitivity to changes in the sequence structure. The theory of the partial ordering is presented together with the proposed distance between Hasse matrices.

5.2 Partial Ordering and Hasse matrices

Partial ordering (PO) is an approach to the ranking where the relationship of "incomparability" is added to the classical relationships of "greater than or equal to" [Halfon and Reggiani (1986), Bruggemann and Bartel (1999), Pavan and Todeschini (2004)]. Given a set **Q** of n elements, each described by a vector **x** of p variables (attributes), the two elements s and t belonging to **Q** can be compared with the following:

$$\mathrm{t} \geq \mathrm{s} \qquad \Leftrightarrow \qquad x_j(t) \geq x_j(s) \qquad \forall j \tag{5.1}$$

The ordering relationships between all the pairs of elements are collected into the Hasse matrix **M**; for each pair of elements s and t, the entry m_{st} of this matrix is:

$$m_{st} = \begin{cases} 1 & \text{if} \quad x_j(s) \geq x_j(t) \quad \forall j = 1, \cdots, p \\ -1 & \text{if} \quad x_j(s) < x_j(t) \quad \forall j = 1, \cdots, p \\ 0 & \text{otherwise} \qquad\qquad \forall j = 1, \cdots, p \end{cases} \tag{5.2}$$

If the entry s-t (m_{st}) contains +1, the entry t-s (m_{ts}) contains -1; if the entry s-t contains 0, also the entry t-s contains 0. Then, if pairs of equal elements are not present, the Hasse matrix is a squared antisymmetric matrix, whose elements take only the values 0, +1 and -1. It is interesting to observe that the Hasse matrix contains an holistic view of all the ordering relationships among the n elements belonging to the set Q. In other words, **M** can be assumed as a fingerprint of the ordering relationships among the n elements.

In order to add more information to the Hasse matrix, the augmented Hasse matrix can be defined by adding into the main diagonal (which

have zero values in the Hasse matrix) any property P of the elements. The property values of each set of n elements can be scaled dividing each value by the maximum property value:

$$m_{ii} = \frac{P_i}{\max(P)} \tag{5.3}$$

5.3 Hasse distances

Be $\mathbf{M}^A$ and $\mathbf{M}^B$ two Hasse matrices with dimensions n times n, obtained by two different retaliations of the variables defining n elements, i.e. representing two partial orderings A and B. The distance between the two partial orderings can be obtained by summing up the differences between the corresponding matrix elements. The distance between A and B can be formulated as the contribution of two terms:

$$d_D(\mathrm{A},\mathrm{B}) = \frac{\sum_{i=1}^{n} |m_{ii}^{\mathrm{A}} - m_{ii}^{\mathrm{B}}|}{n} \tag{5.4}$$

$$d_H(\mathrm{A},\mathrm{B}) = 2\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |m_{ij}^{\mathrm{A}} - m_{ij}^{\mathrm{B}}|}{n(n-1)} \tag{5.5}$$

where the first term d_D is the contribution to the distance due to the diagonal terms (the property values), while the second term d_H is the contribution to the distance due to the off-diagonal terms (the ranking relationships of the Hasse matrix). In both cases, the two distance terms d can range from 0 to 1. This is obvious for the diagonal contribution using scaled values, but not for the off-diagonal contribution. Assuming that only two variables are considered in building the Hasse matrix and the first variable is a totally ordered sequence, the Hasse matrix obtained by using as the second variable another total ordered sequence is constituted only by +1 and -1 values, meaning that a total ranking of the elements exists. If the Hasse matrix is obtained by using as the second variable an ordered sequence, which is inverse with respect to the first one, it will

be constituted only by zero values, meaning that no ordering relationships exist among the elements. Then, it is noticeable that the maximum theoretical distance between these two matrices is $n(n\text{-}1)$.

From the two contributions, a weighted standardised Hasse distance (WSHD, d_W) can be defined as a trade-off between the ranking relationships and the property values:

$$d_W(\mathrm{A},\mathrm{B}) = (1-w).\, d_H(\mathrm{A},\mathrm{B}) + w.\, d_D(\mathrm{A},\mathrm{B}) \tag{5.6}$$

where w is a weighting term ranging between 0 and 1, and the defined distance d_W ranges also between 0 and 1. Using a weight equal to zero, the distance is calculated taking into account only the ranking relationships, while a weight equal to 1 takes into account only the property values. A weight equal to 0.5 takes equally into account both terms, and provides a distance measure where both the ordering relationships among the elements and their property differences are equally considered.

Moreover, WSHD is a Manhattan distance calculated on the corresponding pairs of elements of two Hasse matrices, thus preserving all the metric properties of the Manhattan distance. This distance is straightforwardly interpretable measure of distance (or as percentage d) and as an absolute measure of similarity (s) after the transformation:

$$s = 1 - d_W \tag{5.7}$$

or as a correlation measure r_W after the transformation:

$$r_W = 2(1-d_W) - 1 \qquad -1 \leq r_W \leq +1 \tag{5.8}$$

where on the contrary of the Spearman rank correlation, this correlation takes into account also the presence of incomparabilities. The rank correlation r_W calculated for $w = 0$ (i.e.$d_W = d_H$) coincides with the

Greiner-Kendall rank correlation index (τ), defined as:

$$\tau = \frac{4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^{+}}{n(n-1)} \qquad -1 \leq \tau \leq +1 \tag{5.9}$$

where d_{ij}^{+} is defined as:

$$d_{ij}^{+} = \begin{cases} 1 & \text{if} \quad i < j \quad \text{and} \quad p_i < p_j \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

and p is the number of ranks of the samples. Therefore, the Spearman rank correlation uses more information with respect to the Greiner-Kendall rank correlation: the Spearman index is more suitable if no information has to be discarded while the Greiner-Kendall index is a more robust statistical index.

5.4 Applications on sequential data

Data including an ordering variable can be considered as sequential data. Then sequential data can be characterised by an ordering variable (sequential integer numbers, variable X1) and a property variable (real numbers, variable X2). For example, the intensities of signals (property variable) obtained by mass spectrometry are ordered by increasing masses (their positions constitute the ordering variable); the intensities of IR/UV signals (property variable) are ordered by wave lengths (ordering variable); the intensities of 1DNMRspectra (property variable) are ordered by the chemical shifts (ordering variable); in general, all the spectra achieved along time are intrinsically ordered. Analogously, data based on natural sequences can be also considered as sequential data. In fact, a sequence of integer numbers associated to the elements of the sequence can be used as ordering variable, while a defined property of the elements of sequence, such as the position of a sentence characters in the alphabet (i.e. a=1,..., z = 26), the molecular weights of the four nucleic bases of DNA sequences, some physical-chemical property of the 20 aminoacids of protein sequences,

the most relevant protein abundances of proteomic maps, can be used as property variable.

This kind of data can be easily characterised by Hasse matrices and their similarity/diversity assessed by the previously defined Hasse distance. Considering the presence of incomparabilities, the maximum information content can be obtained by using only two variables, i.e. the ordering variable (X1) and the property variable (X2). In fact, in this case, the incomparabilities between two samples s and t can be due to only one condition, i.e. when the two variables X1 and X2 show an opposite rank:

$$\mathrm{X1(s)} > \mathrm{X1(t)} \quad \textit{and} \quad \mathrm{X2(s)} < \mathrm{X2(t)} \tag{5.11}$$

$$\mathrm{X1(s)} < \mathrm{X1(t)} \quad \textit{and} \quad \mathrm{X2(s)} > \mathrm{X2(t)} \tag{5.12}$$

On the contrary, if three variables are taken into account, the incomparabilities can be obtained by opposite ranks of X1-X2 or X1-X3 or X2-X3, with a loss of information. In fact, in this case, the presence of zero values in the Hasse matrix cannot be univocally related to a specific relationship. All the achieved results [Todeschini *et al.* (2006), Todeschini *et al.* (2007), Ballabio *et al.* (2006b)] suggest that the proposed approach, based on the partial ordering technique and the Hasse distance analysis, allows the characterisation of sequential data.

Moreover, the proposed approach shows some advantages: (**a**) it seems able to link each electronic nose time profile to a meaningful mathematical term (the Hasse matrix), which can be consequently treated and studied by multivariate analysis; (**b**) the Hasse matrices and the corresponding distances are calculated with a simple algorithm; (**c**) the Hasse distance is standardised, allowing a natural interpretation of the results; (**d**) the distances consider the whole time profile, i.e. no parameterisation is needed and the time information is preserved; (**e**) the distances can be obtained by a flexible strategy (the weights) depending on the aim of the analysis. An extended application of these methodologies to

electronic nose data of food products is presented in chapter 9, while the MATLAB modules for calculating the Hasse distances are free and available at www.disat.unimib.it/chm. The code sources of these routines are described in the appendix.

chapter 6

Novel reduction of sequential data dimension

6.1 Introduction

Pattern recognition methods for classification and identification are increasingly used in several fields, such as food chemistry, process monitoring, medical sciences, pharmaceutical chemistry, and social and economic sciences. Classification is one of the fundamental methodologies in chemometrics and consists basically in finding a mathematical model able to recognise the membership of each object to its proper class. Once a classification model has been obtained, the membership of new objects to one of the defined classes can be predicted. Based on different mathematical approaches, several multivariate classification methods have been proposed. However, while the theory for multivariate quantitative calibration is extensive and has been developed over the last thirty years [Martens and Naes (1989)], multivariate classification can be furtherly explored.

Variable selection techniques can be used in order to improve classification models, as well as regression models. Unfortunately variable

selection techniques are in risk of overfitting when a huge number of variables is used especially when the number of samples is limited [Leardi *et al.* (1992),Leardi (2000)]. In order to circumvent this overfitting problem, a new approach based on local data reduction and subsequent variable selection is presented here. In this approach, windows of the original data are compressed using PCA and only significant components are retained. The subsequent variable-selection is then performed on these locally derived score variables rather than the original data.

6.2 Extraction of local scores

As explained, variable selection methods, such as Forward Selection and Genetic Algorithms, can not handle a huge number of variables. Consequently, a reduction of the data dimension is needed, especially when dealing with spectral data. To do so, a new algorithm for the extraction of significant features from high dimensional data is proposed. First the data is reduced using Principal Component Analysis (PCA) for the extraction of local useful information from specific windows of the original spectral profile. The obtained (reduced) dataset can then be easily analysed by means of variable selection techniques coupled with classification or regression methods. Given a data matrix $\mathbf{X}$ with n rows (the samples) and p columns (the variables), the compression approach can be described with the following steps:

1. partition of the entire chromatographic profile in w windows (where the number of the windows can be defined by the user): each window matrix $\mathbf{X_w}$ has dimension n times p/w

2. data pretreatment and application of Principal Component Analysis on each matrix $\mathbf{X_w}$ separately

3. screening of the obtained scores: only the scores with Explained Variance higher then a specified threshold (EV_{thr}) and Squared Sum of Residuals higher then a specified threshold (SSR_{thr}) are retained

4. collection of all the retained scores in the final reduced data matrix $\mathbf{X_r}$, with dimensions n times r, where r is the total number of retained scores.

Fundamentally, this approach permits to collect the local useful information represented by the retained scores in a unique matrix ($\mathbf{X_r}$), while windows without information are removed. In collinear high dimensional data, this solution can perform a significant dimension reduction in the data.

6.3 Coupling with other multivariate approaches

The final reduced data matrix $\mathbf{X_r}$ could represent a starting point for several chemometric applications: in fact, in this matrix, both a data dimension reduction and the presence of local fingerprints can be preserved. Consequently, both a direct regression or classification model can be performed, while variable selection techniques could be also applied in order to better retain the relevant scores, on the basis of the aim of the research.

On the other hand, in order to avoid final models based only on local principal components, the multivariate approaches could be also applied on the original data using the primarily selected windows, i.e. using all the windows where at least one score has been extracted. All this kinds of application are better explained in chapter 13, with examples of real data.

part II

Applications

---

chapter 7

List of applications

7.1 Introduction

In the second part of this thesis, applications of new and classical chemometric methods on several different food data are presented. The Classification And Influence Matrix Analysis (CAIMAN), extensively explained in chapter 4, the similarity/diversity measure based on Hasse distances (chapter 5), the reduction of data dimension of sequential data (chapter 6) and classical chemometric methods (chapter 3) have been used for several applications on food data. Some applications have been grouped on the basis of the applied analytical/chemometric methods and are extensively explained in the next chapters. In Table 7.1 a synthetic list of all the chemometric applications on food data explored during the PhD thesis is presented, with their relative references.

7.2 Brief explanation of some applications

All the applications not treated in the next chapters are just listed in the following paragraphs with a brief resume and the relative bibliographic

Table 7.1: Synthetic list of the chemometric applications on food data explored in this PhD thesis. Bibliographic references, chemometric methods, analytical data and analysed food matrices are reported. ES refers to Electronic Sensors, AM to acoustic-mechanical signals, TI to Time Intensity signals, GC to Gas Chromatography, MS to Mass Spectrometry. In the first column (chap) the relative chapter and paragraphs are reported.

chap	analytical	food mat.	chemometric	reference
8	ES	oil, wine	CAIMAN	Ballabio *et al.* (2006c)
10	ES	oil	CP-ANN	Cosio *et al.* (2006)
10	ES	oil	PCA, DA	Cosio *et al.* (2007)
9	ES	oil	PCA, Hasse	Ballabio *et al.* (2006b)
7.2.1	ES	wine	GA, OLS	Buratti *et al.* (2006)
13	GC	wine	GA, DA	in preparation
7.2.2	GCMS	meat	PCA	Carcione *et al.* (2006)
7.2.4	chemical	beverage	Ranking	Ballabio *et al.* (2004)
11	AM	bread	PCA, DA	Piazza *et al.* (2006)
7.2.3	GC	spirit	CP-ANN	Ballabio *et al.* (2006a)
12	TI	meat	PARAFAC	Reinbach *et al.* (2006)

references.

7.2.1 Prediction of wine sensorial descriptors by means of Genetic Algorithms

It is well known that the sensory characteristics of wine are influenced by a broad spectrum of factors such as type of grape, soil, enological and climatic conditions. Generally the sensory analysis, based on the trained expert panelists, is useful for wine classification and quality control but it is of high-cost, time-consuming and sometimes without any objective estimation. In the last decades many efforts have been done to establish a relationship between sensory attributes and chemical composition of wine in order to understand which components influence the sensory properties and the final quality of the product. These works are based on traditional chemical measurements and on spectrophotometric or chromatographic determinations coupled with classical chemometric techniques such as Principal Component Analysis and Partial Least Square

analysis [Aznar *et al.* (2003), Boselli *et al.* (2004)]. However, these analytical schemes cannot fully replace sensorial examinations and often they are time consuming and require skilled personnel. It is, therefore, of great interest the development of low cost, rapid and non-destructive analytical procedures able to quantify the sensorial descriptors and the overall quality of wines.

In this application, innovative, rapid and objective analytical techniques such as the electronic nose (e-nose) and the electronic tongue (e-tongue) coupled with spectrophotometric methods were used for objective sensorial evaluation of wine [Buratti *et al.* (2006)]. Signals from the instruments were used to build predictive models of sensorial descriptors by means of the Genetic Algorithm. Genetic Algorithms were proposed as an alternative to the mostly used PLS analysis and were employed to select subsets of variables that maximise the predictive power of regression models. The regression models selected were subsequently validated with the bootstrap procedure, that provides a more certain evaluation of their predictive capability while the Williams plots was used to check the presence of outliers.

The results obtained make sure on the possibility to use these innovative techniques in order to describe and predict a large part of the selected sensorial information. The proposed analytical methods have the advantage of being rapid and objective, furthermore the statistical methods applied could be considered a rational operative procedure to build regression models with real predictive capability.

7.2.2 Fatty acid composition as markers of feeding traceabilty

A very important problem in meat inspection is the control of the correspondence between labelling and quality of specimens in the market. In poultry meat products, for example, a significant difference exists between cheaper subjects fed with traditional diet, containing animal fat, and those fed with fully vegetable diet. Several studies confirmed the influence of the

diet on meat composition: a relationship between the fatty acid content of experimental diets and the one of the resulting meat were often established. In this connection total lipids extraction and fatty acid methyl esters (FAME), determination by gas-chromatography (GC) is today a generally used technique in inspection laboratories.

Capons' meat, even if less than broilers, has a relevant place in human nutrition, addressed to leans'. In this application, it has been tested if FAME GC analysis is enough to distinguish from breast muscle specimens coming from animals fed with traditional mixed diet (MD), containing animal and vegetable fats, and those coming from animals fed with completely vegetable diet (VD). The considered diets were not prepared for experimental studies, but really used in a farm to obtain products for sale. For a better comprehension of the eventual significant differences detailed source, separation of the main lipid fractions and chemometric elaboration of the large amount of data were performed [Carcione *et al.* (2006)]. Hence the final stated question is if a previous chromatographic separation, before routine FAME determination, is needed to test feeding traceability in capon breast meat.

7.2.3 Characterisation of Zivania by means of Counter-propagation Artificial Neural Networks

Zivania is a beverage that has been produced for centuries in Cyprus by distillation of grape marc. Quality control is ensured by the Vine Products Industry Rules and Regulations of 1998 for the control of Zivania. As a result of the above, as well as of the increase in the consumption of alcoholic beverages in Cyprus, the quality of Zivania has been improved and it is now produced on an industrial basis, gaining popularity elsewhere in Europe.

There are of course similar products from other countries, which derive their name from the term used to describe the grape marc (the refuse of pressed grapes). Therefore the name Zivania in Cyprus comes from the word "Zivana" meaning grape marc. The authentic taste and aroma

of Zivania is due to its unique chemical composition, which is ensured by the volatile compounds that are transferred from the grape residue during distillation. It is believed that the differentiation between Zivania and other similar alcoholic beverages is related to the unique geological and climatic conditions existing on the island of Cyprus, the methods of production and distillation and the type of local varieties of grapes used in its manufacture.

During recent years there has been increasing interest in the certification of the geographical origin of food products, since authenticity and quality issues can be often associated with a given geographical origin. Therefore, several efforts have been made to authenticate the origin of Zivania by means of chemometric analysis, considering different chemical parameters: in previous publications [Kokkinofta *et al.* (2003), Kokkinofta and Theocharis (2005), Petrakis *et al.* (2005)], Principal Component Analysis and Discriminant Analysis were used to establish chemically the authenticity of the Cypriot traditional spirit Zivania, but these techniques revealed difficulties in making this characterisation.

In the present application, a non-linear classification model has been built by means of Counterpropagation Artificial Neural Networks [Ballabio *et al.* (2006a)]. The aim of this model is the characterization of Zivania and the differentiation of this alcoholic beverage from other, similar, beverages from all over the world, especially Europe. This procedure may be an ideal tool for describing Zivania's uniqueness, since the mapping based on the Neural Networks has shown acceptable predictive capabilities.

Moreover, the role of each variable in the classification model has been considered: Counterpropagation Artificial Neural Network results have been analysed by means of Principal Component Analysis, in order to study which variables have a real discriminant role in the classification model. This procedure appeared as a promising tool to study the relationship between variables and classes in a global way and not variable by variable, and to obtain a multivariate overview of variable behaviour in the classification model.

7.2.4 Multicriteria Decison Making for process monitoring

The application of the quality approach requires managerial tools able to control the system under control. There are several approaches to do this, but since the systems to be monitored are usually complex systems, chemometrics and multivariate analysis can play a fundamental rule for this task. In particular MultiCriteria Decision Making methods [Massart *et al.* (1997), Keller *et al.* (1991), Lewi and Boey (1992), Hendriks *et al.* (1992)] are here proposed as useful tools for the production process monitoring [Ballabio *et al.* (2004)]. These methods are able to suggest the optimal operative choice, on the basis of rankings of the different possibilities, by means of Desirability and Utility indices. In order to do so, it is necessary to select the optimal operative conditions for each considered parameter of the process. MultiCriteria Decision Making methods have been applied for the monitoring of production of an instant sport drink and a simple software has been developed in order to calculate the Desirability and utility indices on these data.

chapter 8

Applications of CAIMAN on food data

8.1 Introduction

During recent years there has been increasing interest in the certification of the geographical origin of food products, since authenticity and quality issues can be often associated with a given geographical origin. The protected denomination of origin (PDO) for agricultural products has been introduced with official European regulations, which allow the labelling of some products with the names of the geographical area of production. Extra virgin olive oil (EVOO) and wine are two products frequently designated with a PDO. In fact, both wine and EVOOs have an important rule in the commercial market, since they are extensively cultivated across the Mediterranean countries and hugely relevant for Mediterranean food industry.

Therefore, an important European regulation allows the PDO labelling of some European EVOOs and this designation guarantees that the quality of the product is closely linked to its geographical origin [Commu-

nity (1992)]. The denomination of origin is also established for wines produced in particular geographical regions, with typical characteristics linked to natural factors, to the environment, and to traditions of the region [Community (1987)]. In the European Union quality wine produced in determined regions are labelled as VQPRD and this is normally used to indicate the denomination of origin of wine. Moreover, the problem of geographic identification of food becomes more attractive when it concerns to restricted production areas: PDO wines and olive oils are frequently produced in small areas and often there are different PDO products located in one region [Brescia *et al.* (2002)].

Given these premises, it's clear how the development of methods for the geographical classification of food is becoming very important. Moreover, reliable techniques for origin authentication are essential, since official analysis of virgin olive oil and wine involves a series of several determinations of chemical and physical parameters that will be of scarce use in the geographical certification. So, in recent years, several efforts have been made to authenticate the origin of wines and olive oils by means of chemometric analysis, with different chemical and physical parameters, and multivariate analysis proved to be a powerful tool for determining the geographical origin, both for wine [Forina *et al.* (1986), Benito *et al.* (1999), Forina and Drava (1997), Gonzalez and Pena-Mendez (2000), Kallithraka *et al.* (2001), Kosir *et al.* (2001), Perez-Magarino *et al.* (2002)] and oil [Alves *et al.* (2005), Aparicio and Luna (2002), Armanino *et al.* (1989a), Bianchi *et al.* (2001), Boggia *et al.* (2002), Brodnjak-Voncina *et al.* (2005), Cerrato Oliveros *et al.* (2005), Eddib and Nickless (1987), Garcia-Gonzalez and Aparicio (2003a), Garcia-Gonzalez and Aparicio (2004), Lanteri *et al.* (2002), Mannina *et al.* (2001), Salter *et al.* (1997), Stella *et al.* (2000), Pinheiro and Esteves da Silva (2005), Tapp *et al.* (2003), Tsimidou *et al.* (1987), Tsimidou and Karakostas (1993), Zupan *et al.* (1994)].

In this work, the geographic classification of wine and olive oil samples has been carried out by means of CAIMAN [Ballabio *et al.* (2006c), Todeschini *et al.* (2005)]. These results have been compared with the classification performances of LDA and QDA, since these are two of the most

commonly used classification techniques, and great attention has been focused on the validation capabilities of the classification models. The geographic characterization has been studied on three different datasets: first, extra virgin olive oils, produced in a small area close to the Garda lake (northern of Italy) and labelled as PDO (the dataset has been called **GarOils**); second, three types of Barbera wines with different denominations of origin, but produced in enclosed geographical areas (dataset **BarWine**); finally, a dataset of olive oils, with 572 samples from 9 different Italian areas of productions (dataset **ItaOils**).

Validation of classification models

Since the main aim of geographical characterisation is the application of the models to unknown samples, great attention has been focused on the predictive capabilities of the classification methods. All the classification models, obtained both by CAIMAN and DA, have been validated using leave-one-out (LOO) and leave-more-out (LMO) procedures. The LOO procedure removes each sample from the data set, one at a time. The classification model is rebuilt and the class of the removed sample is predicted by using the obtained model. All the samples are sequentially removed and reclassified. Finally the percentage of wrong assignments is calculated (Error Rate leave-one-out, $\mathrm{ER_{LOO}}$).

Since LOO can provide overoptimistic results, also a more robust validation technique (LMO) has been applied. In the LMO procedure, a percentage s of samples is randomly removed from the data set; then, the classification model is rebuilt without these objects and the classes of the removed sample are predicted by the obtained model. This procedure is repeated r times, always with a random selection of s samples. Finally the percentage of wrong assignments is calculated (Error Rate leave-more-out, $\mathrm{ER_{LMO}}$). In the following analysis, 500 repetitions ($r = 500$) and a percentage of test samples equal to 20% ($r = 20\%$) have been used. Moreover, in order to check the CAIMAN predictive performances, the leave-more-out technique has been performed for each α value (between 0 and 1, with

step 0.1).

Moreover, a cross validation procedure for the optimisation of parameters has been performed in the following cases: a) the selection of an optimal α value (D-CAIMAN and M-CAIMAN); b) the selection of a leverage threshold (A-CAIMAN); c) the selection of a subset of variables (see next paragraph), both for CAIMAN and DA. Approaching these tasks, the samples are splitted into different cross validation groups. Once at a time, each validation group is removed from the training set, the parameters (i.e. the alpha value, the leverage threshold or the subset of variables) are chosen on the basis of the results achieved by the training set, by minimising the error rate. The optimised model is subsequently tested on the samples of the removed validation group: therefore, for each step, the optimisation is performed without the samples to be predicted [Burden *et al.* (1997)]. At the end of the procedure, the percentage of wrong assignments in the cross validation groups (ER_{CV}) is calculated. If ER_{CV} is satisfactory, the full model is built and the parameter optimisation is performed by taking into account only: a) the range of alpha values achieved in the models obtained by each cross validation groups; b) the range of leverage thresholds achieved with each cross validation groups; c) the subsets of variables selected with each cross validation groups, with a relative frequency higher then 1 [Baumann and Stiefl (2004)]. In the following analysis, five cross validation groups have been used.

Summarising, when parameter optimisation is required, the models are firstly evaluated on the basis of the ER_{CV}, and only the suggested parameters are then used in order to evaluate the parameter optimisation in the full model, i.e. with all the samples. Then, the full model predictive capability is also tested by means of LOO and randomly repeated LMO procedures

Variable reduction

Before approaching the different classification tasks, it has been necessary to reduce the dimension of the data matrices, since QDA and CAIMAN

require a n_g/p ratio greater than 2 or 3, where p is the number of variables and n_g the number of samples for the g-th class. To do so, we employed the forward variable selection technique [Jennrich (1977)]: this method starts with no variables and adds one variable at a time to the model; the inclusion of a variable is based on the $\mathrm{ER_{LOO}}$ value, i.e. the variables will be entered into the model if $\mathrm{ER_{LOO}}$ is minimised.

Multivariate normal distribution

Finally, since LDA and QDA are based on multinormality assumption, all the datasets have been checked for multivariate normality. When several variables are present, checking each variable for univariate normality could not be the only approach, because the variables are correlated and normality of the individual variables does not guarantee joint normality [Rencher (2002)]. Checking for multivariate normality has been carried out with the procedure based on the Mahalanobis distance (MD):

$$MD_i = (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \tag{8.1}$$

where $\mathbf{x}_i$ is the i-th sample, $\bar{\mathbf{x}}$ is the average, and $\mathbf{S}$ the covariance matrix [Johnson and Wichern (1992)]. When the population is multivariate normal, the Mahalanobis distances should behave like chi-square random variables. Therefore, Mahalanobis distances can be ordered from smallest to largest values and then plotted against the percentiles of the chi-square distribution with p degrees of freedom, where p is the number of variables. bla bla bla. The resulting plot (called chi-square plot) should look like a straight line in presence of multivariate normality, while a curved pattern suggests lack of normality. Multivariate normality could be also indicated if approximately half of the Mahalanobis distances are less then or equal to $\chi^2_p(0.5)$, i.e. the 50th percentile of the chi-square distribution with p degrees of freedom.

8.2 Data

The **GarOils** dataset [Cosio *et al.* (2006)] has included 53 samples of EVOOs from five different regions: Garda, 36 samples; Spain, 6 samples; Sardegna, 5 samples; Campania, 4 samples; Abruzzo, 2 samples. The sampling has included also 19 commercial EVOOs: 3 samples labelled as Garda PDO; 3 samples of Garda, not labelled as PDO; 13 samples collected on the market, produced with unknown cultivars. All the 19 commercial samples have been used only to test the classification model. For our purposes, all samples have been divided into 2 classes (Garda and not-Garda), since the Garda samples belong to a small production, located in the lake of Garda (north of Italy) and distinguished with a European Protected Denomination of Origin trademark since 1998. Each sample is described with 31 chemical measurements (free acidity, peroxide value, ultraviolet indices, phenol content and signals collected with an electronic tongue and an electronic nose); therefore a variable reduction has been needed.

The **BarWine** dataset [Buratti *et al.* (2006)] has included 48 samples of three types of Barbera wines with different Protected Denomination of Origin, but produced in enclosed geographical areas: 23 are Barbera Oltrepo produced in the Province of Pavia (Lombardia); 13 are Barbera Piemonte produced in the province of Alessandria, Asti, Cuneo (Piemonte); 12 Barbera Asti produced in the province of Asti and Alessandria (Piemonte). The original data set consists of four classes; however, it has been here modified and the last class has been removed, because of the small number of class objects (5 Barbera Alba produced in the province of Cuneo, Piemonte). Each sample is characterised with 28 chemical measurements (acidity, pH, conductivity, level of alcohol, total extract, oxidation-reduction potential, wine absorbance at four different wavelengths and signals collected with an electronic tongue and an electronic nose): also in this case a variable reduction has been needed.

Finally, the **ItaOils** dataset [Forina *et al.* (1983)] has included 572 samples of olive oils from 9 different Italian areas of productions. Since

Table 8.1: Resume of the characteristics of each dataset: number of variables, samples, classes, and worst class sample/variable ratio

Data	Variables	Objects	Classes	Worst n/p
GarOils	31	53	2	0.6
BarWine	28	48	3	0.4
ItaOils	8	572	9	3.1

each sample is described with the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of the olive oils, in this case no variable reduction has been needed. A summary of the characteristics of each dataset (number of samples, number of variables, number of classes, class partition and worst class samples/variable ratio) is shown in Table 8.1.

8.3 Results

Multivariate normal distribution

Initially, the multivariate normal distribution has been checked for the three datasets, following the procedure based on the Mahalanobis distances. The resultant chi-square plots are shown in Figure 8.1. GarOils appears to be the less multivariate normal distributed: the plot is not straight, even if the 54% of the Mahalanobis distances are minor then the chi-square threshold, $\chi^2_p(0.5)$. In particular, the smallest distances appear to be too small and the largest distances appear to be too large relatively to the expected distances. On the contrary, the BarWine plot is more straight, but only 44% of the Mahalanobis distances are lower then $\chi^2_p(0.5)$, while approximately half of the Mahalanobis distances should be lower then this threshold to assess a multivariate normal distribution. Therefore, also BarWine does not appear to be clearly multivariate normal, even if it is difficult to reach a definitive conclusion. Finally, in the ItaOils plot, two samples with large distances stand out as clearly different from the rest of the pattern, which, apart from these distances,

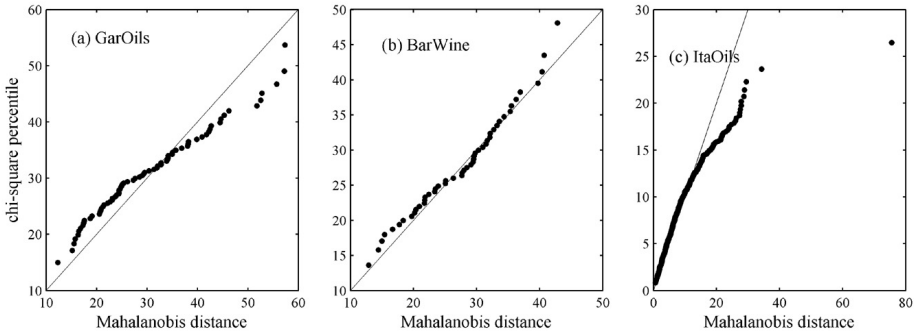

Figure 8.1: Chi-square plots of the analysed data: (a) GarOils, (b) BarWine and (c) ItaOils.

does conform to the expected straight line relation. Only few observations (approximately the 6% of the samples, characterised by the largest distances) do not fit the linear relation. Moreover, almost the 60% of the Mahalanobis distances are minor then the chi-square threshold, $\chi^2_p(0.5)$. Thus, ItaOils can be considered the more multivariate normal distributed, among the three analysed datasets.

GarOils

Since the samples of this dataset has been grouped into two classes on the basis of their provenance (Garda oils and not-Garda oils) and the main aim is the discrimination of the Garda samples from all the other not-Garda oils, the asymmetric classification approach would be more suitable (see next paragraphs). However, D-CAIMAN, M-CAIMAN and DA have been used on these data, in order to compare the classification performances.

The forward selection has been applied to the original variables, in order to reduce the data dimensions. As explained before, five cross validation groups have been used to optimise the a value and the subset of selected variables. The achieved ER_{CV} are equal to 0% both for D-CAIMAN and M-CAIMAN (Table 8.2) and indicate an adequate and consistent cross validation result; all the cross validation groups have indicated an optimal a value equal to zero and only two variables (no.4,13) have been selected

Table 8.2: Comparison of error rates (%) obtained by D-CAIMAN (DC), M-CAIMAN (MC), LDA and QDA for GarOils dataset. CV refers to cross validation, LOO refers to leave-one-out validation, LMO to leave-more-out validation, EXT to the external validation; VS refers to the variable selection performed by the forward technique; for M-CAIMAN, not assigned (%) and confused (%) samples are, respectively, reported in brackets: for example, M-CAIMAN has given an ER_{LMO} = 0%, 2.3% of not assigned and 0% of confused samples. The selected variables (var) and the α values (for CAIMAN) are also reported.

GarOils	**var**	**CV**	**LMO**	**LOO**	**EXT**	α
MC-VS	4,13	0 (0-0)	0 (2.3-0)	0 (0-0)	5.3 (5.3-0)	0
DC-VS	4,13	0	0	0	10.5	0
LDA	All		9.8	7.6	10.5	
LDA-VS	1,10,16,24	3.8	3.1	1.9	0	
QDA-VS	1,13	1.9	0.1	0	15.8	

with a frequency higher then one, both for D-CAIMAN and M-CAIMAN. Therefore, the final models have been built with these settings and their predictive capabilities tested with LOO and LMO procedures. Moreover, the models have been tested also with the external test set of commercial oil samples (Table 8.2).

To better understand how the leverage can be interpreted, the leverage plot of the CAIMAN model with the two selected variables is shown in Figure 8.2. The leverage plot is basically a scatter plot obtained by projecting the samples in the space defined by the leverages referring to two chosen classes. Each axis represents a class and farer the objects are from the axis origin, greater their distance is from the corresponding class. Objects falling in the left bottom corner are close to both the class spaces; objects falling in the left top corner are very close to the class represented by the horizontal axis and far from the other one; on the contrary, objects falling in the right bottom corner are very close to the class represented by the vertical axis and far from the other one; finally, objects falling in the right top corner are far from both the classes. Considering the leverage plot of Figure 8.2 (notice that the two axis do not have the same scale), it can be seen that the leverages from the Garda class of all the objects

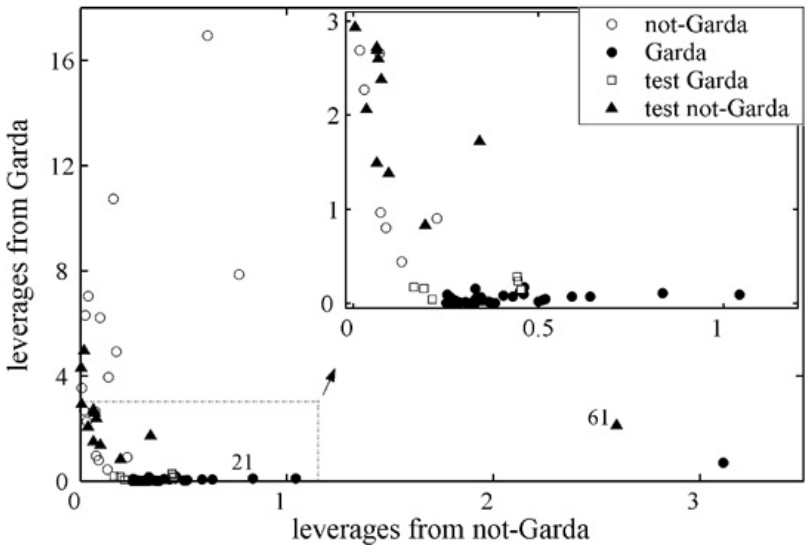

Figure 8.2: CAIMAN on GarOils with the selected subset of variables: leverage plot. Garda and not-Garda samples are plotted with different colours, while test samples are plotted with a different shape. The area marked with the dotted line in the main graph is enlarged in the top-right corner.

effectively belonging to this class are very small, while leverages from the not-Garda class of the same objects are higher. For example, object 21 is typical of Garda, having a low leverage value (0.11) for this class with respect to the not-Garda (0.83). On the contrary, object 61 (a commercial not-Garda oil) has leverages 2.12 and 2.59 from Garda and not-Garda classes, respectively, and is classified as Garda by both D-CAIMAN and M-CAIMAN.

In order to compare and evaluate the quality of the results achieved with CAIMAN, LDA has been applied to the whole dataset, while both QDA and LDA have been applied to the dataset reduced with the forward selection. As before, five cross validation groups have been used to optimise the forward selection. The achieved ER_{CV} are equal to 3.8% and 1.9% for LDA and QDA, respectively. For QDA, only two variables (no. 1, 13) have been retained with a frequency higher then one and con-

sequently the final model has been built considering these two variables. For LDA, five variables have a frequency higher then 1 in the five subsets selected by the cross validation steps; by applying a final forward selection on these five variables with all the samples, only four variables (no. 1, 10, 16, 24) are indicated as the optimal subset. Then, the LDA and QDA final models have been tested with LOO and LMO procedures: the results achieved with the Discriminant Analysis on the selected subsets of variables are summarised in (Table 8.2). As it can be seen, the results obtained by CAIMAN are satisfactory; by considering the different error rates (in cross-validation, ER_{CV}, in leave-one-out, ER_{LOO}, in leave-more-out, ER_{LMO}, and on the external test set, ER_{EXT}) the best approach seems to be M-CAIMAN on the selected variables; the best DA result is LDA with the selected variables ($ER_{LMO} = 3.1\%$, $ER_{LOO} = 1.9\%$ and $ER_{EXT} = 0\%$); QDA on the selected variables have the highest external error rate ($ER_{EXT} = 15.8\%$), while LDA on the original variables gives globally the worst results.

BarWine

As for GarOils, the data dimension of BarWine has been initially reduced with the forward selection. The error rates obtained in cross validation are equal to 4.2% both for D-CAIMAN and M-CAIMAN (Table 8.3); since cross validation results are acceptable, the final optimisation has been performed with all the samples, considering the ranges of alpha values achieved in the models of the cross validation groups and the variables with a relative frequency higher then 1 (5 and 7 for D-CAIMAN and M-CAIMAN, respectively). In this cross validation step, the a values have ranges 0-1 and 0-0.3 for D-CAIMAN and M-CAIMAN respectively: therefore the final optimisation has been performed by considering only these ranges. Two different subsets of three variables have been retained in the final models (variables no. 27, 24, and 26 for D-CAIMAN, variables no. 21, 19, and 26 for M-CAIMAN) and again the validation procedures applied to the reduced data. Then, the data have been classified by means of LDA

Table 8.3: Comparison of error rates (%) obtained by D-CAIMAN (DC), M-CAIMAN (MC), LDA and QDA for BarWine dataset. CV refers to cross validation, LOO refers to leave-one-out validation, LMO to leave-more-out validation; VS refers to the variable selection performed by the forward technique; for M-CAIMAN, not assigned (%) and confused (%) samples are, respectively, reported in brackets. The selected variables (var) and the α values (for CAIMAN) are also reported.

BarWine	**var**	**CV**	**LMO**	**LOO**	α
MC-VS	19,21,26	4.2 (0-0)	0.7 (4.5-0)	0 (0-0)	0.1
DC-VS	24,26,27	4.2	0.4	0	0.3
LDA	All		7.6	4.2	
LDA-VS	6,18,24,27	12.5	4.3	4.2	
QDA-VS	2,8,19,25,27	14.0	8.4	2.1	

with all the original variables, and by means of the subsets of selected variables. During the variable selection, the cross validation has given ER_{CV} equal to 12.5% for LDA and 14.0% for QDA. The results obtained both with CAIMAN and DA are summarised in Table 8.3. The forward selection has enabled both D-CAIMAN and M-CAIMAN to get the best result (ER_{LMO} = 0.4% and 0.7% with a = 0.1 and 0.3, respectively). Also LDA has given a better performance after the variable selection, but its ER_{LMO} (4.3%) is relatively higher then the ER_{LMO} achieved by CAIMAN, while QDA has given the worst results. Since the D-CAIMAN model with three selected variables is the best classification model found on the BarWine data (ER_{LOO} = 0%, ER_{LMO} = 0.4%, with an a value equal to 0.1), the correspondent leverage plots are shown in Figure 8.3. In the first plot (a) the classes Oltrepo and Piemonte are compared, while in the second one (b) the class leverages of Piemonte and Asti are shown. Notice that the Oltrepo samples are not plotted in (b), since their leverages are too high: for example, the leverages of Oltrepo samples from class Piemonte range from 20 to 140 (as it can be seen in the first plot). It is clearly visible that all the objects have small leverages from the membership class, while the leverages from the other classes are significantly higher: this let achieve the best result and no class overlap in the classification.

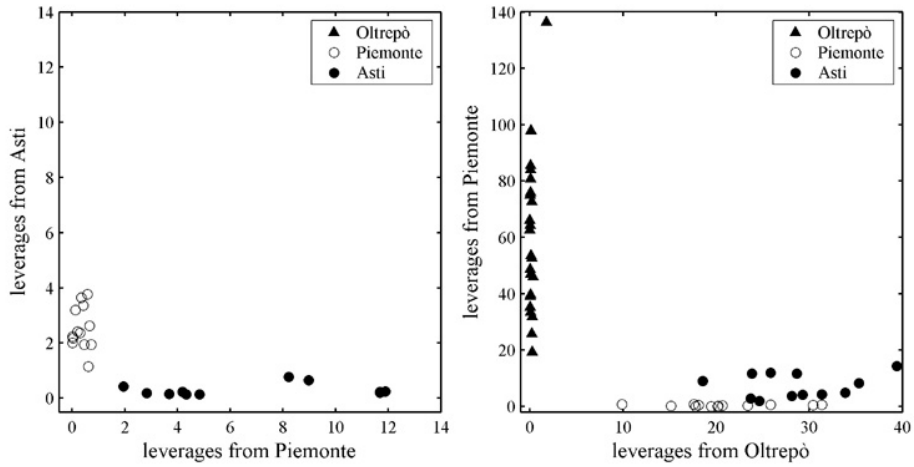

Figure 8.3: CAIMAN on BarWine with the selected subset of variables: leverage plots; (a) on the left leverages from Oltrepo and Piemonte classes; (b) on the right leverages from Piemonte and Asti. Oltrepo samples are not plotted in (b), since their leverages are too high, as explained in the text.

ItaOils

On the contrary of GarOils and BarWine, the number of objects of each class in ItaOils is significantly greater than the number of variables. The worst objects/variables ratio is 3.13 (Table 8.1) and no dimension reduction is theoretically needed. Therefore, D-CAIMAN, M-CAIMAN, LDA, and QDA have been applied using the eight original variables. Also in this case, the a value optimisation has been evaluated by means of cross validation (the relative ER_{CV} is shown in Table 8.4). Considering the results achieved with the cross validation groups, the a values have ranges 0.8-1 and 0.9-1 for D-CAIMAN and M-CAIMAN respectively: the final optimisation has been performed by taking into account these ranges. LDA and QDA have performed better then CAIMAN and QDA has given the best result, with an ER_{LOO} and ER_{LMO} equal to 4.2% and 4.6% respectively, while the Error Rates calculated by CAIMAN are all higher then 10% (Table 8.4). One of the reasons of this behaviour could be due to

Table 8.4: Comparison of error rates (%) obtained by D-CAIMAN (DC), M-CAIMAN (MC), LDA and QDA for ItaOils dataset. CV refers to cross validation, LOO refers to leave-one-out validation, LMO to leave-more-out validation; VS refers to the variable selection performed by the forward technique; for M-CAIMAN, not assigned (%) and confused (%) samples are, respectively, reported in brackets. The selected variables (var) and the α values (for CAIMAN) are also reported.

ItaOils	**var**	**CV**	**LMO**	**LOO**	α
MC	All	13.5 (0-1.9)	12.8 (0-1.6)	10.8 (0-2.1)	1
MC-VS	2,5,6	15.9 (0-3.3)	13 (0-2.3)	10.9 (0-4.9)	0.3
DC	All	14.3	13.5	11.9	1
DC-VS	4,5,6	18.0	13.5	12.9	1
LDA	All		6.4	6.5	
QDA	All		4.6	4.2	

the fact that ItaOils can be considered the more multivariate normal distributed (Figure 8.1), among the three analysed cases: unlike the approach of CAIMAN, Discriminant Analysis is based on multinormality assumption, exploits this condition and better classifies the ItaOil samples. In order to improve the CAIMAN classification performances, the forward variable selection has been also considered. Subsets of three variables (no. 2, 5, 6 and 4, 5, 6 for M-CAIMAN and D-CAIMAN, respectively) have been selected. With respect to the full models, the prediction capabilities have not been improved; on the other hand a simplification of the models has been obtained: for example, considering the results achieved by D-CAIMAN, ER_{LMO} and ER_{LOO} are substantially comparable to the ones achieved with all the variables, but in the reduced model only three of eight available variables have been used, i.e. a simpler and easier classification model has been carried out.

GarOils: asymmetric classification case

Finally, Asymmetric CAIMAN has been applied to GarOils. The aim of this model is the geographical characterization of Garda samples and the differentiation of these olive oils from all the others. This procedure may be

an ideal tool for describing PDO Garda oil uniqueness and, consequently, for the protection of Garda PDO against adulteration. So, the class Garda is here taken as the reference class to be modelled; all the other samples are considered as not-Garda samples.

Therefore, only one model has been evaluated for the Garda class with the corresponding leverage values. Due to the sample/variable ratio, the data dimension has been reduced by selecting the variables. As explained before, five cross validation groups have been used to optimise the leverage threshold and perform variable selection. Since A-CAIMAN is an asymmetric classification method, the geometric mean of specificity and sensitivity of the modelled class (instead of error rate) has been used for the parameter optimisation. The achieved geometric mean in cross validation is equal to 95% and represents an adequate result. All the cross validation groups have indicated an optimal leverage threshold in the range 0.3-3 and only two variables (no. 11, 13) have been always selected. Therefore, the final A-CAIMAN model has been built with these two variables, selecting a threshold into the suggested range. Consequently, the leave-one-out validation performed with all the samples has proposed an optimal leverage threshold equal to 1.6: this simply means that all the samples with a leverage from the Garda class lower then 1.6 will be classify as Garda, otherwise as not-Garda. With this threshold, specificity equal to 100% and sensitivity equal to 100% have been reached on the training set. The Garda class has been well described, the entire Garda samples have been correctly classified and not-Garda samples have never been assigned to the Garda class during the leave-one-out iterations. Then, the leverages for the external test set samples (composed by 19 commercial oils) have been calculated: all the commercial Garda and not-Garda oils have been correctly assigned.

To better clarify the procedure, Figure 8.4 and Figure 8.5 are shown. In Figure 8.4, sensitivity and specificity are reported for different values of the leverage threshold: the selection of the leverage threshold is based on this graph. The threshold maximising the class sensitivity (100%) is 0.3. With this threshold also the class specificity is equal to 100%. On

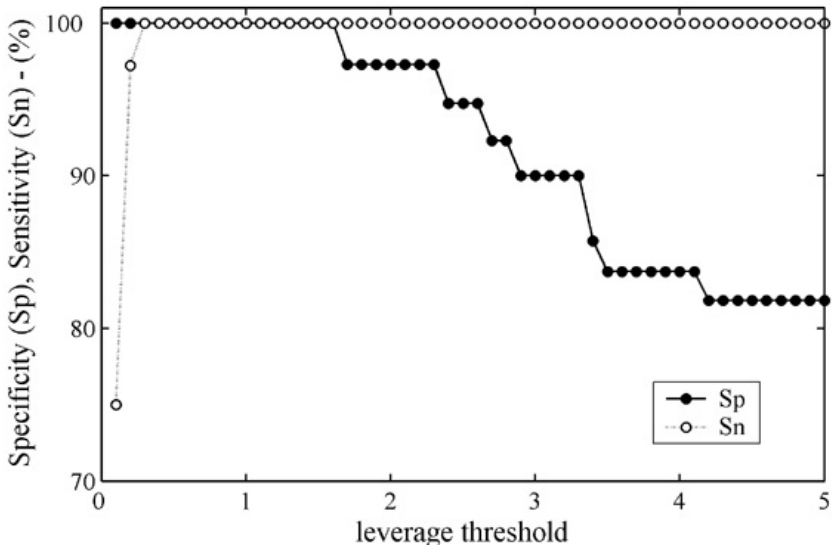

Figure 8.4: A-CAIMAN on GarOils with the selected subset of variables: sensitivity and specificity (y axis) for different values of the leverage threshold (x axis).

the other hand, the maximum specificity is reached for a threshold lower then 1.6, and this threshold has been selected as the optimal one. The geometric mean, the Jaccard's coefficient and the Pearson's F coefficient have followed a similar behaviour and confirmed the optimality of the selected threshold. Depending on the aim of the analysis, a different leverage threshold could be selected in the range 0.3-1.6.

In Figure 8.5 the leverages from Garda class are plotted; on the horizontal axis the object identification number is reported. In order to get an optimal graphical resolution, samples no. 43 and 45 (not-Garda) have not been plotted, since their leverages are high (16.2 and 23.2, respectively). The dotted line represents the selected threshold. As it can be seen, all the Garda oil samples have a leverage value significantly lower then 2, while all the not-Garda sample leverages are higher then the specified threshold and correctly assigned. The same conclusions can be done on the commercial test samples and therefore specificity and sensitivity of 100% have

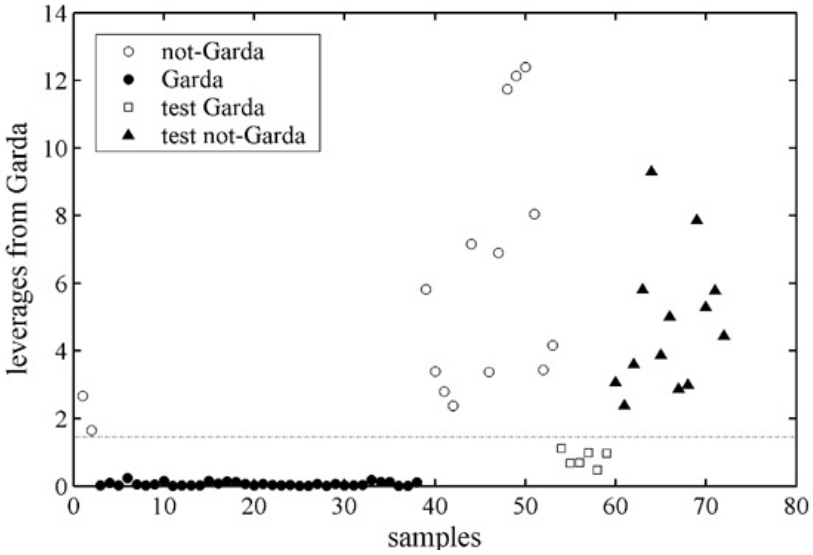

Figure 8.5: A-CAIMAN on GarOils with the selected subset of variables: leverages from Garda class for all the data together with the optimal leverage threshold. Garda and not-Garda samples are plotted with different colours, while test samples are plotted with a different shape.

been reached on the test set.

8.4 Conclusions

CAIMAN is a classification method based on a simple mathematical approach where the results can be easily interpreted by analysing the leverage and hyper-leverage values and no assumption on the multinormal distribution of the data is required. The application of this new classification approach to the geographical origin identification of food samples seems to offer several advantages: first of all, D-CAIMAN and M-CAIMAN show - on an average basis - good performance when compared to two other methods, such as Linear and Quadratic Discriminant Analysis.

Moreover, asymmetric CAIMAN is able to deal in a simple and easily interpretable way classification problems related to tipicity, authenticity,

and uniqueness characterization, which are increasingly common in food quality issues. CAIMAN requires, in the same way as QDA, the number of class objects to be significantly greater than the number of variables: it is possible to overcome this limit with a reliable dimension reduction.

chapter 9

Electronic sensor selection based on Hasse approach

9.1 Introduction

During the last decade, electronic nose has increased in uses, capabilities and applications in food science [Bartlett *et al.* (1997), Gardner and Bartlett (1993)]. Basically, the principle involved in the electronic nose is the transfer of the total headspace of a sample to a sensor array, where each sensor has partial specificity to a wide range of aroma molecules. These non-selective gas sensors are theoretically able to simulate human sensing and give an objective tool of detecting aromatic fingerprints. Since electronic noses offer several advantages (cheapness, quickness, simplicity, little or no prior sample preparation), they have been used in food science for a variety of applications: assessment of food properties [Brezmes *et al.* (2001), Garcia-Gonzalez and Aparicio (2003b), Guadarrama *et al.* (2000), Hines *et al.* (1999), Llobet *et al.* (1999), Vinaixa *et al.* (2004), Vinaixa *et al.* (2005)], detection of adulteration [Cerrato Oliveros *et al.* (2002)], sensory properties prediction [Buratti *et al.* (2006), Cozzolino *et al.* (2003)], classi-

fication of different food matrices [Gallardo *et al.* (2005), Gonzalez Martin *et al.* (1999), Pilar Marti *et al.* (2004), Stella *et al.* (2000)].

Unlike traditional analytical methods, electronic nose sensor responses do not provide information on the nature of the compounds under investigation, but only give a digital fingerprint of the food product, which can be subsequently investigated by means of chemometric methods. In fact, multivariate analysis proved to be especially able to handle the large amounts of data produced by modern analytical techniques and has been successfully applied on electronic nose data.

However, even if each sensor is linked to specific classes of compounds, not all the sensors contribute to the characterisation of the analysed product. Therefore, the selection of sensors can be of paramount interest to maintain only the sensors that contain significant information for the specific task [Eklov *et al.* (1999)]. By removing irrelevant sensors, the fingerprint description can be simplified and more qualitative information can be extracted.

Since all the spectra achieved along time are intrinsically ordered, the data provided by electronic noses can be also considered as sequential data (where time is considered the ordering variable) and characterised by means of Hasse matrices (chapter 5). This application deals with the above mentioned topics: extra virgin olive oil samples of different geographical origin have been taken into account and Hasse distances have been used in order to characterise the fingerprint of each electronic nose sensor and select the sensors which appear more able to discriminate the olive oil origins [Ballabio *et al.* (2006b)].

9.2 Hasse distances and electronic nose data

Also electronic nose data can be characterised by means of Hasse matrices and their similarity/diversity measures. In fact, even if parameters are usually extracted from the electronic nose spectra, it is not necessary to transform the measured time profile into univariate features [Skov and Bro (2005)]. As a result of this parameterisation, relevant information from

the raw data could be lost, while by preserving the time information, in some cases more information can be extracted. Therefore, if the whole signal among time is considered, i.e. the sequential property of the data is preserved, the Hasse approach can be easily applied.

For electronic nose data, only the property variable is used: in this way, incomparabilities cannot be present and only Hasse matrices with +1 or -1 off-diagonal values can be obtained (equation 5.2). Hence, for each sample (s) of each electronic nose sensor (e) an Hasse matrix ($\mathbf{M}^{se}$) can be calculated, by considering as unique property variable the intensities among time, i.e. the time profile of the sample. If a total number of t times are taken into account, each $\mathbf{M}^{se}$ matrix, built on the basis of the ordering relationships of these t times, has dimensions t times t, and each element m_{ij} represents the relationship between the signal intensities at i-th and j-th times, i.e. it says if intensity in time i is lower (m_{ij} = -1) or higher ($m_{ij} = 1$) then intensity in time j.

Consequently, a total ordering relationship is obtained, since no incomparabilities will be present. However, in this way, the pairwise ordering relationships are also taken into account when the similarity measure is calculated between two different samples. Concerning the augmented Hasse matrix (equation 5.3), the considered property added into the main diagonal of each Hasse Matrix is just the intensity (scaled on the maximum value), i.e. in this case:

$$m_{ii} = \frac{I_i}{\max(I)} \tag{9.1}$$

where I_i is the intensity of the i-th time. Therefore, the achieved Hasse matrix can be interpreted as a fingerprint of the time profile of each sample, i.e. a mathematical representation of the electronic nose signal, which takes into account both information of the curve shape and the intensity.

In order to clarify the Hasse approach, four curves (A, B, C and D) have been built as different Gamma probability density functions and the Hasse distances calculated. These four curves can represent the time profiles of

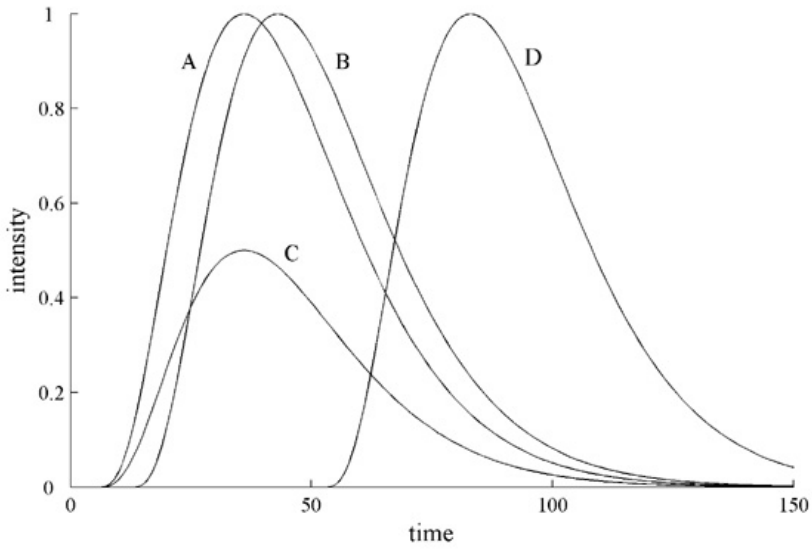

Figure 9.1: Plot of four simulated curves.

four samples, achieved by means of a singular electronic nose sensor. In Figure 9.1, each curve is shown as intensity values (I) plotted versus 150 times values. The curves B, C and D have been built by a small shift of curve A (B), by a big shift of curve A (C) and by its intensity reduction (D), respectively. In Table 9.1, some intensity (not consecutive) values for curve A are reported: the intensity column is the input of the Hasse analysis, as explained above. For example, taking into account the signals no. 33 (0.98), 34 (0.99), and 41 (0.96), the Hasse matrix elements are the following: $\mathbf{M}^A(33,34)$ =-1, while $\mathbf{M}^A(34,\ 33)$ = +1, since I(33) is lower than I(34); $\mathbf{M}^A(33,\ 41)$ = +1,$\mathbf{M}^A(41,\ 33)$ =-1, and so on.

In Table 9.2, a partial augmented Hasse matrix, relative to the data of Table 9.1 (times 3343), is showed. In the main diagonal, zero values have been replaced with the scaled intensities. After the whole Hasse matrices ($\mathbf{M}^A$, $\mathbf{M}^B$, $\mathbf{M}^C$ and $\mathbf{M}^D$) have been calculated, the weighted Hasse distances between the four curves have been carried out (Table 9.3), by using a weight w equal to 0 (i.e. taking into account only the ranking

Table 9.1: Time and intensity values of curve A. Ranges 1:3, 33:43 and 148:150 are reported.

Time	Intensity
1	0.00
2	0.00
3	0.00
...	...
33	0.98
34	0.99
35	1.00
36	1.00
37	1.00
38	0.99
39	0.99
40	0.98
41	0.96
42	0.95
43	0.93
...	...
148	0.00
149	0.00
150	0.00

Table 9.2: Augmented Hasse matrix, relative to the data of Table 9.1 (times 3343).

	33	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**
33	0.98	-1	-1	-1	-1	-1	-1	1	1	1	1
34	1	0.99	-1	-1	-1	1	1	1	1	1	1
35	1	1	1	1	1	1	1	1	1	1	1
36	1	1	1	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1	1	1
38	1	1	-1	-1	-1	0.99	1	1	1	1	1
39	1	1	-1	-1	-1	1	0.99	1	1	1	1
40	1	-1	-1	-1	-1	-1	-1	0.98	1	1	1
41	-1	-1	-1	-1	-1	-1	-1	-1	0.96	1	1
42	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.95	1
43	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.93

relationships, see equation 5.6) and equal to 1 (i.e. taking into account only the intensities). In the first case ($w = 0$), d_W(A, C), i.e. the weighted

Table 9.3: Weighted standardised Hasse distances (d_W) between the four simulated curves, calculated with weight $w = 0$ (not-italic) and $w = 1$ (italic).

	A	**B**	**C**	**D**
A	0.00	0.09	0.00	0.53
B	*0.09*	0.00	0.09	0.46
C	*0.15*	*0.17*	0.00	0.53
D	*0.48*	*0.43*	*0.38*	0.00

distance calculated between A and C, is equal to 0, since the curves A and C have exactly the same shape and do not shift; d_W(A, B) is small (with respect to the range 01), since the shift between A and B is small and d_W(A, D) is greater, since the shift between A and D is greater; d_W(C, D) is equal to 0.53 and is the greatest distance calculated with $w = 0$. The sensibility of the Hasse approach can be highlighted by considering also the results achieved with $w = 1$. In fact, the distance between curves D and C should be expected to be the greatest one, since the profiles differ in time and intensity, but this is not confirmed by considering only the intensities, since with $w = 1$, d_W(A, D) and d_W(B, D) are greater than d_W(C, D). Hence, the combination of the two approaches could represent an optimal solution.

9.3 Sensor selection based on Hasse class distance index

As explained in the previous paragraph, each sample analysed by means of each electronic nose sensor provides a Hasse matrix, which can be considered its fingerprint. A distance matrix $\mathbf{D}^e$ can be calculated for each e sensor, where each element d_e(s,t) is the weighted Hasse distance between samples s and t. If a number of G different classes of samples is present, it is possible to calculate for each class the maximum intra-class distance (with-in class distance), i.e. the maximum distance between the samples of the same class, and the minimum inter-class distance (between-

in class distance), i.e. the minimum distance between the samples of the considered class and all the other samples.

A class separability index (Hasse class distance index, HCD) can be easily calculated as the ratio between the average of G maximum intra-class distances and the average of G minimum inter-class distances. Finally, sensors can be ranked on the basis of this index, since it should be minimised in order to obtain a better class separation.

9.4 Data

The characterisation of electronic nose sensors with the Hasse distance has been studied on real data: olive oils samples, produced in different geographical areas, have been considered in order to select the sensors more able to distinguish their origin [Cosio *et al.* (2006)]. The dataset includes 52 samples of extra virgin olive oils, divided into two classes: oils produced in the Garda Lake region (Garda class, 35 samples) and oils produced in other different areas (not-Garda class, 17 samples). The signals collected by four metal oxide semiconductor field effect transistors (MOSFET) and 11 Taguchi type (metal oxide semiconductors, MOS) sensors among 100 times have been considered. In order to facilitate the exposition, a label (Sn, where n represents the number of the sensor) has been assigned to each sensor.

Summarising, the dataset include 52 samples belonging to two different classes; each sample is described by 15 time profiles (comprising 100 times) provided by the sensors. The data have been pretreated with a baseline correction [Skov and Bro (2005), Distante *et al.* (2002)], i.e. by calculating the difference between the signal and the baseline signal.

Calculations have been performed by MATLAB 6.5 (Mathworks), with routines built by the authors. The MATLAB modules for calculating the Hasse distances are free and available at www.disat.unimib.it/chm. The code sources of these routines are described in the appendix.

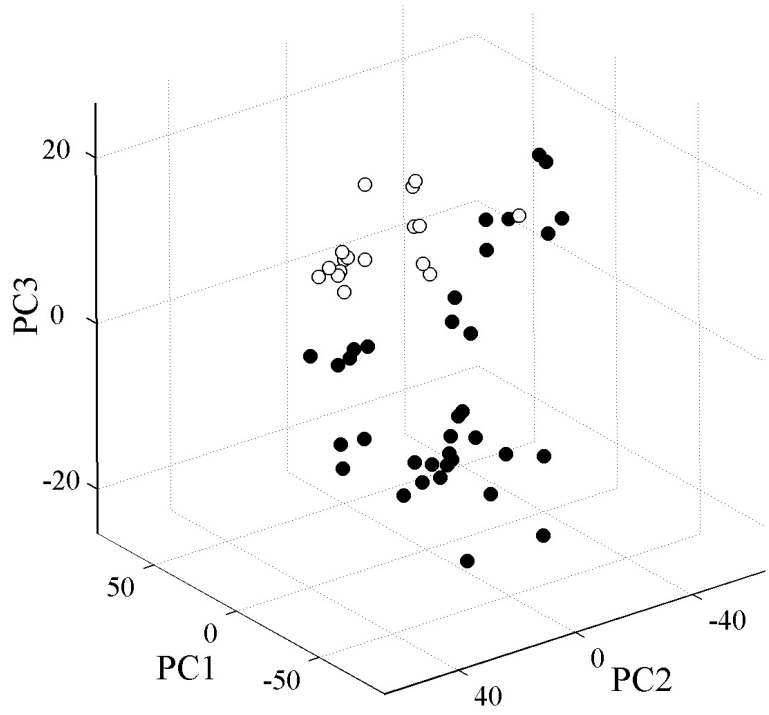

Figure 9.2: PCA on the unfolded data with all the sensors: score plot of the first three components (explained variance = 85%). Garda samples are drawn in black and not-Garda samples in white.

9.5 Results

Before approaching the sensor selection tasks, principal component analysis has been applied, in order to establish if the sensors, all together, were able to well separate the classes of samples. Hence, data have been unfolded: a matrix with 52 rows (samples) and 1500 columns (15 sensors times 100 signal values) has been built and analysed. The variables have been autoscaled and the first three components, which explain 85the total variance, have been considered. Examining the score plot (Figure 9.2), the not-Garda samples are placed in the space where the scores of the three retained components are positive, but this distinction is rough, since some samples have negative scores, especially on components 2 and 3. Hence,

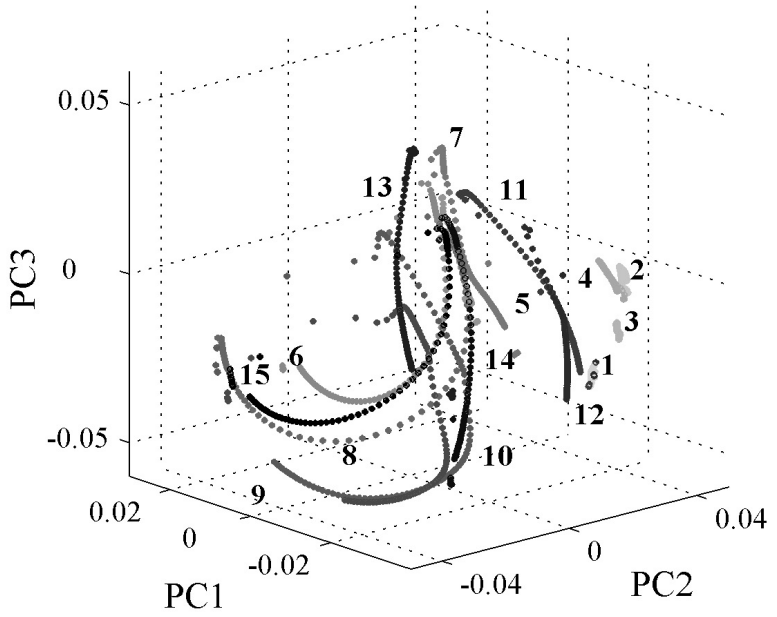

Figure 9.3: PCA on the unfolded data with all the sensors: loading plot of the first three components (explained variance = 85%). The different sensors are numbered (from 1 to 15) and marked with colour of different intensities (from grey to black).

it can be established that a trend of separation is present and the two classes could be approximately distinguished (with the exception of one not-Garda sample).

Considering the loading plot of the three retained principal components (Figure 9.3), firstly the time sequence relationships are clearly highlighted by the spatial sequences of the variables. Then, different behaviour of the sensors can be observed: sensors S1, S2, S3 and S4 have loadings really condensed in a confined area, with a small influence on the first and third components; the loadings of sensors S5, S7, S12, S13, S14 and partially S11 are more scattered on the third component; finally, the remaining sensors (S6, S8, S9, S10, S15) have much relevant loadings on components 1 and 3 and their values spread over the loading space.

Afterwards, the weighted standardised Hasse distances have been cal-

Table 9.4: Hasse class distance index (HCD) of the considered electronic nose sensors (S1S15), calculated with different weights: $w = 0$, 0.5 and 1.

	w=0		w=0.5		w=1	
rank	**sensors**	**HCD**	**sensors**	**HCD**	**sensors**	**HCD**
1	S10	4.60	S9	4.30	S9	5.50
2	S2	6.70	S10	4.40	S10	6.10
3	S9	6.80	S8	13.40	S6	28.30
4	S1	7.60	S2	16.20	S15	42.50
5	S4	8.50	S6	19.90	S8	44.50
6	S8	14.80	S15	21.60	S12	53.20
7	S14	15.40	S1	22.90	S2	55.70
8	S6	17.70	S4	27.30	S1	57.40
9	S7	20.20	S13	30.40	S7	57.50
10	S3	20.20	S7	33.70	S13	67.50
11	S5	23.10	S14	36.30	S4	75.20
12	S13	25.50	S12	41.10	S5	80.30
13	S11	27.50	S11	43.80	S14	84.70
14	S15	30.70	S3	43.90	S3	89.40
15	S12	58.30	S5	49.70	S11	108.20

culated between the samples, for each sensor; the relative Hasse class distance index (HCD) has been calculated and the sensors have been ranked on the basis of the achieved indexes (Table 9.4). Sensors S9 and S10 are placed in the first ranks in all the three cases, while only sensor S2 has a better rank position with respect to S9, when w is equal to 0. In order to get an objective evaluation of each sensor performance, it is possible to sum their three rank positions: also in this case, it is evident how S9 (3+1+1=5) and S10 (1+2+2=5) appear much better then all the other ones, since the third best sensor would be S2, with a sum of its rank positions equal to 13 (2+4+7). As discussed above for the simulated data, the optimal approach is to retain informations on both the ordering relationships and the scaled intensities; therefore, the results with $w = 0.5$ have been considered and, as showed, sensors S9 and S10 have values of HCD (4.3 and 4.4, respectively) significantly lower than the other sensors for this w value. On the basis of these results, sensors S9 and S10 have

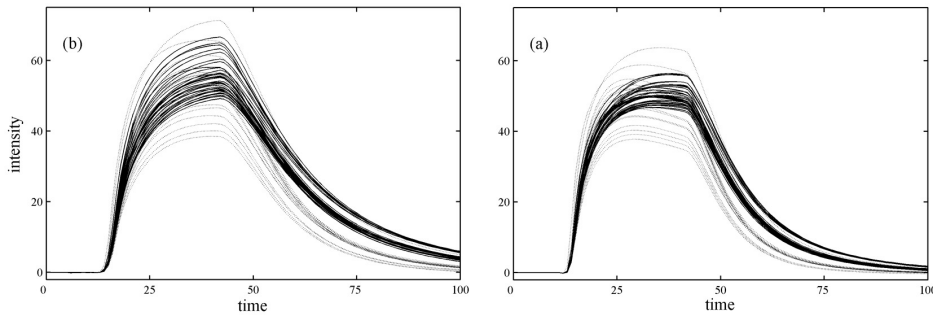

Figure 9.4: Time profiles of sensors S9 (a) and S10 (b): intensity values vs. time. Samples of Garda and not-Garda classes are drawn with different lines (Garda samples: solid line and not-Garda samples: dotted line).

been selected as the best sensors for class separation.

Considering the electronic nose signals of these selected sensors (Figure 9.4), it can be observed how the two classes are really characterised by different time profiles: regarding sensor S9, Garda oil samples have a maximum intensity value after time 40, while not-Garda samples (with the exception of one oil sample) have the maximum intensity in-between time 20 and 30. On the other hand, the time profiles of the two classes in S10 have the maximum value in the same time (approximately 45), but the curves of Garda samples are more inclined when approaching the maximum and less inclined after the maximum. Also the intensities of the two classes differ, especially for sensor S9. Hence, the selection of sensors S9 and S10 as the best sensor for class discrimination seems to be reliable.

Concerning sensor S2 (Figure 9.5), it also seems to have a different profile shape for each class. In fact, its HCD index is high when $w = 1$ and low when $w = 0$: in this lat case S2 has a better HCD value than S9, i.e. S2 appears more able than S9 to distinguish the two classes if only the profile shape is considered and the intensities are not taken into account. The difference of HCD indexes (6.8 for S9 and 6.7 for S2, Table 9.4) is not great and is due to the lower value of the maximum intraclass distance of

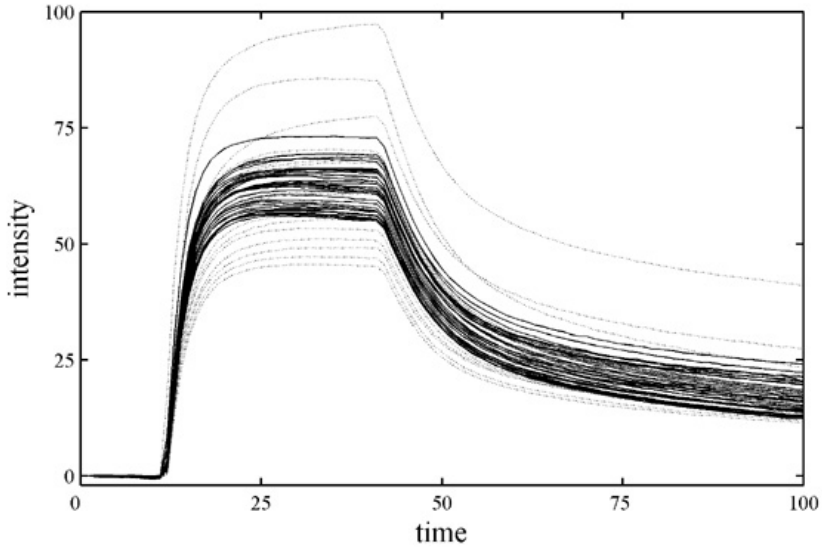

Figure 9.5: Time profiles of sensors S2: intensity values vs. time. Samples of Garda and not-Garda classes are drawn with different lines (Garda samples: solid line and not-Garda samples: dotted line).

not-Garda class samples, while the maximum intra-class distance of Garda class and the minimum inter-class distances are similar. By observing the time profiles, the difference between the shape of the profiles of the two classes is not obviously perceptible; but, considering the first derivative of sensor S2 (Figure 9.6), the Garda samples are more similar, while the first derivatives of not-Garda samples have more differences and, therefore, a different behaviour with respect to the Garda class exists.

However, in order to confirm the discrimination capability of the selected sensors (S9 and S10), PCA has been applied on a reduced dataset, where only sensors S9 and S10 have been retained. Hence, the data have been unfolded and a matrix with 52 rows (samples) and 200 columns (two sensors times 100 signal values) has been built. The variables have been autoscaled and the first three components, which explain together 86% of the data variance, have been retained. Comparing the score plots achieved with all the considered sensors (Figure 9.2) and with the selected sensors

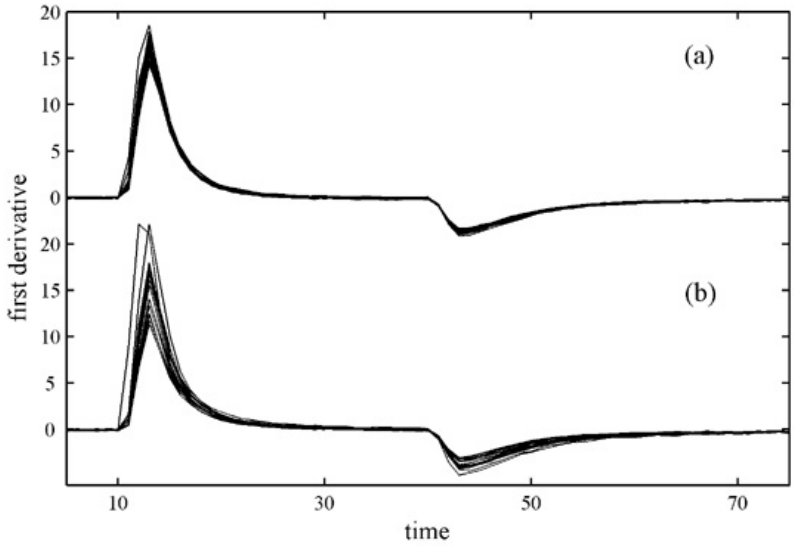

Figure 9.6: First derivative of time profiles of sensors S2. Samples of Garda and not- Garda classes are plotted with a vertical shift: Garda samples are in the upper part (a) and not-Garda samples in the lower part (b).

(Figure 9.7), a manifest class separation is obtained after the sensor selection: in Figure 9.7 the two classes appear better separated on the three components, with the exception of a not-Garda sample placed far from both the two classes. In the previous PCA analysis (Figure 9.2), a trend of separation was present, but not supported by a clear class separation; on the other hand, after the sensor selection, not only the class separation is present, but it appears also clearer, since the two classes are placed in different areas, do not overlap and are not close one to each other. This confirms the class discrimination capability of sensors S9 and S10, selected by means of the similarity/diversity approach based on the Hasse distance.

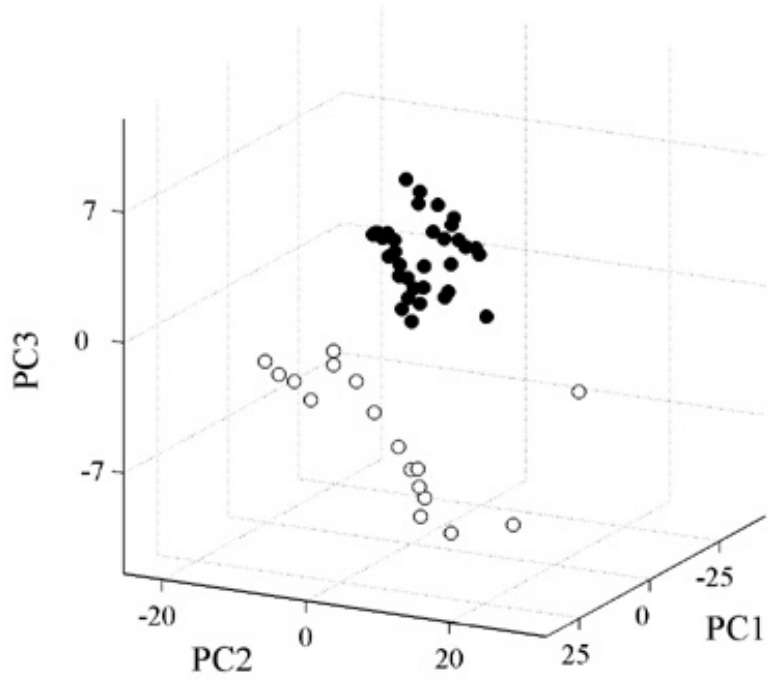

Figure 9.7: PCA on the unfolded data with the selected sensors (S9 and S10): score plot of the first three components (explained variance = 86%). Garda samples are drawn in black and not-Garda samples in white.

9.6 Conclusions

The new proposed approach, based on the partial ordering technique and the Hasse distance analysis, seems to allow the characterisation of chemical fingerprints of electronic nose sensors. In fact, the selection of sensors with more class discrimination capability, obtained by means of the Hasse distance class index, has given interesting results, confirmed by principal component analysis.

Moreover, the proposed approach shows some advantages: (**a**) it seems able to link each electronic nose time profile to a meaningful mathematical term (the Hasse matrix), which can be consequently treated and studied

by multivariate analysis; (**b**) the Hasse matrices and the corresponding distances are calculated with a simple algorithm; (**c**) the Hasse distance is standardised, allowing a natural interpretation of the results; (**d**) the distances consider the whole time profile, i.e. no parameterisation is needed and the time information is preserved; (**e**) the distances can be obtained by a flexible strategy (the weights) depending on the aim of the analysis.

chapter 10

Applications on electronic sensors

In the present chapter, two different chemometric applications on electronic sensor data are presented, the first dealing with geographical characterisation of extra virgin olive oil by means of neural networks, the second one dealing with the characterisation of storage conditions of olive oils.

10.1 Geographical characterisation by means of neural networks

10.1.1 Introduction

The quality and uniqueness of specific extra virgin olive oils (EVOOs) is the result of different factors such as cultivar, environment and cultural practises. Moreover, an important act of legislation [Community (1992)] allows the European Protected Denomination of Origin (PDO) labelling of some European EVOOs with the names of the areas where they are produced. This designation guarantees that the quality of the product is closely linked to its geographical origin.

PDO olive oils are considered the best among EVOOs on the basis of their authenticity and specified organoleptic characteristics. As a consequence, PDO olive oils have a much higher market price and therefore are subjected to frauds: the addition of cheap oils and/or the marketing of oils from one region as those from another [Salter *et al.* (1997)]. Consumers are also more and more oriented towards purchasing food products of a certified genuineness and geographical origin. Detailed percentages of specified cultivar olives, cultural practises, circumscribed geographical production areas, chemical and sensorial properties are required in order to obtain the PDO label, as indicated in the Production Disciplinary.

However, at present no analytical parameters exist that enable the PDO oil to be distinguished from similar products of other regions. The development of methods for the classification of oils is becoming very important for the assignment of a denomination of origin trademark. Since official analysis of virgin olive oils involves a series of several determinations of chemical and physical constant that will be of scarce use in the geographical certification of the oil samples, reliable methods of geographical origin authentication of oil are essential.

In recent years, several attempts have been made in order to verify the declared geographical origin of olive oils by means of suitable chemical parameters, such as triglyceride and fatty acid profiles [Tsimidou *et al.* (1987)] or by means of NMR spectroscopy [Mannina *et al.* (2001)]. These techniques usually require time-consuming measurements, sample preparation and a qualified staff. Consequently, there is the necessity of quick and simple methods to characterise the origin of PDO extra virgin olive oils. Actually, electronic nose and electronic tongue, combined with pattern recognition techniques, offer a fast, simple and efficient tool for classification purposes.

The electronic nose consists of an array of gas sensors, a signal collecting unit and a pattern-recognition software [Gardner and Bartlett (1993)]. The principle involved in the electronic nose is the transfer of the total headspace containing different chemical volatile compounds to a sensor array, where each sensor has partial specificity to a wide range

of aroma molecules. In fact in the headspace of virgin olive oils different volatile compounds are present (such as aldehydes, alcohols, khetones, esters, ethers, etc.). In literature, there are several examples that demonstrate the possibility of using an electronic nose for the characterization of vegetable oils [Cerrato Oliveros *et al.* (2002), Stella *et al.* (2000)] and for the quality control of olive oil aroma [Guadarrama *et al.* (2001)].

The principle of the electronic tongue is similar to that of the electronic nose, except for the array of sensors, which is designed for liquids. There are various techniques that can be used for an electronic tongue, such as conductimetric, potentiometric or electrochemical techniques. In the present paper, the electrochemical technique has been used, where the sensor array measures the current at fixed potentials and provides a pattern of signals. In the literature attempts of using electronic tongue for foodstuffs analysis have been reported [Gallardo *et al.* (2005), Legin *et al.* (2003)], but no study on olive oils have been done. Otherwise, electrochemical sensors can detect some redox active compounds, such as poliphenols and tocopherols, present in EVOOs [Mannino *et al.* (1999), Campanella *et al.* (1999)].

Contrary to traditional analytical methods, electronic nose and electronic tongue sensor responses do not provide information on the nature of the compounds under investigation, but only give a digital fingerprint of the food product that can be investigated by means of multivariate statistical analysis. In fact, multivariate analysis proved to be a powerful tool for determining the geographical origin of food and is especially able to handle the large amounts of data produced by modern analytical techniques. Moreover, multivariate analysis has been successfully applied on olive oil data [Eddib and Nickless (1987), Armanino *et al.* (1989a), Tsimidou and Karakostas (1993), Guimet *et al.* (2005), Guimet *et al.* (2004)].

In this application, EVOO Garda Bresciano has been considered: this is a product made in the Garda lake, a circumscribed area in the north Italian region of Lombardia and distinguished as PDO since 1998. An electronic nose and an electronic tongue have been used to characterise the geographical origin of Garda EVOOs. Classical chemical parameters

have also been determined, in order to evaluate the olive oil quality. Total phenols, important for organoleptic and antioxidant characteristics, have been evaluated to verify a possible correlation with the geographical origin of oil samples. A classification model has been built by means of Counterpropagation Artificial Neural Networks (CP-ANN) in order to differentiate 36 EVOOs of Garda, from 17 EVOOs of several regions of Italy and Europe. CP-ANN are a well-known classification technique, already used for the characterisation of oil data [Goodacre *et al.* (1992), Zupan *et al.* (1994)]. A classification model with two classes (Garda and not-Garda) can represent an appropriate tool to describe PDO Garda EVOO uniqueness and to develop a useful method for the protection of Garda PDO against adulteration. For this purpose, the classification model has been also tested with 19 commercial EVOOs. Counterpropagation Neural Network results have been analysed by means of Principal Component Analysis, in order to select the variables with a real discriminant rule and to improve the previous classification model.

10.1.2 Oil samples

The data set has included 53 samples of monovarietal EVOOs obtained from several olive cultivars and grown in five different regions: Garda, 36 samples; Spain, 6 samples; Sardegna, 5 samples; Campania, 4 samples; Abruzzo, 2 samples. Olives have been harvested in the period from November to December 2003. Once collected the olives have been washed and processed using a micro-oil press equipped with an hammer crusher, a vertical mixer and a two phase decanter (Alfa Laval, Firenze, Italy). The sampling has included also 19 commercial and multivarietal EVOOs: 3 samples labelled as Garda PDO, produced with cultivars (55% of Casaliva, Leccino and Frantoio and 45% of other Garda cultivars) allowed by the Garda Production Disciplinary (Decreto Ministeriale, 1998, Settembre 17); three samples of Garda, not labelled as PDO; 13 samples collected on the market, produced with unknown cultivars. All the commercial samples have been used only to test the classification model. For our purposes,

Table 10.1: Origin of extra virgin olive oil samples: number of samples, class and rule in the classification model are reported for each origin group.

Origin	**Samples**	**Class**	**Traing/test**
Abruzzo	2	Not-Garda	Training
Garda	36	Garda	Training
Campania	4	Not-Garda	Training
Sardegna	5	Not-Garda	Training
Spain	6	Not-Garda	Training
Commercial	6	Garda	Test
Commercial	13	Not-Garda	Test

all the samples have been divided into two classes: Garda class and not-Garda class (Table 10.1). A refined olive oil was purchased from a local market and used as internal standard for the calibration method of the electronic nose. This standard was still frozen and did not alter during the whole time of the experiments.

10.1.3 Chemical analyses

The chemical analyses have included the measurement of several parameters. The free acidity (FA), which is indicative of the free fatty acid content of the oil expressed as oleic acid (%); the peroxide value (PV), which is a measure of the amount of hydroperoxides (mequiv. O2/kg) formed through autoxidation during storage; the absorbances UV (K_{232}, K_{270}, and ΔK), which provide a measurement of the state of oxidation of the oils.

A commercial electronic nose has been used (model 3320 Applied Sensor Lab Emission Analyser, Applied Sensor Co., Linkoping, Sweden), comprising three parts: automatic sampling apparatus, detector unit containing the array of sensors, and software for data recording [Cosio *et al.* (2006)]. The automatic sampling system supports a carousel of 12 sites for loading the samples and permits the control of internal temperature. Twenty-two different sensors compose the sensor array: 10 sensors are Metal Oxide Semiconductor Field Effect Transistors (MOSFET) and 12

Table 10.2: List of the variables considered in the experimentation. The number of variables (chemical analyses, phenols, electronic nose and electronic tongue) and the variable code are reported.

	No. Var	Variables	Code
Chemical	5	free acidity (%)	FA
		peroxide value (meq/kg)	PV
		Absorbance UV	K_{232}, K_{270}, ΔK
		at 232 nm, 270 nm and ΔK	
Phenols	1	total phenol	TP
E-nose	22	10 MOSFET sensors	FE
		12 MOS sensors	MO
E-tongue	3	3 Carbon electrode (+0.5,	P500, P600, P800
		+0.6,+ 0.8 V(vs Ag/AgCl)	

are Taguchi type sensors (Metal Oxide Semiconductors MOS).

Regarding the electronic tongue, a measurement system based on flow injection analysis (FIA) with two amperometric detectors has been set up. The FIA apparatus consist of a Jasco (Tokyo, Japan) model 880 PU pump and two EG&G Princeton Applied Research (Princeton, NJ, USA) Model 400 thin-layer electrochemical detectors connected in series. Each detector is equipped with a working electrode (a dual and a single glassy carbon electrode, respectively), a reference (Ag/AgCl saturated) electrode and a platinum counter electrode. Data are recorded using a Philips (Eindhoven, The Netherlands) PM 8252 recorder. In the flow system, a carrier solution is continuously pumped through the amperometric detectors and the samples are injected into the flow stream. The amperometric detectors permit the oxidation of electroactive compounds at the working electrode while a constant potential is applied.

10.1.4 Data analysis

A data matrix with 53 rows (EVOO samples) and 31 columns (chemical variables, total phenols, electronic nose and electronic tongue sensor signals) has been built. The variable description is provided in Table 10.2. Samples have been divided in two classes (Garda and not-Garda) and

Counterpropagation Artificial Neural Networks (CP-ANN) have been applied on this dataset in order to find a predictive classification model. The original variables have been treated by scaling in the 01 range in order to make them comparable with the networks weights.

The classification model has been validated using a cross validation procedure. Eleven samples, which correspond to the 20% of the total number of samples, have been randomly removed from the data set. The classification model has been rebuilt without these objects and the removed sample have been classified in the new model. This procedure has been repeated 100 times, always with a random selection of 11 removed samples. Finally, a percentage of correct classification has been calculated. The quality of the classification model has been considered on the basis of the validation results. Moreover, the final classification model has been tested also with 19 commercial oil samples. These samples have not been used to build the final classification model, but just to test it. In fact, the aim of the model is the classification of new commercial oil samples and these samples have been used to test this purpose.

In order to examine the relationship between variables and classes, the weights of the Kohonen layer have been analysed by means of Principal Component Analysis (PCA). The weights of the Kohonen layer can be seen as a data matrix $\mathbf{W}$ with r rows and p columns, where r is the number of neurons, p is the number of variables. Therefore, each element w_{ij} of the matrix $\mathbf{W}$ is the weight of the j-th variable in the i-th neuron. By applying PCA on $\mathbf{W}$, a loading matrix and a score matrix are obtained: the loading matrix has dimension p times f, where f is the number of significant principal components, while the score matrix has dimension r times f. In this way, the relationship between variables and neurons can be shown: since each neuron is assigned to a class (or unclassified) on the basis of the weights of the output layer, the relationship between variables and classes can be highlighted. The Kohonen weights have values in a range from 0 to 1, therefore no scaling is needed and PCA has been performed without a data pretreatment. Principal Component Analysis has been performed using the statistical package SCAN SCAN (Minitab,

State College, PA), and CP-ANN using KOALA (Milano Chemometrics and QSAR Research Group, Milano-Bicocca University,).

10.1.5 Results

The olive oils samples have widely respected the EC Regulation limits, consequently they can be considered as EVOOs. Total phenols, the responses obtained by the electronic nose (22 sensors), the electronic tongue signals (three sensors) in addiction to the chemical parameters (five variables) have been considered all together for multivariate statistical analysis.

In order to find a predictive classification model able to characterise the EVOOs on the bases of their geographical origin, the 53 oil samples (Table 10.1) have been used to build and cross validate a Counter Propagation Artificial Neural Network, while the 19 commercial oil samples (Table 10.1) have been used to test the classification model. The purpose of the classification model is the differentiation of Garda and not-Garda oils. For the CP-ANN parameter optimisation, several networks have been evaluated, by varying the number of neurons (from 36 to 81) and training epochs (from 20 to 300). The best classification model has been given by a CP-ANN with 77 neurons and trained with 200 epochs.A Mean Error Rate in prediction equal to 1% has been obtained, and two test samples have not been assigned to a class (unclassified). Therefore, the selected CP-ANN appears able to describe the classes, since the model classifies the samples with significant predictive performances.

The Kohonen map or Top map, permits to analyse the sample behaviour and the relationship between samples and classes. On the Top map shown in Figure 10.1 the two classes have been drawn with different marks and the samples labelled with their identification numbers. It is important to remark that the Kohonen map is a toroidal space, i.e. the neuron in the bottom-left corner of the graphic is close to the ones in the opposite corners, and that samples are usually inclined to spread over the whole map. Garda and not-Garda classes have been well separated in the Kohonen map, since all the neurons around the class spaces are empty

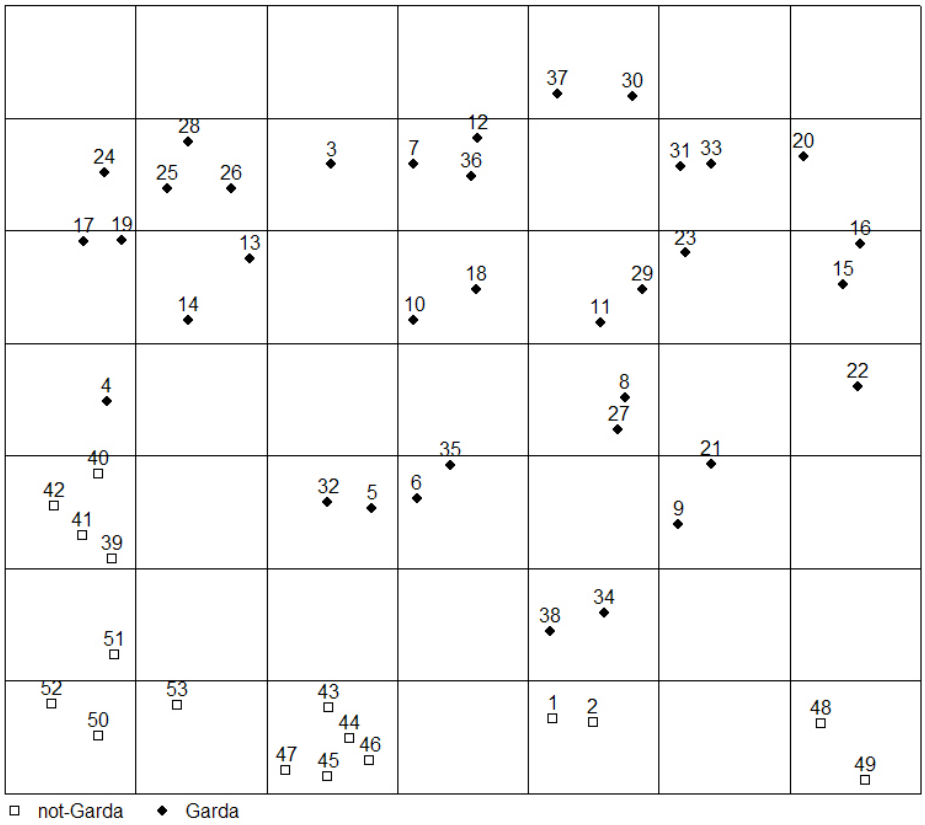

Figure 10.1: Counterpropagation Neural Network with all the variables: Kohonen map, trained with 49 neurons and 200 epochs. Garda samples are drawn in black.

and without samples, with the exception of two neurons. In fact, as can be seen in Figure 10.1, samples no. 1 and 2 belong to not-Garda class but are closer to the Garda oils than all other not-Garda oils. The same happens for sample no. 4, which belong to the Garda class and it is close to four not-Garda samples (no. 39, 40, 41, and 42). Nevertheless, samples no. 1, 2 and 4 have been well classified, due to the CP-ANN properties. All other samples of the two different classes are at least separated by one empty neuron. Since the goal of the model is the classification of new commercial samples, the Counterpropagation Neural Network has been tested

with 19 commercial oils. Even in this case, the model has shown a good behaviour: the six Garda commercial samples have been properly classified in accord with their geographical origin; two not-Garda commercial samples (no. 61 and 68) have been assigned to none of the two classes by the neural network, while all the other 11 not-Garda commercial oils have been correctly classified as not-Garda. This result could suggest how these methods are suitable for the geographical recognition of real oil samples.

In order to investigate the relationship between variables and classes, the Kohonen weights should be analysed variable by variable.With the aim of facilitating this analysis and obtaining a multivariate overview of variable behaviour in the classification model, the weights of the Kohonen layer have been analysed by means of Principal Component Analysis. In this case, the complete score and loading matrices have dimensions 4931 and 3131, respectively. The score plot for components 1 and 2 are shown in Figure 10.2.

The points in this plot represent the 49 Kohonen neurons: since each neuron can be classified in a class or not classified, the points have been marked in function of their assignments. As can be seen in the score plot, neurons of different classes are well described by components 1 and 2, which explain the 77% of the total variance. Since these components have separated the two classes, it has been possible to show which variables describe the Garda and not-Garda class in the classification model. Looking at the loadings of components 1 and 2 (Figure 10.3), it is clear how some chemical parameters do not have a relevant rule in the class discrimination: K_{270} is placed in the middle of the plot, close to the axis origin; PV and ΔK have a small influence on both components, while free acidity (FA) and K_{232} are relevant on the second component. The same conclusion can be done on the total phenols (TP). The electronic tongue features (P500, P600, P800) have a relevant rule only on the second component, while the electronic nose sensors have globally high loadings on both components. Since the two classes can be easily separated by a linear combination of the two first components, the electronic nose sensors appear appropriate for the characterization of the two classes.

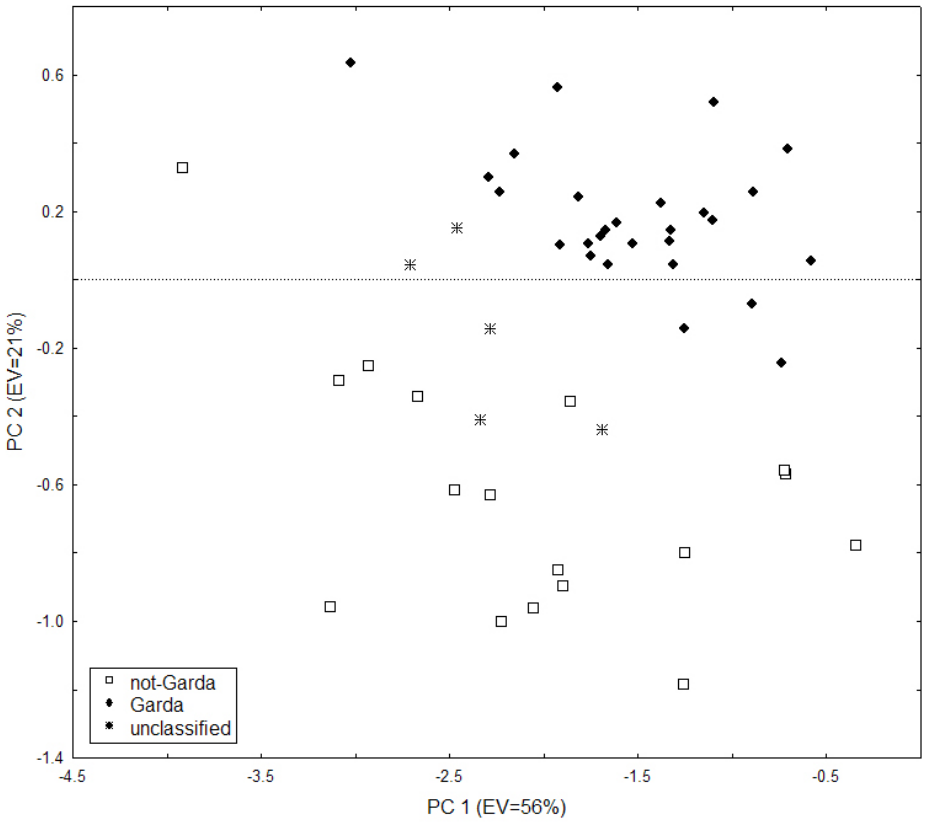

Figure 10.2: PCA on Kohonen weights: score plot between components 1 and 2. Garda, not-Garda and unclassified neurons are shown with different shapes.

Even if the previous neural network has performed a good classification result, the possibility of using only electronic tongue or electronic nose sensors for the sample classification has been considered. In this way, the classification tool can be simplified and improved, and the electronic sensors have been chosen since they are more reliable, cheap and fast with respect to the considered chemical analyses. Initially, CP-ANNs have been built by using the three electronic tongue variables. The best classification model has been obtained again with a CP-ANN with 77 neurons and trained with 100 epochs. The Mean Error Rate in prediction has been

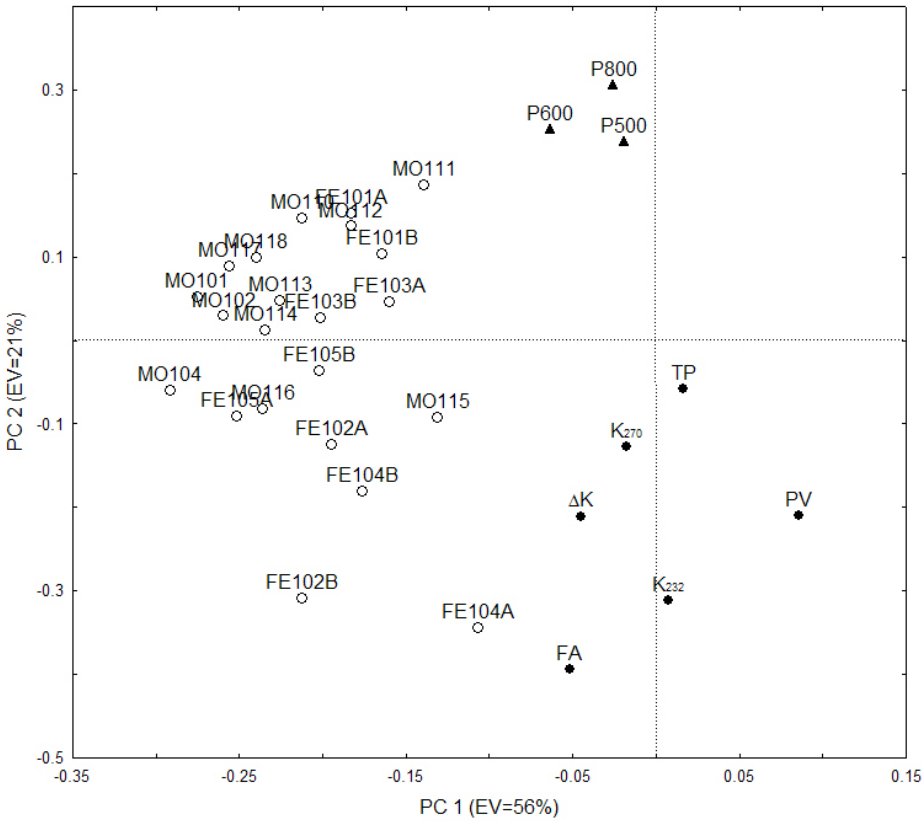

Figure 10.3: PCA on Kohonen weights: loading plot between components 1 and 2. Electronic nose sensors are marked with white circles; electronic tongue sensors with black triangles; chemical variables and total phenols with black circles.

equal to 4.2%, one test sample has been assigned to the wrong class, while four test samples have not been assigned to a class (unclassified), consequently the obtained classification model has performed worse than the previous one. With respect to the electronic nose, four sensors (FE104A, FE102B, MO104, MO111) have been selected, since they have a relevant rule in the loading plot (Figure 10.3). CP-ANNs have been built by using only these four variables. The best classification model has been obtained again with a CP-ANN with 77 neurons and trained with 200 epochs. The Mean Error Rate in prediction has been equal to 0.2%, and all the test

objects have been correctly assigned. On the basis of these results, the classification model built with the electronic nose selected variables appears better than the model with all the variables, since both error rate and unclassified samples have been reduced. The neural network output weights for the 53 training samples rank from 0 to 1 (Table 10.3), therefore these values can be seen as class membership probabilities: since the true class membership probability for the major part of the training samples is equal to 1, this result also shows the good classification performance of the neural network.

Moreover, Garda and not-Garda classes have been well separated in the Kohonen map (Figure 10.4). All the samples of the two classes are at least separated by one empty neuron, with the exception of samples no. 1 and 2 (as in the previous CP-ANN) and sample no. 10. The neural network built with the selected variables has also been tested with the commercial oil samples.

Even in this case, the new model has shown a better behaviour than the previous ones: the six Garda commercial samples have been properly classified in accord with their geographical origin, while all the not-Garda commercial oils have been correctly classified as not-Garda. On the contrary of the previous models, no test samples have been unclassified now. In Table 10.4, a comparison between the classification model with all the variables and with the four electronic nose sensors is given, by showing the Garda class membership probability of the 19 test samples.

All the Garda commercial oils (samples no. 5459) have been perfectly classified as Garda samples with a probability equal to 1 in the CP-ANN built with the four selected variables, while the sample no. 58 had been classified with a lower probability (0.57) in the CP-ANN built with all the 31 variables. Furthermore, the variable selection has also permitted to globally reduce the Garda membership probabilities of the not-Garda commercial oil samples. As it can be seen, samples no. 60, 62, 69 have increased their probabilities, samples no. 64, 70, 72 have not increased the probabilities in a significant way: for example sample 64 has a Garda probability of zero with all the variables and 0.01 with the selected vari-

Table 10.3: Garda class membership probability for the 53 training samples in the CP-ANN with the selected electronic nose variables.

sample	class	prob	sample	class	prob
1	not-Garda	0.00	28	Garda	1.00
2	not-Garda	0.00	29	Garda	1.00
3	Garda	1.00	30	Garda	1.00
4	Garda	1.00	31	Garda	1.00
5	Garda	1.00	32	Garda	1.00
6	Garda	1.00	33	Garda	1.00
7	Garda	1.00	34	Garda	0.97
8	Garda	1.00	35	Garda	1.00
9	Garda	1.00	36	Garda	1.00
10	Garda	0.97	37	Garda	1.00
11	Garda	1.00	38	Garda	1.00
12	Garda	1.00	39	not-Garda	0.00
13	Garda	1.00	40	not-Garda	0.00
14	Garda	1.00	41	not-Garda	0.00
15	Garda	1.00	42	not-Garda	0.00
16	Garda	1.00	43	not-Garda	0.00
17	Garda	1.00	44	not-Garda	0.00
18	Garda	1.00	45	not-Garda	0.00
19	Garda	1.00	46	not-Garda	0.00
20	Garda	1.00	47	not-Garda	0.00
21	Garda	1.00	48	not-Garda	0.00
22	Garda	1.00	49	not-Garda	0.00
23	Garda	1.00	50	not-Garda	0.00
24	Garda	1.00	51	not-Garda	0.00
25	Garda	1.00	52	not-Garda	0.07
26	Garda	1.00	53	not-Garda	0.00
27	Garda	1.00			

ables, but this increasing is not significant, since probabilities range from 0 to 1. On the other hand samples no. 61, 63, 66, 68 have decreased their probabilities. Looking at the total difference of not-Garda sample probability (-0.4), i.e. the sum of the differences between the probabilities achieved with all the variables and with the selected variables, it can be seen that probabilities are higher by using all the variables.

In conclusion, electronic nose combined with neural networks could represent a reliable, cheaper and faster classification tool, able to verify

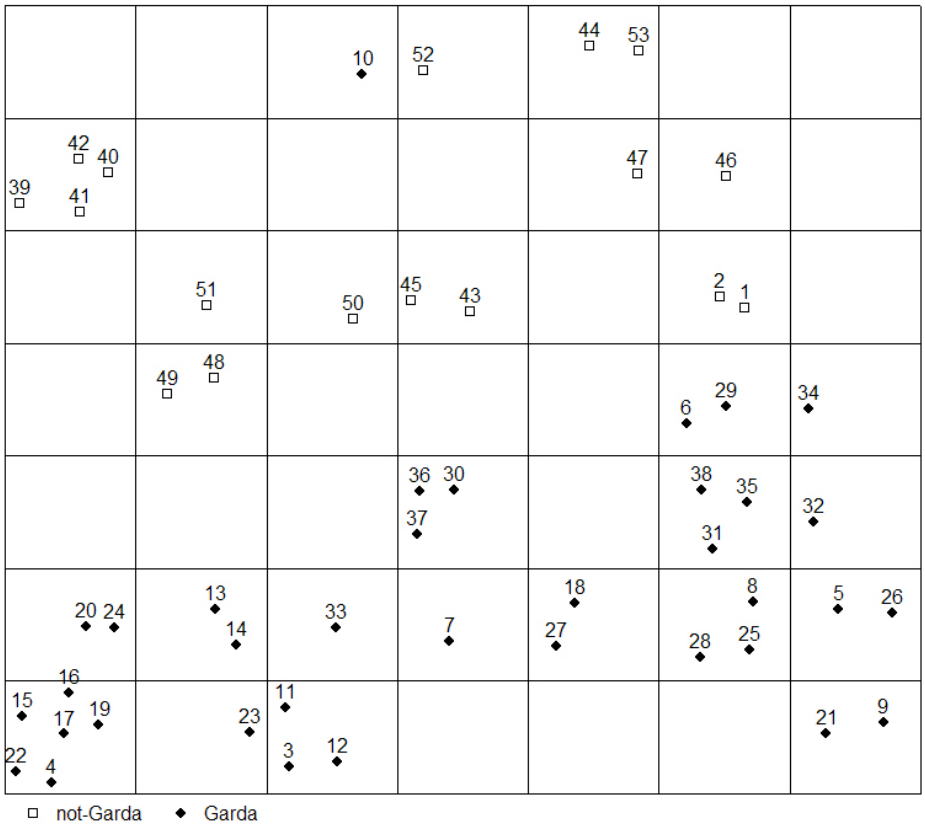

Figure 10.4: Counterpropagation neural networks with four electronic nose variables: Kohonen map, trained with 49 neurons and 200 epochs. Garda samples are drawn in black.

the geographical origin of extra virgin olive oils from a restricted area. The classification model built with the selected electronic nose sensors appears better than the model built with all the variables, both for cross-validation and external validation results. Finally, the classification model built with the electronic tongue sensors gives acceptable results, but its prediction performances are worse then the performances obtained by means of the electronic nose.

Table 10.4: Garda class membership probability for the test objects in the CP-ANN with all the variables and in the CP-ANN with the selected electronic nose variables. Differences in the results are also shown.

sample	class	Sel Var	All Var	Diff
54	Garda	1.00	1.00	0.00
55	Garda	1.00	1.00	0.00
56	Garda	1.00	1.00	0.00
57	Garda PDO	1.00	1.00	0.00
58	Garda PDO	1.00	0.57	0.43
59	Garda PDO	1.00	1.00	0.00
	Garda			0.43
60	not-Garda	0.38	0.14	0.24
61	not-Garda	0.07	0.49	-0.42
62	not-Garda	0.38	0.00	0.38
63	not-Garda	0.01	0.39	-0.38
64	not-Garda	0.01	0.00	0.01
65	not-Garda	0.07	0.09	-0.03
66	not-Garda	0.01	0.14	-0.14
67	not-Garda	0.00	0.00	0.00
68	not-Garda	0.38	0.49	-0.11
69	not-Garda	0.15	0.09	0.05
70	not-Garda	0.01	0.00	0.01
71	not-Garda	0.07	0.09	-0.03
72	not-Garda	0.01	0.00	0.01
	not-Garda			-0.41

10.2 Evaluation of different storage conditions of olive oil

10.2.1 Introduction

Extra virgin olive oil is properly processed from fresh and mature high quality olives and presents a complex flavour. Flavour is usually divided into the subsets of aroma and taste, which are perceived in the nose and in the mouth, respectively. Many authors in fact have clearly demonstrated that the flavour is mainly produced by volatile and phenol compounds [Flath *et al.* (1973)], most of which have been identified and quantified in different extra virgin olive oils [Tsimidou *et al.* (1992), Vichi *et al.*

(2003)]. Lipolysis and oxidation are the processes leading to the most serious deterioration of olive oil. Lipolysis usually starts when the oil is still in the fruit, while the oxidation begins at the processing stage and proceeds during storage influenced by exposition to air, heat, light and metals. Though extra virgin olive oil is considered to be a stable oil due to the presence of natural antioxidants, it is also susceptible to oxidation. Volatile compounds are the main responsible of the pleasant flavour and change in off-flavours during the storage [Angerosa *et al.* (1999), Morales *et al.* (1997)]. At present the classical methods used to ascertain extra virgin olive oil quality are based on chemical analysis [Community (1991)] and sensory analysis [Community (2002)].

However, these methods are expensive and time consuming. Recently HPLC/GCMS was applied to detect changes in the chemical composition of olive oil during the storage. HPLC with different detection systems has been used for hydroperoxide analysis [Oshima *et al.* (1996)]. GC-MS was used to detect hydroxy fatty acids and volatile compounds originated from hydroperoxide degradation [Morales *et al.* (1997)]. Each of these analyses only gives partial information about the extent of oxidation and there is a large demand for rapid, cheap and effective techniques for quality control of extra virgin olive oils. In recent years, considerable efforts have been devoted to the development of innovative analytical instrumentation such as the electronic nose and electronic tongue, which can mimic the human sense of olfaction and of taste and provide low-cost and rapid sensory information for monitoring food quality and state of a process.

The electronic nose consists of an array of gas sensors with different selectivity, a signal collecting unit and a suitable pattern recognition software. It is particularly useful for the quality control in food or beverage production for monitoring flavour changes [Bartlett *et al.* (1997)].

In the literature, there are several examples that demonstrate the possibility of using an electronic nose for the characterization of vegetable oils [Gonzalez Martin *et al.* (1999)] and for the quality control of olive oil aroma [Guadarrama *et al.* (2000)], while information about the use of an electronic nose to predict shelf life of vegetable oils or to monitor oil

oxidation under real life storage conditions are not frequent.

The principle of the electronic tongue is similar to that of the electronic nose, except for the array of sensors, which is designed for liquids. Many publications report the application of the array of electrochemical sensors for beverage analysis and wine discrimination [Gallardo *et al.* (2005), Legin *et al.* (2003)]. Olive oils contain some redox active compounds such as polyphenols, tocopherols, etc. that have an important relevance in their organoleptic characteristics and antioxidant properties and could be analysed by means of electrochemical sensors [Campanella *et al.* (1999), Mannino *et al.* (1999)].

The aim of the present research is to show how non destructive techniques in combination with multivariate statistical analysis can represent an effective device for the evaluation of the oxidative status of an extra virgin olive oil. Furthermore, the study has been conducted with the use of real life storage conditions and not by applying an accelerated thermoxidation process. In comparison to classical techniques, this approach could represent a faster and cheaper recognition tool for monitoring oil oxidation.

10.2.2 Oil samples

Fresh and mature olive samples were collected during years 2001 and 2002, from different cultivars and different Italian areas. From these olives, 61 monovarietal samples of extra-virgin olive oil were properly obtained using a micro-oil press equipped with a hammer crusher, a vertical mixer and a two phase decanter (Alfa Laval). Each 2002 oil sample was divided in two aliquots: the first aliquot was stored under normal light, the second one under dark for one year. Instead, each 2001 oil sample was stored for two years under dark. All the samples were left at room temperature in 200 ml amber and transparent glass bottles (dark and light condition).

Table 10.5: List of the variables considered in the experimentation.

	No. Var	Variables	Code
Chemical	5	free acidity (%)	FA
		peroxide value (meq/kg)	PV
		Absorbance UV	K_{232}, K_{270}, ΔK
		at 232 nm, 270 nm and ΔK	
E-nose	22	10 MOSFET sensors	FE
		12 MOS sensors	MO
E-tongue	3	3 Carbon electrode (+0.5,	P500, P600, P800
		+0.6,+ 0.8 V(vs Ag/AgCl)	

10.2.3 Chemical analysis

The chemical analysis included the measurement of several parameters. The acidity (acidity), which is indicative of the free fatty acid content of the oil expressed as oleic acid (%); the peroxide value (PV), which is a measure of the amount (meqO2/kg) of the hydroperoxides formed through oxidation during storage; the absorbances UV at 232 and 270 nm (K_{232}, K_{270} and ΔK) provide a measurement of the state of oxidation of the oils. Regarding the used settings and apparatus for the analysis conducted by means of electronic nose and tongue, more details are given in [Cosio *et al.* (2007)].

10.2.4 Data analysis

A data matrix with 61 rows (oil samples) and 30 columns (variables) was built. The variable description is provided in Table 10.5. Initially, this matrix was analysed by means of principal component analysis (PCA), in order to display the structure of the multivariate data. Since the variables have been measured in different units, the original variables were autoscaled. Afterwards, samples were divided in three classes (Table 10.6) and linear discriminant analysis (LDA) was applied in order to find a predictive classification model. The quality of the LDA classification model was considered on the basis of the validation results. Principal component analysis was performed by using the statistical package SCAN (Minitab

Table 10.6: Class definition. The storage condition, the storage period and the number of samples of each class are reported.

class code	storage	period	samples
class 1	dark	1 year	16
class 2	light	1 year	16
class 3	dark	2 years	29

Inc., PA, 1995), and linear discriminant analysis by using SPSS (Inc., Chicago, 2004).

10.2.5 Results

The quality of extra-virgin olive oils was ascertained with the following analytical parameters: acidity, PV, K_{232}, K_{270} and ΔK. As suggested by Regulation EEC/ 2568/91, these parameters are valuable olive oil freshness indices and the following limits for extra-virgin olive oils are established: acidity 6 0.8, PV 6 20, K_{232} 6 2.4, K_{270} 6 0.22 and ΔK 6 0.01 [Community (2003)]. The 61 oil samples, analysed before the storage, widely respected the limits of the before mentioned Regulation, confirming a good overall quality: these oils could be labelled as extra-virgin according to the European legislation.

Then the samples were analysed after 1 year of storage under dark (class 1), under normal light (class 2), and after 2 years under dark (class 3). All samples of the three classes presented an acidity value lower than 0.4, PV from 16 to 20 (class 1), from 17 to 61 (class 2) and from 17 to 39 (class 3). Except for class 1, most of the samples of class 2 and class 3 had UV values higher then the law limits.

At the end of their storage period, all the oil samples were also analysed with alternative and innovative techniques (electronic nose and electronic tongue). The responses obtained with the electronic nose (22 sensors) and the electronic tongue (three sensors) together with the classical chemical determinations (five parameters) calculated at the end of the sample storage period were considered all together and used for statistical analysis.

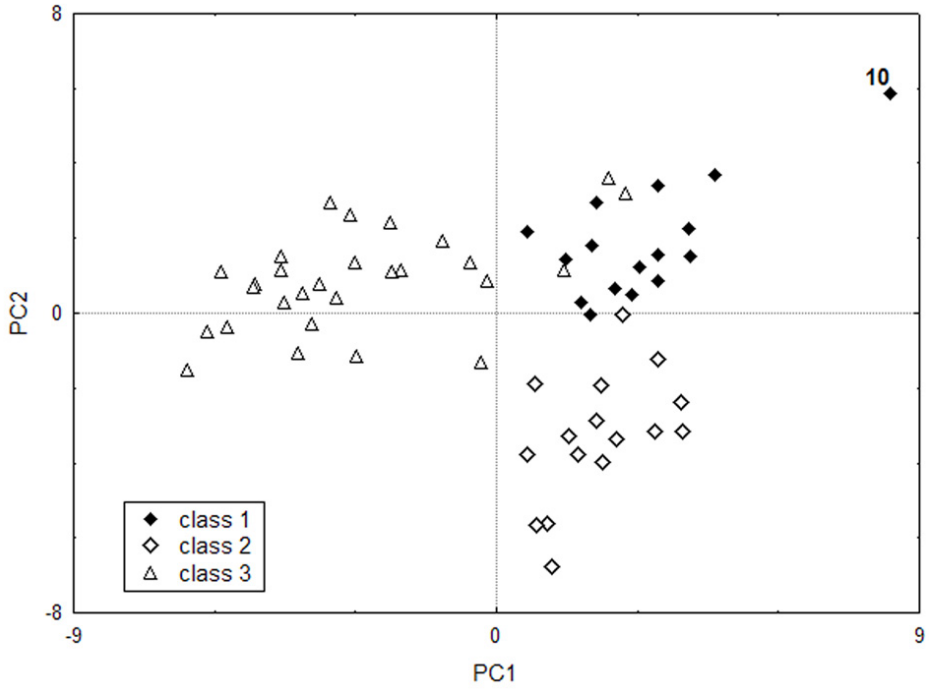

Figure 10.5: PCA on autoscaled data: score plot. Classes are shown with different symbols.

Initially, the data matrix with 61 rows (oil samples) and 30 columns (variables) was analysed by means of PCA, in order to study how the different storage conditions characterised the oil samples. The first principal component and the second principal component were enough to display the data structure, since they explained 61% of the total variance. Examining the score plot (Figure 10.5) in the area defined by the first two principal components, a separation of the samples into three groups was found according to the different storage conditions and storage periods. Only few samples belonging to class 3 were projected in the middle of class 1, but this does not affect the effectiveness of the plot.

Furthermore, on the basis of the position of each group in the plot, it was possible to assign a particular meaning to each component. The first component was able to separate the oil samples belonging to class

3 (characterised by negative values) from all other samples, i.e., the first component was able to characterise the samples on the basis of the storage period. In fact, samples belonging to class 3 were characterised by a storage period of two years, while all the other samples by a storage period of only one year. On the second principal component, oil samples of class 2 had negative values, while all other samples had positive values, i.e., the second principal component was able to describe the samples on the basis of the storage conditions. In fact, class 2 samples were stored under light, while all other samples under dark.

Finally, a sample belonging to class 1 appeared far from its class space in the score plot. The highest scores on the first and the second component characterised this sample, labelled in the score plot as sample no. 10. As described before, the meaning of each component is related to the quality of the storage period and conditions. The highest positive scores on the two components were associated to the best storage situation, i.e., conservation under dark for one year. The behaviour of sample no. 10 confirmed this hypothesis: in fact, all the values of classical chemical parameters for this sample respected widely the law limits and allowed it to be considered as a extra-virgin olive oil. All other samples of class 1 could be considered as extravirgin olive oils but presented PV and UV values near the law limits.

Since samples were well described in the score plot, the loading plot was analysed in order to show which variables influenced the group separation. As can be seen in the loading plot of the first two principal components (Figure 10.6), the majority of electronic nose sensors characterised the first principal component, while electronic tongue sensors, two sensors of electronic nose (FE101A and FE101B), and the PV were relevant on the second component. First of all, it is clear how the electronic tongue sensors were correlated giving the same information, as expected. The two types of electronic nose sensors (MOSFET and MOS) appeared different, since they grouped in different areas on the second component.

Furthermore, MOSFET sensors appeared more informative, since they showed high loading values on both components. It is important to notice

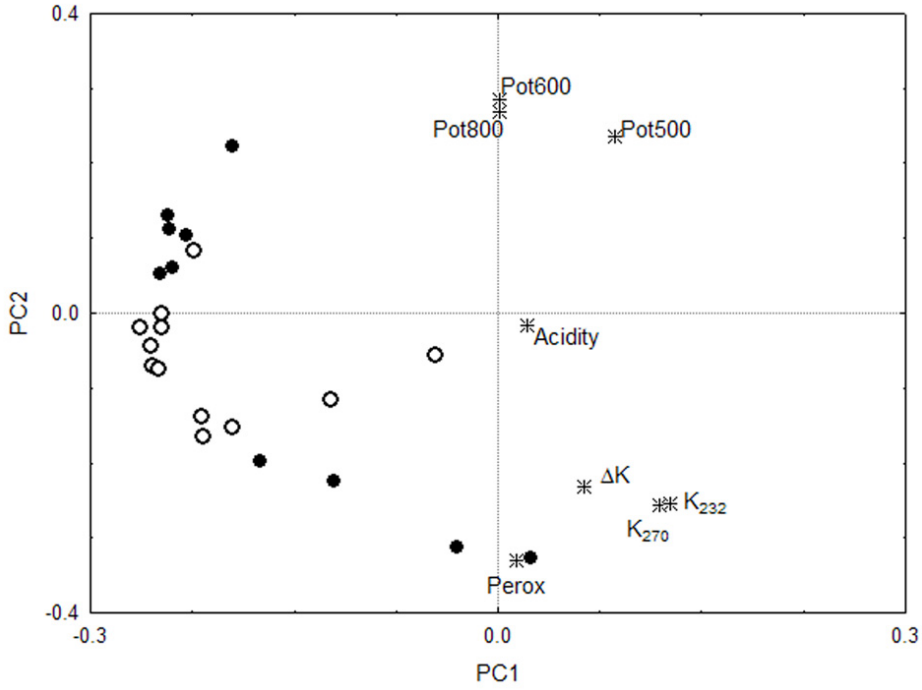

Figure 10.6: PCA on autoscaled data: loading plot. Variables are shown with different symbols: electronic nose MOSFET sensors (dark circle), electronic nose MOS sensors (white circle), electronic tongue sensors and classical chemical variables (white square).

that classical chemical variables did not appear relevant for the description of the samples under study. Acidity was placed in the middle of the loading plot: this variable did not have an impact on the group separation, i.e., in the description of the storage period and the storage conditions. Among classical chemical variables, the PV is the only with a high loading value, but, as can be seen in the loading plot, two electronic nose sensors (FE101A and FE101B) were placed close to it. This means that all these three variables had the same information, i.e., the PV could be removed without a decrease of discrimination capability. In conclusion, electronic nose sensors and electronic tongue sensors gave sufficient information to describe the different storage conditions and storage periods and appeared to be enough for the characterization of the three classes of oil samples.

Table 10.7: Confusion matrix of the LDA classification model with all the variables (fitting and validation results are both reported). Rows represent the true class, columns represent the assigned class.

	Classes	**1**	**2**	**3**	**Total**
Fitting	1	16	0	0	16
	2	0	16	0	16
	3	0	0	29	29
Cross-Validation	1	16	0	0	16
	2	0	16	0	16
	3	0	1	28	29

Since the data structure analysis gave a good sample characterization, a classification model was built. LDA analysis was applied in order to find a predictive classification model, able to separate the three described classes. In Table 10.7, the results of LDA and leave one out cross validation are reported. As can be seen, LDA applied to the complete data set gave a recognition percentage of 100%, while only one oil sample was not correctly classified in the validation procedure. Even if this model performed a good classification result, the classification after selection of a minimum number of variables was also considered.

In fact, the PCA loading plot highlighted how the classical chemical variables did not appear relevant for the class discrimination or appeared correlated to electronic sensors and nose sensors. For this reason and in order to simplify the classification model by reducing the number of the considered variables, LDA was repeated by considering only the electronic nose and electronic tongue features. The classification model gave again 100% correct classification for three classes and only one wrong assignment in the validation procedure. The discriminant scores for the classification model with only the electronic nose and electronic tongue features (Figure 10.7) showed a clear class separation.

More conclusions can be obtained by observing the plot of the standardised canonical discriminant function coefficients (Figure 10.8). In this plot the behaviour and the rule of each variable in the classification model can be analysed. It is clear that few sensors had high canonical discrim-

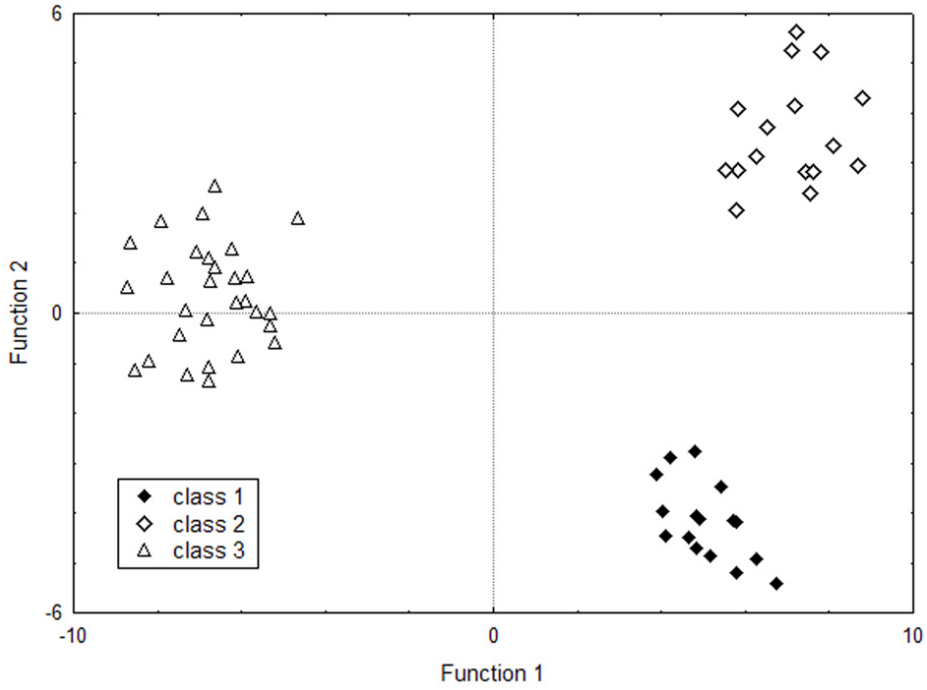

Figure 10.7: LDA classification model with the electronic nose and electronic tongue sensors: discriminant scores. Classes are shown with different symbols.

inant function coefficients and that the electronic tongue sensors did not show a relevant role in the classification model, since they were placed in the middle of the plot, close to the axis origin. This result suggested the possibility of removing the electronic tongue sensors from the model, without a decreasing of its classification capability. Therefore, in order to simplify once more the classification model, LDA was repeated by considering only the electronic nose features. As expected, the classification model gave the same results as before, i.e., a recognition percentage of 100%, and only one wrong assignation in the validation procedure.

Since an equal classification performance was obtained by considering only the electronic nose sensors, it is evident that chemical analyses and electronic tongue sensors were not required in order to achieve a better sample discrimination, i.e., chemical analyses and electronic tongue sen-

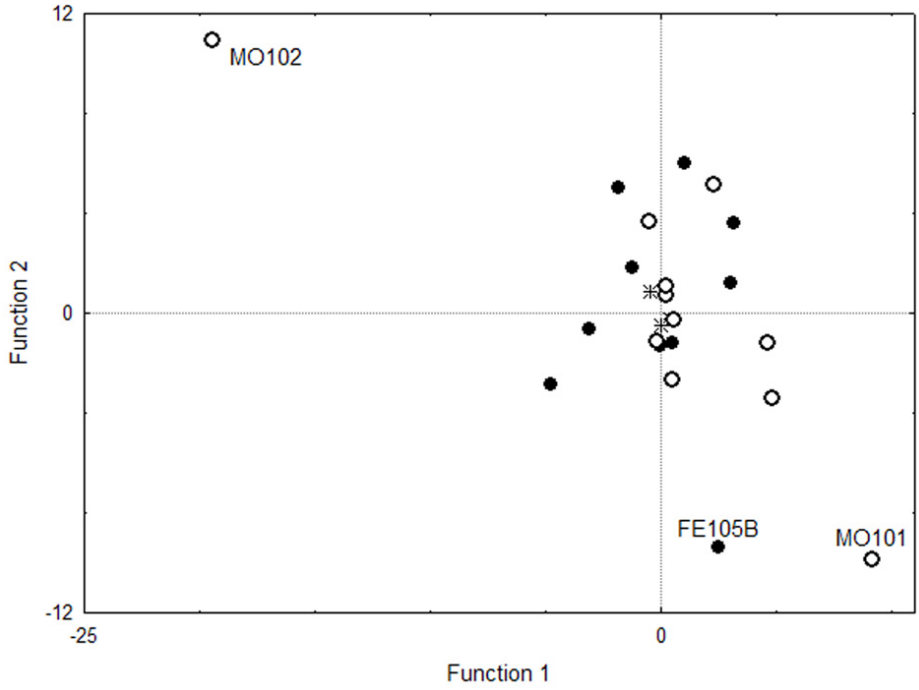

Figure 10.8: LDA classification model with the electronic nose and electronic tongue sensors: standardised canonical discriminant function coefficients. Variables are shown with different symbols: electronic nose MOSFET sensors (dark circle), electronic nose MOS sensors (white circle), electronic tongue sensors (white square).

sors did not improve the classification model.

In conclusion, this study evaluated the possibility of differentiate olive oil samples stored in different conditions and periods by using innovative analytical techniques, such as the electronic nose and electronic tongue, in combination with multivariate analysis. Chemical parameters and electronic tongue did not appear as relevant in the LDA classification model. In fact, it has been showed that by removing chemical analysis and electronic tongue sensors, the classification performance is preserved and a more applicable model is obtained. The final classification model built by means of the electronic nose sensors was able to describe the samples storage conditions and could represent a simpler, faster, and cheaper

recognition tool, since a minor number of variables must be determined. This new approach could offer a valid alternative to the difficult and time-consuming traditional analytical methods and could be a useful tool for on line or routine determination of olive oil storage conditions.

chapter 11

Acoustic and Mechanical data of crispy products

11.1 Introduction

Structured foods are complex and heterogeneous materials whose performances are backwards and straight related to their properties, their structure and, at last, to the production processes. Thermal and mechanical histories along all the manufacturing steps lead to the tailoring of the food structure and to its physical properties, which are finally related to the sensory-appreciated food attributes. The application here presented reports, as an example, the results of a physical instrumental analysis that has been performed in order to objectively quantify the most characterising sensory attributes of dry bakery crispy products, i.e. the combined measure of mechanical and acoustic stimuli that produce textural sensations during mastication [Piazza *et al.* (2006)].

Crispness is synonymous of freshness. For crispy products, the crispness loss due to the adsorption of ambient moisture or due to the water mass transfer from neighbouring components is a major cause of rejec-

tion by the consumers. Most low moisture baked or extruded products (biscuits, crackers, bread-sticks, toasted sliced breads, etc.) are predominantly glassy starch based materials and have a crispy texture. They are generally described as brittle cellular materials [Roudaut *et al.* (2002)]. When a force is applied to a crisp item, its structure is stressed until a critical point is reached: the action of external force causes the rupture of the brittle walls of the cellular structure which start to vibrate. The vibration is transmitted trough the air as acoustic waves, which generates the sound. Sensory crispness is therefore the perception of deformation and time events but, almost and primarily, of their simultaneous acoustical effects.

Sensory methods are the primary means of measuring the range of textural characteristics of food that are important to consumer acceptance. The highly labour-intensive nature of sensory analysis has inevitably led to the development of instrumental methods designed to measure food properties that relate to relevant sensory characteristics. Mechanical measurements of crispness are performed on instruments originally developed for material science, providing physical parameters with fundamental significance in terms of rheological properties, i.e. the materials' response to the applied force. The texture analysers typically have a crosshead containing a load cell, which is driven vertically at a range of constant speed. Probes can be attached to the crosshead for penetrating, shearing or crushing food. The load is recorded relative to time or to deformation distance, and displayed as force-deformation plot. The compression mode is the most commonly employed deformation mode because of its similarities with the mastication process.

Models for the compression behaviour of cellular solids in general have been proposed by Gibson and Ashby [Gibson and Ashby (1988)] and include: a linear elastic region in which force rises proportionally with deformation; a plateau region where force remains at a fairly constant level due to the breakage of cell walls; a densification region in which force rises rapidly due to compaction of the structure. Brittle porous structure has a characteristic jaggedness superimposed upon these curves that results

from the abrupt fracturing of cell walls. The parameterisation of the crispness texture attributes may differ in the way the force/deformation plot are analysed, either by extracting some parameters, or by considering the signal as a whole. Moreover, data analysis of compression test may be concentrated only in the linear region of the force-deformation plot, and reflect the mechanical properties with a material science approach or, as it will be assumed in this work, information can be collected from the jagged part of the force-deformation curves.

In literature three main ways are reported: extracting parameters from each peak, calculating the power spectrum by using the Fast Fourier Transformation analysis and determining the fractal dimension of the signal, mainly using the Kolmogorov algorithm [Gibson and Ashby (1988)]. While mechanical tests are a well-experienced approach in food science to analyse the structure of foods, the relationship with the consequent acoustic emissions, in particular for crisp products, have still to be fully developed. Sound wave propagation is made of alternate regions of compression (increased density of molecules) and rarefaction (decreased density of molecules) moving through the air [Speaks (1999)]. The amplitude of displacement of molecules from the equilibrium is recorded by acoustic recording during the mechanical test and is plotted versus time (amplitude-time plot). When a microphone is used, acoustic pressure is proportional to the voltage signal from the microphone and is plotted as sound pressure level (SPL) versus time or deformation.

As previously described for the mechanical jagged force/deformation relationships, many different properties of sound waves have been measured and have been correlated with sensory properties: parameters have been extracted to characterise sound curve, fractal analysis has been used and finally Fourier analysis has been applied [Duizer (2001)]. Therefore, it is possible to surmise that the sound is emitted by the same sequence of minor and major failure events recorded in the mechanical spectrum. Results by researchers have otherwise shown that the degree of jaggedness taken from the mechanical spectrum can not be treated as a universal textural characteristic by itself. It has been found, in fact, that the com-

bination of acoustic and mechanical techniques more adequately describes food sounds than either technique alone [Duizer (2001), Roudaut *et al.* (2002)].

The sound emission technique has been used as an objective texture measurement of brittle food products since the sixties [Drake (1963), Vickers and Bourne (1976)]. In 1976 Vickers and Bourne [Vickers and Bourne (1976)] proposed a psycho acoustical theory of crispness, while in seventies most of the studies on friable bakery products, potato chips and crunch twists correlate sound emission to sensory measurements of textural crispness descriptors [Seymour and Hammann (1988)]. An improvement in instrumental assessment of crispness could arise if the acoustical data are collected alongside time in order to provide an additional source of information on texture.

Unfortunately, a huge amount of data is obtained by combining the acoustic and mechanical simultaneous tests, making difficult to extract pertinent information. Multivariate analysis has never been applied to acoustic-mechanical data, even if it has been proved to be able to handle the large amounts of data produced by modern analytical techniques and appears useful for studying this kind of data. The acoustic and mechanical properties of five different brands of sliced toasted bread were studied in this work. The final goal was to combine the acoustic-mechanical analysis with chemometrics and to show how this approach is a powerful tool with a discriminating potential, useful for the quality assessment and structural characterization of crispy products.

11.2 Analysis and samples

Acoustic-mechanical combined analysis

Five different brands of toasted sliced bread from soft wheat (named A, B, C, D and E, respectively) were purchased and analysed just after being unpackaged. The industrial technologies for the production of toasted sliced breads are highly standardised and, from the textural point of view, all

Table 11.1: Mechanical and acoustic parameters extracted from the acoustic-mechanical spectra.

Typology	**name**	**label**	**unit**
Mechanical	area	m-area	MJ m-3
Mechanical	numbers of peaks	m-peaks	-
Mechanical	linear distance	m-ldist	-
Mechanical	average stress	m-mean	MPa
Acoustic	numbers of peaks	a-peaks	-
Acoustic	linear distance	a-ldist	-
Acoustic	average Sound Pressure Level	a-mean	dB

the five brands look very similar even if they are differently appreciated by the consumers. The mechanical test was performed with the TA.XT.plus Texture Analyser and the consequent, simultaneous acoustic emission was measured with the AED acoustic envelope detector (Stable Micro Systems, Godalming, UK), which is combined with the texture analyser. The AED Acoustic Envelope Detector is an acoustic emission monitoring systems, consisting of an electro-acoustic transducer, a preamplifier, a signal conditioning system and a data acquisition system; the AED works in the frequency range of 3.125-12 kHz. The microphone was placed as close as possible to the sample in order to further improve the acquisition of the acoustic signal. For the mechanical compression test, the samples of each brand of toasted sliced breads were square shaped (side: 20 mm) and then compressed to 27% of deformation by means of a parallel plates geometry, at crosshead speeds of 10 mm/min and 600 mm/min. A load cell of 490.3325 N was used. The raw mechanical data were expressed as stress (force on unit area, MPa) vs. strain (percentage engineering deformation) while the acoustic data (that were collected simultaneously) were expressed as sound pressure level (ratio between the measured acoustic pressure level and the reference acoustic pressure level in a logarithmic scale, dB) vs. strain. Since the crosshead speed is constant, it is possible to establish a relationship between strain and time. A total number of 7363 stress/strains input were collected with the low speed (10 mm/min) protocol, while 155 stress/strains data were acquired with the high speed

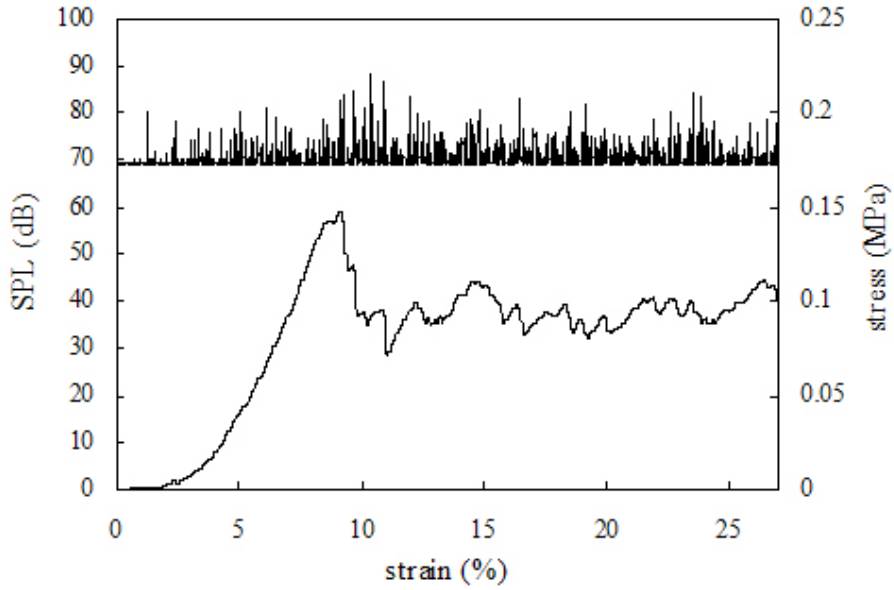

Figure 11.1: Examples of acoustic (upper) and mechanical (lower) spectra obtained at 10 mm/min compression speed, (low speed, LS).

Table 11.2: Characteristics of the analysed datasets (LS and HS): number of variables, samples, classes and the number of samples in each class (class partition) are reported.

dataset	var	samples	classes	class partition
Low Speed (LS)	7	106	5	20, 23, 21, 21, 21
High Speed (HS)	7	104	5	21, 22, 22, 21, 18

(600 mm/min) protocol. In Figure 11.1 and Figure 11.2 typical acoustic-mechanical spectra are shown both for low speed and high speed protocols, respectively. From the compression test spectra, 4 mechanical and 3 acoustic discrete parameters were extracted by means of the software Texture Exponent Exceed TEE32 (Stable Micro Systems, Godalming, UK). A full description of the parameters is given in Table 11.1.

Summarising, two datasets were analysed: the first (called LS) was obtained from the low compression speed test (10 mm/min) and included 106 samples of the 5 different brands, described by 7 discrete parameters;

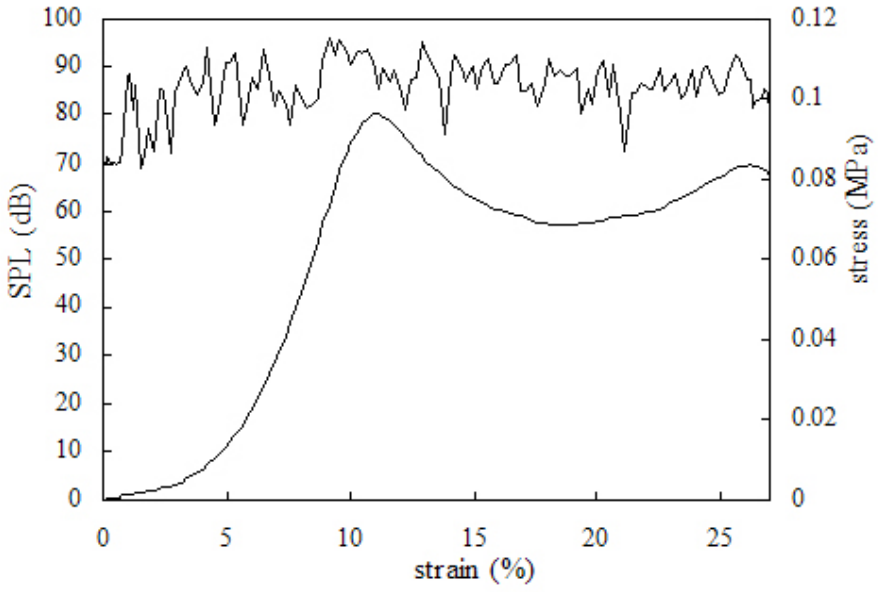

Figure 11.2: Examples of acoustic (upper) and mechanical (lower) spectra obtained at 600 mm/min compression speed, (high speed HS).

the second one (called HS) was obtained from the high compression speed test (600mm/min) and included 104 samples of the 5 different brands, described by the same discrete parameters. A summary of the characteristics of each dataset (number of samples, number of variables, number of classes, class partition) is shown in Table 11.2.

Image analysis

In order to perform a morphological characterization of toasted sliced breads, images of 24 cross sections of each brand of toasted sliced bread were acquired and elaborated. The images were acquired through a video microscope (Microflex Classic) with an enlargement of 20; the program MATROX PC-VCR supplied the digitalisation of the images (JPG format) with a resolution of 300 dpi and saved the images on. The image analysis software Image Pro PlusTM (vers. 5.0, Media Cybernetics, Silver Spring, MD, USA) was used to standardise and elaborate the digital im-

ages, previously converted into grey scale. The raw data were expressed as frequency of grey levels vs. a 0-255 pixels scale, were 0 conventionally represents the black while 255 is the white. The intermediate values represent the graduation of grey. The grey levels distributions for the 5 commercial brands were analysed by means of the Peak Fit software (Jandel Scientific): raw data distribution was smoothed according to a FFT procedure and deconvoluted in an appropriate number of Gaussian distributions.

Multivariate analysis

The data structure was analysed by means of Principal Component Analysis (PCA), applied both on the parameterised datasets (i.e. on the matrices with dimensions 1067 and 1047, respectively, for low and high compression speed) and on the original spectra obtained with the low compression speed (1067363). In the first case (parameterised datasets), autoscaling was applied, since the variables were expressed in different units (Table 11.1). In the second case (original spectra) a different data pretreatment was needed. In fact, when dealing with huge spectra, i.e. an high number of variables, the number of variables could be previously reduced by applying windows of size n, where each window is the average of the intensities at n consecutive points [Leardi and Norgaard (2004)]. In the presence of spectra with narrow peaks this approach could be quite dangerous, since some spectral features can be smoothed and lose their importance. To overcome this problem, small windows of size 5 were used, in order to reduce the data matrix and avoid an information loss, and no scaling was applied on the data. When PCA has been applied on the parameterised data, only the components with relative eigenvalues greater than 1 have been taken into account, while on the windowed original spectra the components have been retained on the basis of the explained variance.

Then, LDA and QDA have been applied in order to find predictive classification model. Both for LDA and QDA an a-priori class probability proportional to the number of objects in the class was used. Great

attention was focused on the predictive capabilities of the classification methods. All the classification models were validated using leave-one-out (LOO) and leave-more-out (LMO) procedures: in this second case, 500 repetitions and a percentage of test samples equal to 20% were always used.

Before approaching the different classification tasks, data reduction was also applied, in order to evaluate which variables had the greatest discrimination capability. Hence, all subset selection method was employed. Usually, an exhaustive search of all possible solutions is not feasible, but in this case the small number of variables (7) allowed it. In fact, with 7 variables, a total of 128 combinations (models) could be obtained. In order to overcome the presence of overfitting during the variable selection, the following procedure was used. The samples were divided into four cross validation groups; once at a time, each validation group was removed from the training set, the classification model with the smallest ER on the training set was chosen; the optimised model was subsequently tested on the samples of the removed validation group: therefore, for each step, the optimisation was performed without the samples to be predicted. At the end of the procedure, the percentage of wrong assignments in the cross validation groups (ER_{CV}) was calculated. If ER_{CV} was satisfactory, the full model was built by taking into account only the subsets of variables selected with each cross validation groups, with a relative frequency higher than 1 [Baumann and Stiefl (2004)]. Then, the full model predictive capability was also tested by means of LOO (ER_{LOO}) and randomly repeated LMO (ER_{LMO}) procedures. Multivariate analysis was performed by means of MATLAB 6.5 (Mathworks).

11.3 Results

A number of factors should be considered before setting up a protocol for measuring the textural properties of a particular food by means of the texture analyser: (1) the nature of the product (liquid, solid, brittle, plastic, homogeneous, heterogeneous material); (2) the purpose of the test, in the

sense that in some cases a single-point measurement is adequate, while, in other cases, a multipoint measurements is needed; (3) the test principles, i.e. the deformation mode (compression, tensile, flexural, bending modes); (4) the operative test conditions (sample size, deformation speed, deforming forces) in order to give the best resolution between different samples. For routine quality control purposes, where a rapid test is essential, sophistication is sacrificed for the sake of rapidity. On the other hand, when a more fundamental approach is followed in research laboratories or in a new product development, a higher rigor in test making, in data acquisition, and treatment are needed.

The compression rate is fundamental in testing brittle cellular material like toasted sliced breads. Two widely different compression rates were considered: the lower, (10 mm/min) mainly suitable for the characterization of toasted sliced breads structure, the higher (600 mm/min) preferably useful for classification purposes. Therefore, the two data set LS and HS were initially considered together, i.e. a matrix with 210 rows (samples) and 7 columns (variables) was built. PCA was applied on this matrix to study the differences of the acoustic-mechanical outcomes resulting from the two deformation rate conditions. The variables were autoscaled and the first two components, which accumulated together 85% of the data variance, were considered.

Examining the score plot (Figure 11.3) it is clear how the samples are separated on the first component: all the samples tested at high compression rate have scores greater than 0, while the low rate compressed ones have score lower than 0. Along the second component a trend of class separations is present, but there are several overlaps between the classes; for examples the classes C and D both for HS and LS have respectively the highest and the lowest scores, but especially for LS there is a clear superposition among the classes.

In Figure 11.4 the correspondent loading plot is shown. The seven variables are clearly grouped: on the first component only a-mean has a positive loading, while m-peaks, a-peaks, a-ldist, m-ldist have the highest negative values. The two remaining variables (m-area and m-mean) have

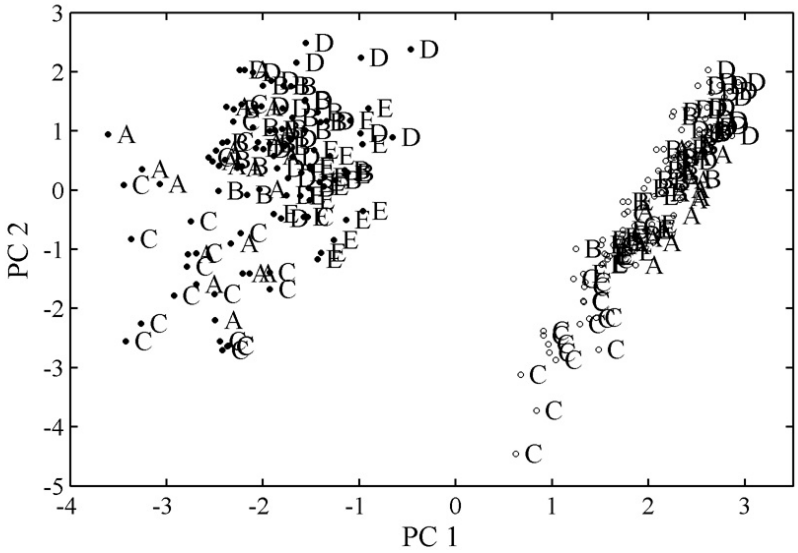

Figure 11.3: PCA on the two data set LS and HS: score plot of components 1 and 2 (Explained Data Variance=85%). Samples compressed at high speed (HS) are marked with white circles, while samples compressed at low speed (LS) are marked with dark circles. Each sample is labelled with its class code (A, B, C, D and E).

a small influence on the first component, but they are the only meaningful variables on the second component. From the interpretation of this graph, significant observations for the acoustic-mechanical test arise. The first component accounts for the variables taken from the jagged spectrum, which is the peculiar response of the cellular structure. The shape and dimension of cells, their density and the mechanical properties of their walls determine, at macroscopic level, the mechanical and acoustic response of the structure during crushing. The breaking energy applied to the cell walls increases with the speed of deformation. The first component then separates samples according to the compression rate, while the total mechanical deformation energy (m-area) and the average mechanical stress value (m-mean) express bulk properties of the matrix: it is therefore justified, from the mechanical point of view, the role of these variables in characterising the classes of objects. It is interesting to notice that the

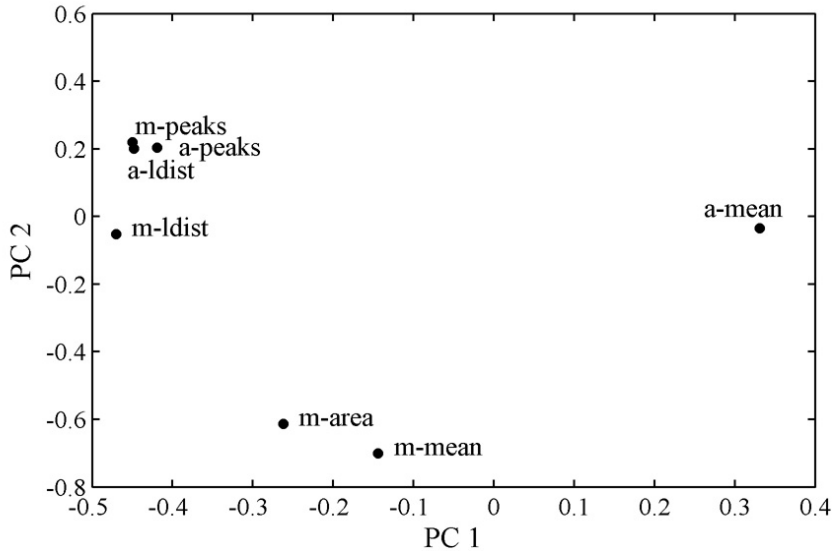

Figure 11.4: separated on the first component: all the samples tested at high PCA on the two data set LS and HS: loading plot of components 1 and 2 (Explained Data Variance=85%).

parameters a-peaks, m-peaks, a-ldist and m-ldist have high loading values on the first component and strongly characterise the LS data; from a rheological point of view, high values of these variables mean high resolution in defining the irregularity of the force-deformation relationship. Since the pattern carries information that is relevant to structure, it can be interpreted in structural terms: the degree of irregularity of the pattern is itself a manifestation of the deformation mechanism and may provide information not directly related to the bulk crispness attribute, but to the structural basis of this attribute. The first result from the explorative PCA is that digitalized force-displacement relationships are not fractal objects in mathematical sense: they have a finite resolution and different extracted variables characterise, from the chemometric point of view, LS and HS data sets. These observations confirm Peleg and co-workers choice for an apparent fractal dimension as an effective measure of jaggedness [Borges and Peleg (1996)]. After that HS and LS were analysed separately and in

Table 11.3: Comparison of error rates (%) obtained by LDA and QDA with the LS dataset by considering all the classes. M refers to the subset of mechanical parameters, A to the subset of acoustic parameters, A+M to both mechanical and acoustic parameters, All Subset to the selection achieved by means of the All Subset method; ER_{LOO} refers to leave-one-out validation, ER_{LMO} to leave-more-out validation; no.var refers to the number of used variables (for the All Subset method, the first value is related to the number of variables selected by LDA, while the second one by QDA); mean is the average of the error rates achieved with each subset of variables.

		LDA		**QDA**		**mean**
	no.var	ER_{LOO}	ER_{LMO}	ER_{LOO}	ER_{LMO}	
M	4	51.9	53.1	55.7	54.9	53.9
A	3	52.8	53.8	52.8	54.9	53.6
A + M	7	37.7	39.2	35.9	40.9	38.4
All Subset	4, 4	36.8	37.8	36.7	39.9	37.8

the following paragraphs results are shown.

11.3.1 Low compression speed data

One of the primary goals of instrument evaluation of food texture attributes is to provide suitable means for food authentication and quality assessment. From this point of view, classification models are convenient tools to be proposed. Classification models were therefore built in order to evaluate the separations degree between the five classes of commercial brands of toasted sliced breads according to their acoustic and mechanical properties. LDA and QDA were applied on three different subsets of variables: (1) the four mechanical parameters, (2) the three acoustic parameters, (3) both mechanical and acoustic ones. In order to improve the models and to better understand which variables have more discriminating power, also variable selection by means of all subset model technique was applied. All the models were evaluated on the basis of their predictive performances, i.e. the results achieved with LOO and LMO procedures (Table 11.3).

The classification with either the mechanical subset or the acoustic one

gave error rates higher than 50%, while in combining these two subsets better models are achieved. The models found with all subset selection confirmed this behaviour: both LDA (m-area, m-peaks, a-peaks, a-ldist) and QDA (m-area, m-mean, a-peaks, a-ldist) models involved mechanical within acoustic variables. It is interesting to observe that considering both mechanical and acoustic parameters increases the classification. This means that, by adding the acoustic dimension to the classical mechanical measurements, a new information content is obtained, due to the synergic effect of the two kinds of variables.

However, the obtained error rates show that a potential class separation is present but it is not achieved in an optimal way. In particular, results mislay the macroscopic differences in the organisation of the gas cellular structure of A, B, C, D and E brands of toasted slice breads. A regular design of the cells and homogeneity in cell distribution on the single slice and between samples of the same brands are in fact considered a perceivable quality attribute. It is to underline that the five commercial brands are hardly separable relatively to their morphology, since the producers progressively move together to a common and standardised technology of production. These quality properties were investigated by means of image analysis (Figure 11.5) and PCA (Figure 11.6).

In Figure 11.5, the grey levels distribution of brand A is shown. The same behaviour was also observed for the brands B, C, and D. The deconvolution in four Gaussian distributions singles out four gas-cells classes. It can be concluded that the toasted sliced breads of the brand A (and similarly brands B, C and D) are characterised by inhomogeneous gas cell distributions. On the other hand one Gaussian describes the grey level distribution of brand E (Figure 11.5) and this allows concluding that this rigid foam is homogeneous and regular in the structure.

Further data modelling was therefore needed to support the visual remarks. The original acoustic spectra were divided in windows (as previously described in Material and methods) and PCA applied without scaling. The first fifty components, which accumulated together 72% of the data variance, were considered: the centroid of each class was calcu-

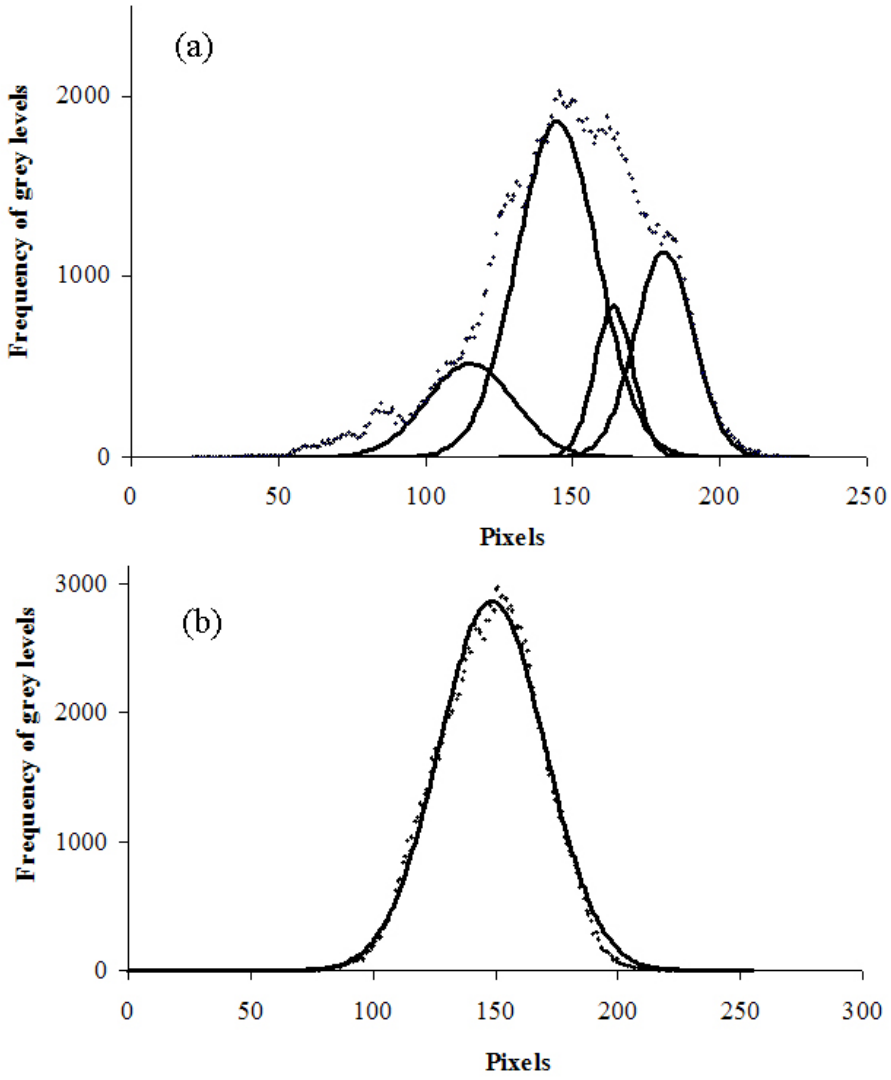

Figure 11.5: Image analysis: grey levels distribution (dotted line) of class A (a) and E (b) and their deconvolution in Gaussian distributions (solid line).

lated in the score space and the distance of each sample from its class centroid carried out.

In Figure 11.6, the mean distance of the samples of each class is shown: it is clear how class E has a considerably lower distance with respect to the

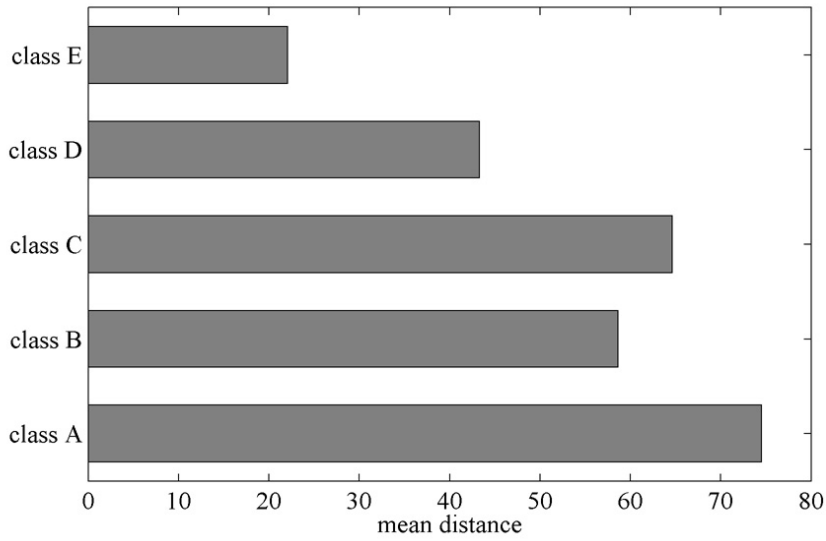

Figure 11.6: PCA on the original acoustic spectra: mean distance of the samples of each class from the class centroid in the score space.

other classes, i.e. the class can be considered more homogeneous among samples. Finally, classification was applied in order to verify if mechanical and acoustic parameters confirm these structural quality properties. Hence all the samples were splitted in two classes: class E (21 samples) and class not-E (85 samples), which includes all the analysed classes other than E. The classification models were built following the previously described steps, i.e. LDA, QDA applied on different subsets of variables, validated by means of LOO and LMO procedures (Table 11.4).

The classification achieved by using only the mechanical parameters gives the worst results (error rates always higher than 20%) while the acoustical variables carried out acceptable models (ER_{LMO} equal to 7% for LDA). The best model found by means of all subsets selection applied on QDA (ER_{CV} equal to 8.3% and ER_{LMO} equal to 8.0%) involved two mechanical parameters (m-peaks, m-mean) and one acoustic parameter (a-mean), i.e. once again the combination of the two kinds of data seems appropriate for samples description: two of them (average mechanical

Table 11.4: Comparison of error rates (%) obtained by LDA and QDA with the LS dataset, by considering class E versus all the other classes. See Table 11.3 for symbols and acronyms.

		LDA		**QDA**		**mean**
	no.var	ER_{LOO}	ER_{LMO}	ER_{LOO}	ER_{LMO}	
M	4	21.7	21	21.7	21.9	21.9
A	3	7.6	7	10.4	10	10
A + M	7	10.4	10.1	11.3	11.8	11.8
All Subset	3, 3	8.5	8.8	7.5	8	8.2

stress and average sound pressure level) account for force and energy, i.e. textural features, the latter (mechanical numbers of peaks) is capable of identifying fracture events, with specific references to structure details.

11.3.2 High compression speed data

Typically, compression tests on crispy products are performed at a deformation rate ranging between 10 and 20 mm/min, i.e. low compression speed. In this work the samples were compressed also at high deformation rate, i.e. 600 mm/min, in order to simulate the range of human mastication rates. Since this methodological approach is rarely applied, PCA was initially performed on the obtained HS data in order to study the data structure. The variables were autoscaled and the first two components, which explain together 70% of the data variance, were considered. In the score plot (Figure 11.7) it is possible to observe how some of the five classes are separated among them, while other classes overlap. For example, class C overlaps with class E, but it is unconnected with all the other classes; class D overlaps with class A and B, but it is clearly separated from classes C and E; class B appears as the more confused class in the score plot. In the correspondent loading plot (Figure 11.8), the seven variables appear clearly separated: all the acoustic parameters (a-ldist, a-mean, a-peaks) have negative loadings on the first component and appear grouped, such as three mechanical parameters (m-ldist, m-mean, m-area), which, on the contrary, have positive loadings. Finally m-peaks has no influence on the

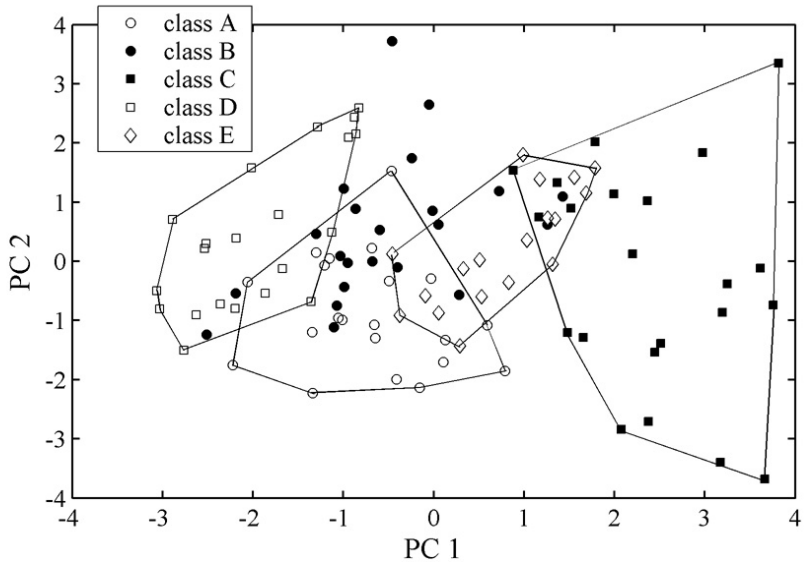

Figure 11.7: PCA on the high compression speed data: score plot of components 1 and 2 (Explained Data Variance=70%). The five classes are shown with different colours and shapes. The class spaces (except for class B) are marked with solid lines.

first component and appear different from all the other variables.

Summarising, all the variables (except probably m-peaks) appear relevant in describing the data structure and significant for the class separation. Therefore, as it has been presented for the LS dataset, classification models were built by means of LDA and QDA in order to separate the five classes according to their acoustic and mechanical properties. All the models were evaluated on the basis of their predictive capabilities: the achieved results are shown in (Table 11.5). The acoustic parameters did not give acceptable classification models (error rates always higher than 65%) while the mechanical ones provided better models but with error rates over 30% as well as all the seven variables combined together. The best achieved model was obtained by means of all subset selection applied to LDA (ER_{CV} equal to 30.8%, ER_{LOO} and ER_{LMO} equal to 27.9%), which selected two mechanical variables (m-area and m-mean) and one acoustic parameter (a-ldist): from the force-deformation traces, in this case, in-

Table 11.5: Comparison of Error Rates obtained by LDA and QDA with the HS dataset, by considering all the classes. See Table 11.3 for symbols and acronyms.

		LDA		**QDA**		**mean**
	no.var	ER_{LOO}	ER_{LMO}	ER_{LOO}	ER_{LMO}	
M	4	31.7	33.6	34.6	34.6	33.6
A	3	66.4	67.8	73.1	75	70.6
A + M	7	30.7	31.1	41.4	42.4	36.4
All Subset	3, 3	27.9	27.9	28.8	29.9	28.6

formation are extracted essentially about the specimen stiffness, which is associated with the absolute force. These properties should therefore be treated as textural descriptors. Once again the combination of acoustic and mechanical parameters can improve the data analysis. Moreover this model is better than all the classification models obtained with the LS dataset (Table 11.3) since the mean error rate decreased more than 9% (from 37.8% to 28.6%): the high deformation rate protocol appears more appropriated for brand characterization purposes with respect to the low deformation rate protocol.

Finally, the classification of class E samples was considered. The classification models were built following the previously described procedures and the achieved results are shown in Table 11.6. With respect to the models built with the LS dataset (Table 11.4), the obtained error rates were significantly higher: the best ER_{LMO} found on the HS data is equal to 12.3% (QDA with three mechanical variables selected by means of all subset method). Therefore, low speed protocol appears more, indicated in evaluating the structure properties related to the quality of the sliced toasted breads.

These results have outlined an upgrading of the fundamental instrumental texture analysis by means of chemometrics. Until now chemometrics has been poorly applied in this field, but a lot of applications can be done and it can represent a new opportunity of data analysis for rheological issues. Chemometric procedures have been here presented for the treat-

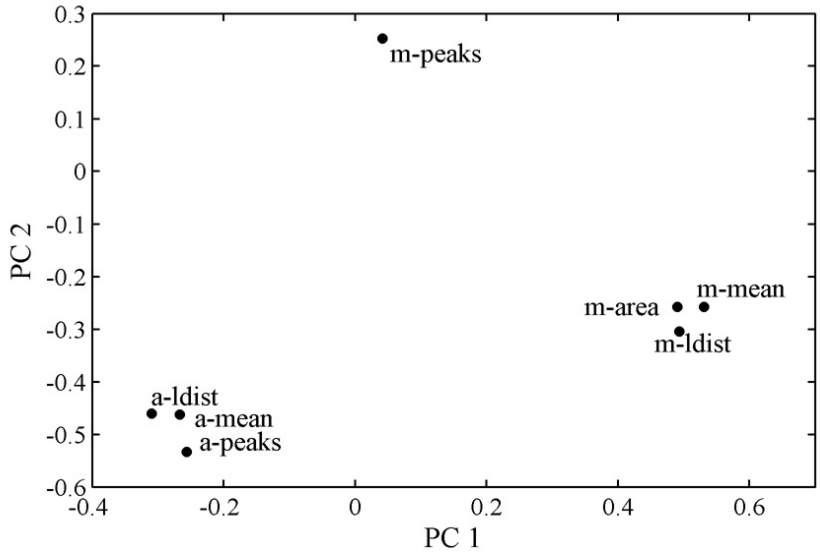

Figure 11.8: PCA on the high compression speed data: loading plot of components 1 and 2 (Explained Data Variance=70%).

Table 11.6: Comparison of Error Rates obtained by LDA and QDA with the HS dataset, by considering class E versus all the other classes. See Table 11.3 for symbols and acronyms.

		LDA		**QDA**		**mean**
	no.var	ER_{LOO}	ER_{LMO}	ER_{LOO}	ER_{LMO}	
M	4	21.2	22.1	12.5	13.5	17.3
A	3	17.3	19.7	17.3	20.1	18.6
A + M	7	22.1	23.8	14.4	16.7	19.3
All Subset	3, 3	18.7	19.8	10.6	12.3	15.3

ment of data collected during acoustic-mechanical test performed with a universal testing machine, in order to evaluate typical texture properties of crisp bakery products.

The final goal was to show how this approach is a powerful tool with a discriminating potential, useful for the quality assessment and structural characterization of food products: basic chemometric tools, such as Principal Component Analysis and Discriminant Analysis, proved to be able to

extract relevant information and offer an easy approach for the interpretation of acoustic-mechanical parameters. In fact PCA allowed to derive useful observations on the texture of the analysed toasted sliced bread samples, while acceptable classification models, both from chemometrics and rheological points of view, were built by means of Discriminant Analysis. Jagged force-deformation relationship, despite their irreproducible shape, can yield fairly consistent texture parameters, by following appropriate chemometric modelling. Further investigation on the nature of the data will be developed by means of chemometrics, by considering the whole acoustic and mechanical spectra, in order to join or overcome the parameters extraction.

chapter 12

Evaluation of sensory time-intensity signals

12.1 Introduction

The Time Intensity (TI) is a dynamic sensory method, in which oral burn can be evaluated as it changes over time [Piggott (2000)]. The time-intensity curves are based on the responses of the assessors and are made by an increasing part, a plateau and a final decreasing part. In order to analyse TI data with the classical multivariate techniques, a weighted average of the individual curves is usually calculated or a parameter-extractions is applied. It is clear that in this way, a part of information could be lost. TI data can be also seen as a 3-way matrix, where the first dimension (called also mode) is represented by samples, the second mode by times and the third mode by assessors. Therefore, the multi-way approach allows information about each mode to be obtained at the same time: it has the ability to show the global, and at the same time, particular information in an easily interpretable way. In fact, in this way all the interrelations in-between modes can also be analysed, while with a 2-way

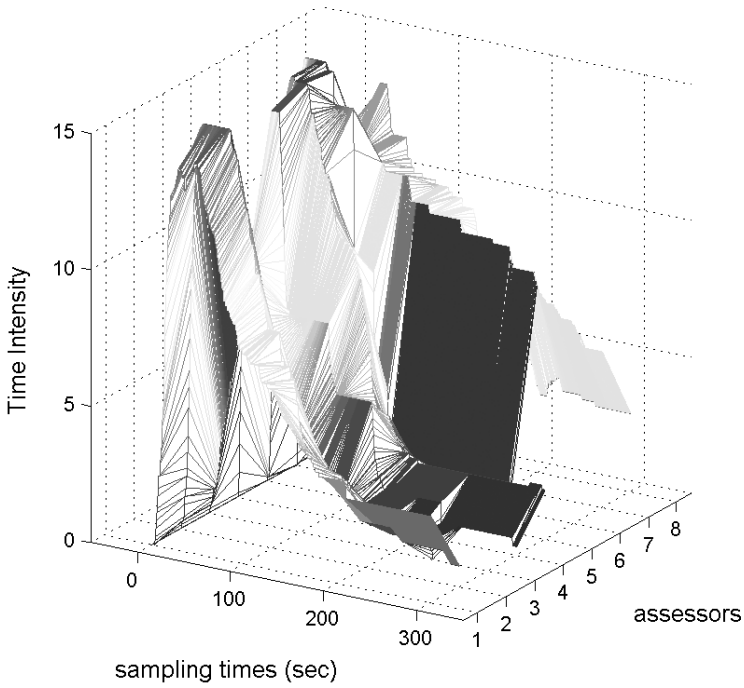

Figure 12.1: Graphical representation of a TI sample, assessed by several assessors during time.

approach a part of this information is necessarily lost.

The most common technique for 3-way data structure analysis is the PARAFAC model [Smilde *et al.* (2004)], which can be compared to a bilinear PCA analysis. An evolution of the PARAFAC model is the so called PARAFAC2 model [Smilde *et al.* (2004)]. The PARAFAC2 model has the advantage of providing a set of time loadings for each assessor: this permits to overcame the presence of shifts in the time mode, i.e. different behaviour of the assessors during the sampling time. The correlation between the oral burn from chilli, the meat taste, and the texture in pork patties have been studied by means of 3-way analysis. Pork patties have been prepared with two chilli levels and two texture levels, in order to compare the assessor responses with regard to perceived oral burn and

Table 12.1: Experimental design and considered samples (+: high level, -: low level, 0: no chilli)

sample code	chilli	texture
T1	0	-
T2	0	+
T1C1	-	-
T1C2	+	-
T2C1	-	+
T2C2	+	+

effects on meat taste perception.

12.2 Experimental design and TI datasets

The panel consisted of eight assessors and the panelists were well trained in the assessment of different meat products, but they were not familiar with the Time-Intensity methodology. The training on Time-Intensity methodology was conducted following the guidelines proposed by Peyvieux and Dijksterhuis [Peyvieux and Dijksterhuis (2001)].

Meat taste and chilli burn have been evaluated on pork patties with two texture levels and two chilli levels (Table 12.1). The meat taste has been evaluated twice in all six types of patties, and the chilli burn twice in the four types of chilli patties. Chilli burn has not been evaluated on samples without chilli (samples labelled as T1 and T2 in Table 12.1). The panelists have evaluated the samples in duplicates, during 8 sessions. The assessment begun when the assessor put the pork patty in the mouth and was stopped after 300 seconds. The assessors have evaluated one attribute at each serving on a 15 cm unstructured line scale, by using a marker (mouse). The position of the marker has been recorded every two seconds, i.e. 150 time intensity values have been obtained for each sample assessed by each panelist.

Two datasets were created, one for each attribute. For the chilli burn the 3-way cube has been arranged with the modes: 8 samples times 150

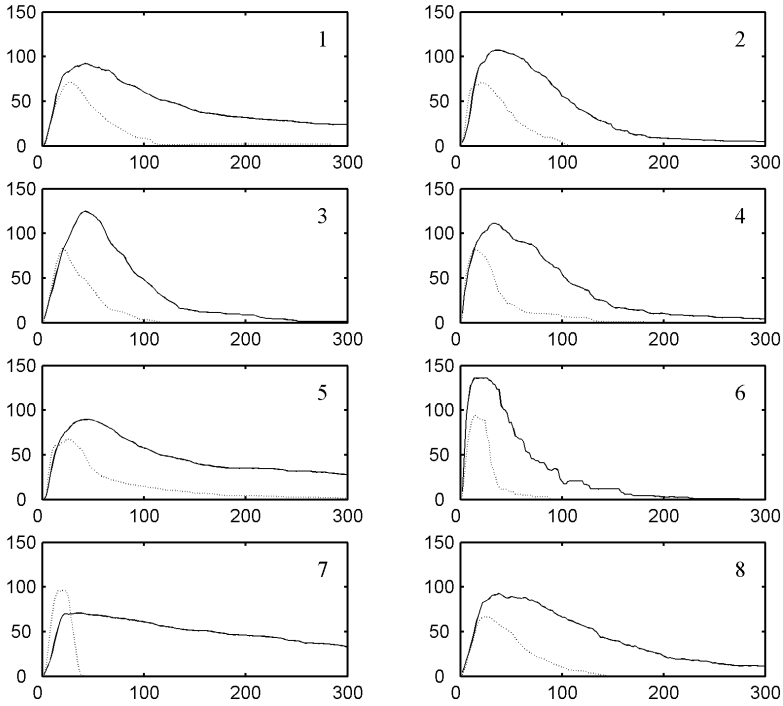

Figure 12.2: Time loadings for the two PARAFAC2 (PC 1) models for each assessor. Time in sec. (X-axis) loading (Y-axis), meat taste (dotted line) and chilli burn (black line).

recordings times 8 assessors. For the meat taste with the modes: 12 samples times 150 recordings times 8 assessors. The datasets have been analysed by means of PARAFAC2 model: in fact, the PARAFAC2 models has recently been shown to be successfully applicable on TI data [Ovejero-Lopez *et al.* (2005)]. Since the time profiles can differ from assessor to assessor, for each assessors scoring, a corresponding set of time profile loadings is obtained with the PARAFAC2 model, i.e. the PARAFAC2 model allows that every assessor can have its distinct set of time loadings. Hence, there is not only one loading matrix for the time profiles as there would be in a PARAFAC model (see chapter 3). Multi-way analysis has

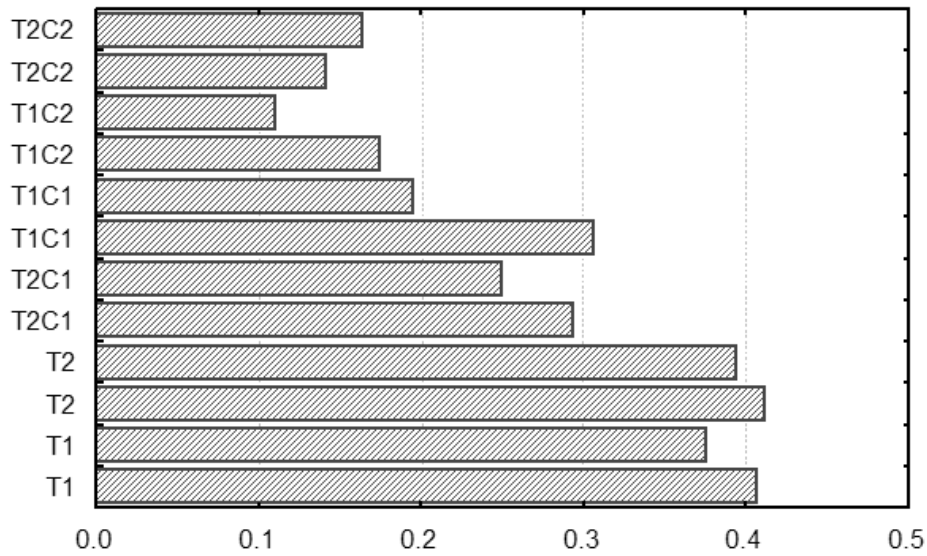

Figure 12.3: Sample loadings for the PARAFAC2 model on taste of meat.

been performed in MatLab 6.5 (Math-Works) with the N-way Toolbox [Andersson and Bro (2000)].

12.3 Results

For each attribute (meat taste and chilli burn) a PARAFAC2 models has been calculated. One component has been chosen for both models, the model on chilli burn explains 94% of the variation, while the model on meat taste explains 75% of the variation. Time loadings for the two models are shown for each assessor in Figure 12.2.

The benefit of using a multi-way approach is now clear: from Figure 12.2, some important characteristics can be easily seen. In fact, the evaluation of chilli burn has a longer duration, though the profiles of the two curves within each assessor show the same behaviour. Assessor 1, 5, and 7 have a slower decrease when evaluating chilli. Assessors 3 and 6 have the highest loadings with regard to evaluation of chilli burn, which means,

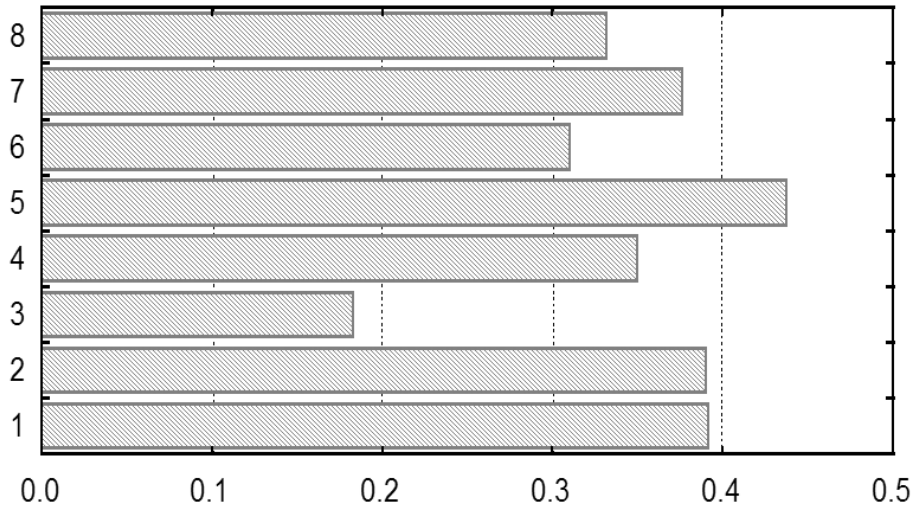

Figure 12.4: Assessor loadings for the PARAFAC2 model on taste of meat.

that these two assessors are better at perceiving the difference between the chilli levels compared to the other assessors.

The sample scores evaluated for chilli burn show a small difference between the two chilli levels, and no correlation to texture. Loadings of samples evaluated for meat taste (Figure 12.3) show a greater degree of difference, with highest loadings for samples without chilli, meaning these patties had more meat taste than the patties with added chilli. It can further be seen, that patties with high chilli level have the lowest loading values. Useful information can be obtained also with the assessor loadings. It can be seen (Figure 12.4), that the assessors agree upon their evaluations (the loadings have the same sign), but assessor 3 has lower loadings than the other panelists, and assessor 5 has the highest loading. Hence, a comparison of these 2 assessors can be evaluated on the raw data (Figure 12.5). Finally, the ability of the assessors to replicate the assessments of meat flavour and chili burn was studied from a contour plot (Figure 12.6). In general higher repeatability was observed for meat flavour than for chili burn and only few replicates were assessed with low

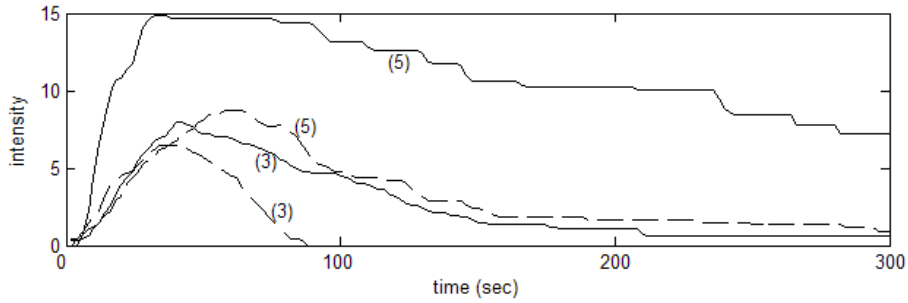

Figure 12.5: Comparison of the evaluation of chilli burn by assessors 3 and 5. Solid line: sample T1C2, dashed line: sample T1C1.

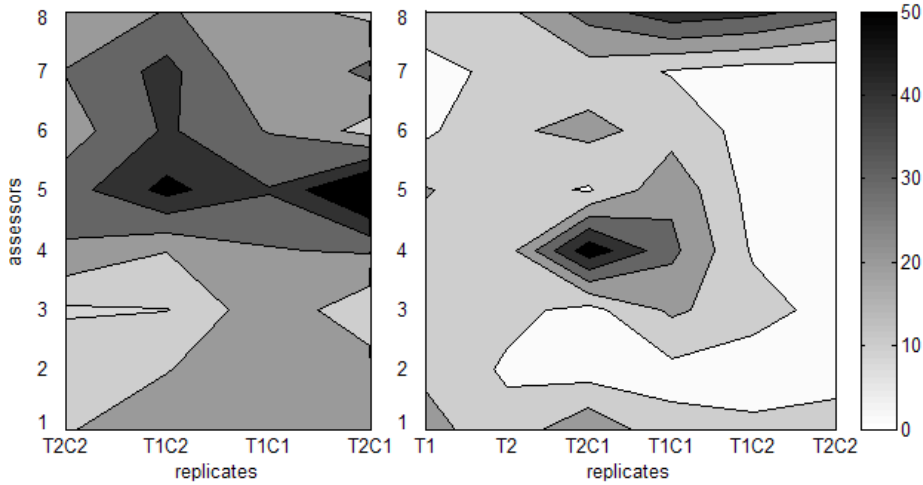

Figure 12.6: Contour plot of replicates on burning sensation (left) and taste of meat (right); the darker the colour, the longer the distance between the two replicates.

repeatability (dark grey to black areas). Especially the meat flavour in pork patties without chili showed high repeatability, indicating that the evaluation task becomes simpler when chili burn does not interfere with the meat flavour evaluation. The lower repeatability observed for chili burn evaluations (especially for assessor 4-8), could indicate that the longer chili burn evaluations make the task more demanding than when doing meat flavour evaluation.

In conclusion, multi-way analysis has been shown as a good tool for Time Intensity data analysis. The data have been arranged in a 3-way cube and analysed by means of the PARAFAC2 model. The model provided loadings matrices for samples, assessors and time intensity profiles. In this way useful information has been easily extracted and evaluated. Therefore, the PARAFAC2 model allowed information about each mode to be obtained at the same time: it has the ability to show the global, and at the same time, particular information in an easily interpretable way.

chapter 13

Compression and variable selection on wine data

13.1 Introduction

Since the coupling of classification and variable selection methods can represent an optimal solution for preserving the relevant information for the sample classification, the classification performances on the geographical origin of 62 wine samples for several classification methods have been compared, coupled with different variable selection approaches.

The wine samples have been analysed by means of dynamic headspace gas chromatography mass spectrometry (HS-GC-MS) and the entire chromatographic profile has been considered to build the models. Normally only peak areas are used to characterise chromatographic profiles, but as a result of such a feature-extraction before modelling, relevant information from the raw data could be lost. More information can potentially be extracted by preserving the entire profile as a digital fingerprint of the analysed samples and perform the feature selection by means of variable selection. Since variable selection techniques can provide over fit when

Table 13.1: Origin and classes of the analysed samples.

origin	class	no. samples
Argentina	South America	7
Chile	South America	16
Australia	Australia	20
South Africa	South Africa	19

dealing with a huge number of variables (like in this case), also a data reduction has been applied, by means of the methodology explained in chapter 6.

13.2 Materials and methods

Wine samples

Sixty-two samples of red wine, produced from the same grape (Cabernet Sauvignon) and belonging to different geographical areas have been collected from local supermarkets and analysed by means of HS-GC-MS. Details on the sample origins and classes are given in Table 13.1.

Instrumental analysis

In both the sample preparation and GC run the samples were randomised to minimise the introduction of systematic effects in data.

10 mL of each wine sample, without sample preparation, were added directly into the 100 mL purge flask and 2 mL 4-methyl-1-penthanol (50 mg/L) was added as internal standard. The samples were equilibrated to 30°C in a circulating water bath and then purged with nitrogen (75 mL/min) for 20 minutes. The volatile compounds were collected on a Tenax-TA trap. The trap contained 250 mg of Tenax-TA with mesh size 60/80.

The trapped volatiles were desorbed using an automatic thermal desorption unit (ATD 400, Perkin Elmer, Norwalk, USA). Primary desorption was carried out by heating the trap to 250°C with a flow (60 mL/min)

of carrier gas (He) for 15.0 min. The stripped volatiles were trapped in a Tenax TA cold trap (5°C), which was subsequently heated at 300°C (secondary desorption). This allowed for rapid transfer of volatiles to a gas chromatograph-mass spectrometer (see below) through a heated (225°C) transfer line.

Separation of aroma compounds was carried out on gas chromatography system (HP 6890 GC with an autosampler for liquid samples) with a 30 m long DB-Wax capillary column with an internal diameter of 0.25 mm, and with a film (thickness). The column flow rate was 1.0 mL/min using helium as a carrier gas. The column temperature program was: 10 min at 45°C, from 45°C to 240°C at 6°C/min, and finally 10 min at 240°C. A split ratio of 1:50 was used for all the experiments.

Data was initially explored and analysed in Agilent ChemStation software. Subsequently, the data was exported to MATLAB where advanced and comprehensive data analyses (baseline correction, peak alignment, modelling, classification etc.) were conducted.

Summarising, a data matrix with 62 rows (wine samples) and 2700 columns (points of chromatographic profile) has been considered and analysed.

Classification and variable selection methods

Three different approaches have been evaluated in order to build the classification models: Partial Least Square Discriminant Analysis (PLS-DA), Extended Canonical Variates Analysis (ECVA), and Linear Discriminant Analysis (LDA). In the present work, PLS-DA models have been investigated both considering all the available classes at the same time and considering each class at a time, i.e. building a classification model for each class versus all the others. The performance evaluation of the presented classification models has been based on Non-Error Rate (NER) values, i.e. on the percentage of correctly assigned samples, evaluated both on cross validation groups and external test samples.

Forward Selection and Genetic Algorithms have been both used, cou-

pled with LDA, in order to select the retained scores, which maximise the classification performances of the models. Concerning Genetic Algorithms, the approach used in this work is an evolution of the algorithm described by Leardi and Lupianez Gonzalez [Leardi and Lupianez (1998)], where the basic difference is that the Genetic Algorithm is coupled directly with LDA so that the selected variables in each step are evaluated by means of Linear Discriminant Analysis. The selection of variables is performed by repeating a GA t times (runs) and then including the variables on the basis of the frequencies of selection of each variable in the best model of each run and on the basis of NER values as a function of the number of selected variables.

In the following paragraphs, the term "variable" will refer to a single column of the original dataset, i.e. to a point of the original chromatographic profile; the term "window" will refer to a group of consecutive variables, while the term "score" will refer to each score extracted with the proposed approach, i.e. each score represents a single column of the reduced data matrix.

Software

Calculations have been performed in MATLAB 6.5 (Mathworks). The PLS toolbox (Eigenvector Research, Inc., Manson, Washington) has been used for PLS-DA; ECVA have been performed with the MATLAB modules available at www.models.kvl.dk; GAs have been applied by means of MATLAB modules given by R. Leardi (University of Genova), while the score extraction has been performed with routines built by the authors.

13.3 Results

Data pretreatment and test set selection

Prior to the extraction of scores the chromatographic profiles were aligned using the automated aligned approach by Skov *et al.* [Skov *et al.* (2006)]. Having aligned the data, the score extraction using PCA will be more

efficient, as the aligned data now can be explained by a low-rank bilinear model. This allows the PCA model to focus on the variation between classes and not on the peak shifting behaviour due to the chromatographic process

Afterwards, PCA has been applied on the original data matrix and the first two principal components, which accumulate together the 65% of Explained Variance , considered as significant. The retained scores have been subsequently used to split the 62 samples into training (46 samples) and test (16 samples) sets, on the basis of the Kennard-Stone algorithm [Rajer-Kanduc *et al.* (2003), Wu *et al.* (1996), Kennard and Stone (1969)]. This was done in a way such that all the three classes were proportionally represented in the test set. This procedure assures that representative samples, internal to the data domain, are selected as test objects. Any subsequent scaling on the test set (mean centering before applying PCA, PLS-DA and ECVA) has been performed by using the parameters obtained by the training set.

Concerning the training samples, these have been divided into five groups, where the classes were equally represented, in order to perform an internal cross validation procedure on all the classification models.

PLS-DA and ECVA without variable selection

Initially, PLS-DA and ECVA have been applied on the entire chromatographic profile, i.e. a data matrix with 62 rows and 2700 variables. The number of optimal components for both the procedures has been selected on the basis of the cross-validated results and then the models have been used to predict the classes of the test samples.

The obtained NER are shown in Table 13.2. PLS-DA gives the best model using six factors (NER equal to 60% and 68% for the cross-validation groups and the test samples, respectively), while when the three classes are separately considered the performances decrease. ECVA gives the best classification results and performs better then PLS-DA. In fact, considering that the three classes overlap and are not clearly and completely

Table 13.2: Non Error Rate (NER, % of correctly assigned samples) achieved with different classification methods. FIT refers to the training samples, CV to the cross validated samples, TEST to the external test samples. The number of used factors is also reported: for the PLS-DA model built separately on each class, the number of factors for the three models is reported.

method	**Factors**	**FIT**	**CV**	**TEST**
PLS-DA	3	67	50	56
PLS-DA	6	84	60	68
PLS-DA with each class	2, 2, 3	67	41	50
ECVA	14	97	87	81

separated, error rates equal to 13% on the cross validation groups and 19% on the test set can be considered as close to optimal in this situation.

Score extraction

The extraction of scores has been performed by using a window size equal to 10 points, i.e. 270 different windows have been analysed, since the total number of chromatographic points was 2700. The window size has been selected by scrutinising the chromatographic peaks: the majority of the represented peaks are included in a range of 10 points, i.e. each window could contain and represent a single peak. Regardless, if peaks were larger than the selected window size or if the window did not include the entire chromatographic peak, the described phenomena should be preserved, since in a chromatographic peak the ascending and the descending sides conserve the same information, i.e. the same chemical fingerprint. Alternatively, the window size may be selected manually by the user, in order to correctly include all the known phenomena in the specified windows. After windowing the profile, all the obtained windows have been mean centered and PCA has been performed; then, the scores have been retained when Explained Variance was higher than 10% and Squared Sum of Residuals higher then 0.0001. These selected thresholds are obviously subjective and depend on the data and the problem.

However, these thresholds have been chosen on the basis of the phe-

nomena linked to each window, i.e. maximising the number of retained components related to windows where the chromatographic peaks were present and minimising the number of components related to noisy windows or windows without any peaks. Finally, 63 scores (representing 56 windows) have been retained by using the training samples, while the test samples have been subsequently projected in the score spaces.

Variable selection and ECVA on the retained scores

Genetic algorithms, coupled with LDA, have been applied on the retained score matrix. Initially, a screening test has been performed in order to optimise the number of evaluations (generations) to be used for each GA run. This is done as GAs start by modelling the main information and then gradually overfit more and more. To verify when overfitting becomes pronounced, a series of 40 runs has been performed, i.e. GAs have been repeated 40 times considering a maximum number of evaluations equal to 200: in the first 20 runs the class vector is the original one, while in the second 20 runs the class vector has been randomised. Then, a vector of the differences between the mean performances obtained with the randomised vectors and the original class vectors has been computed. When working with a good dataset, these differences increase and reach a plateau, that corresponds to the optimal number of evaluation to be used [Leardi and Lupianez (1998)].

In Figure 13.1 the differences as a function of the number of evaluations are shown: 120 has been chosen as the optimal number of evaluations, since after this value the differences do not increase in a significant way. Then, GAs have been performed on the retained score matrix of the training samples, by using the settings summarised in Table 13.3.

In Figure 13.2 the histogram of frequency of selection of each retained score on the 500 runs is showed, while in Figure 13.3 the cross validated NER as a function of included scores is represented. On the basis of these results, 6 scores have been included in the final LDA model, where the first considered score is the one with the highest frequency, the second

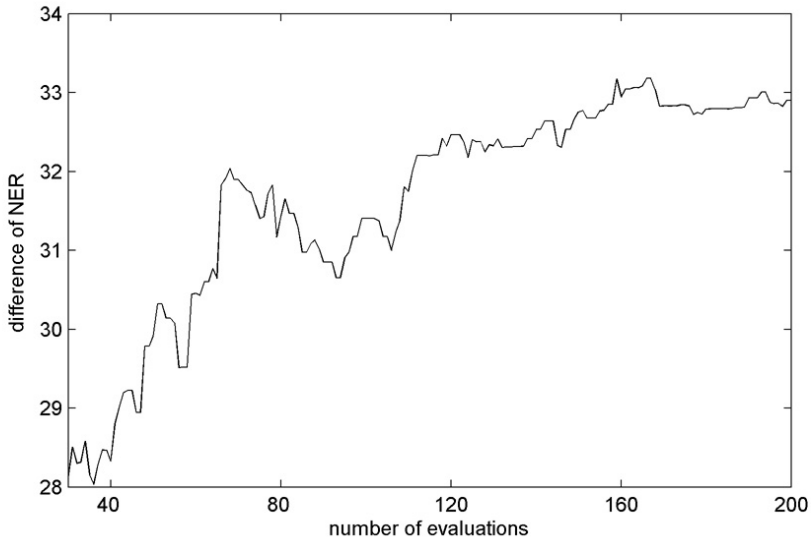

Figure 13.1: Genetic Algorithms: differences between the averages of NER obtained with the true and the randomised runs as a function of the number of evaluations.

Table 13.3: List of settings and parameters used for the Genetic Algorithms.

Parameter	
response to be maximized	NER
classification method	LDA
population size	30 chromosomes
average number of variables	
selected in the starting population	5
mutation probability	0.01
cross over probability	0.5
stop criterion	maximum no. evaluations
maximum no. of evaluations	120
number of runs	500

score is the one with the second highest frequency and so on.

Forward selection, coupled with LDA, has been also applied on the retained score matrix. The selection algorithm has been repeated five times, by removing and predicting each validation group each time and

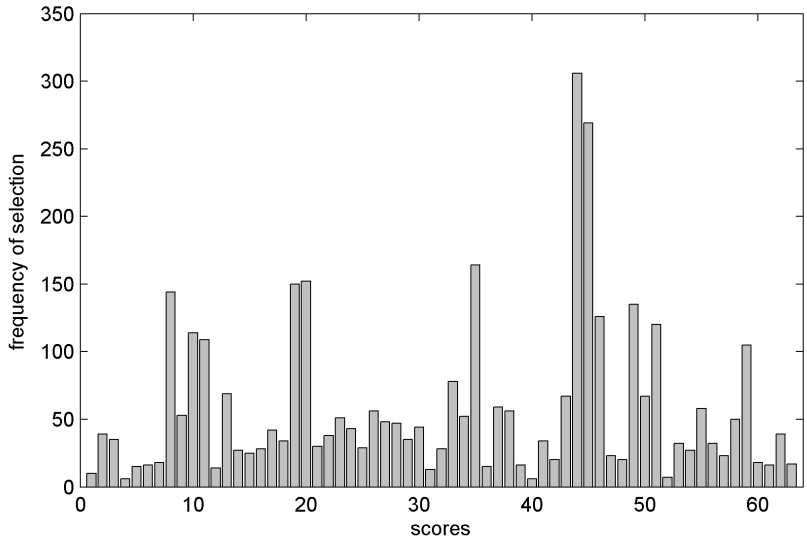

Figure 13.2: Genetic Algorithms: histogram of frequency of score selection.

adding up to 15 scores. Then the scores have been ranked on the basis of their frequencies of selection (Figure 13.4), weighting their frequencies in a proportional way with respect to the position reached in the forward selection sequence: the highest rank value (15) was assigned to the score selected as the first one, the second highest value (15-1=14) to the score selected as the second one, and so on. As it is possible to see, score no. 8 has always been selected as the first best score for all the 5 cross validation steps, getting the highest value of 75 (75=15x5).

In Figure 13.5 the cross validated NER as a function of included scores is represented. On the basis of these results, 10 scores have been included in the final LDA model, where the first considered score is the one with the highest weighted frequency, the second score is the one with the second highest weighted frequency and so on.

ECVA has also been applied on the same matrix of retained scores, in order to compare its performances on the entire chromatogram and on the retained scores: 11 components have been selected as significant, on

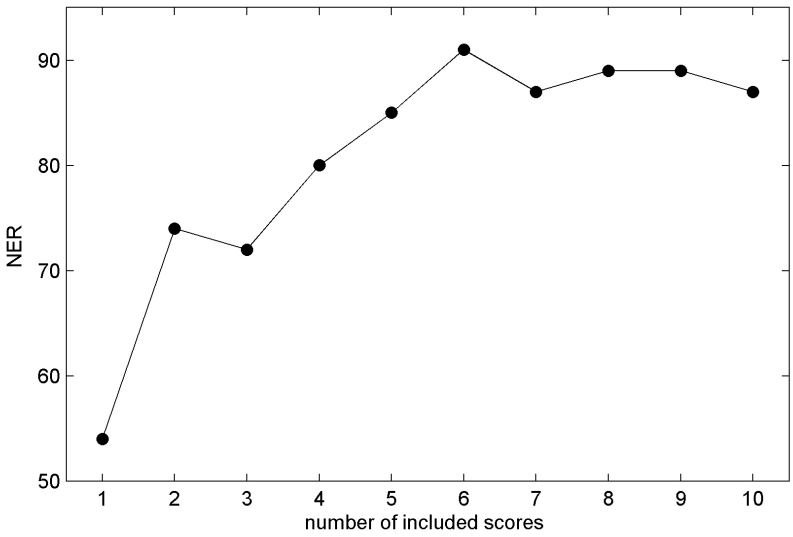

Figure 13.3: Genetic Algorithms: cross validated NER as a function of included scores.

Table 13.4: Non Error Rate (NER, % of correctly assigned samples) achieved with different classification methods. FIT refers to the training samples, CV to the cross validated samples, TEST to the external test samples. The number of used factors (ECVA) or selected scores (GAs and Forward Selection) is also reported.

method	**factor or scores**	**FIT**	**CV**	**TEST**
GA-LDA	6	93	91	81
Forward LDA	10	89	80	89
ECVA	11	96	89	81

the basis of the cross validated NER. The achieved NER of the described classification methods are showed in Table 13.4.

The three methods give acceptable models; despite that GAs and ECVA give best results on the training set (NER equal to 93% and 96%, respectively) and on the cross validation groups (91% and 89%), while the model achieved by means of Forward Selection has a better performance on the test set. With respect to the models achieved on the entire

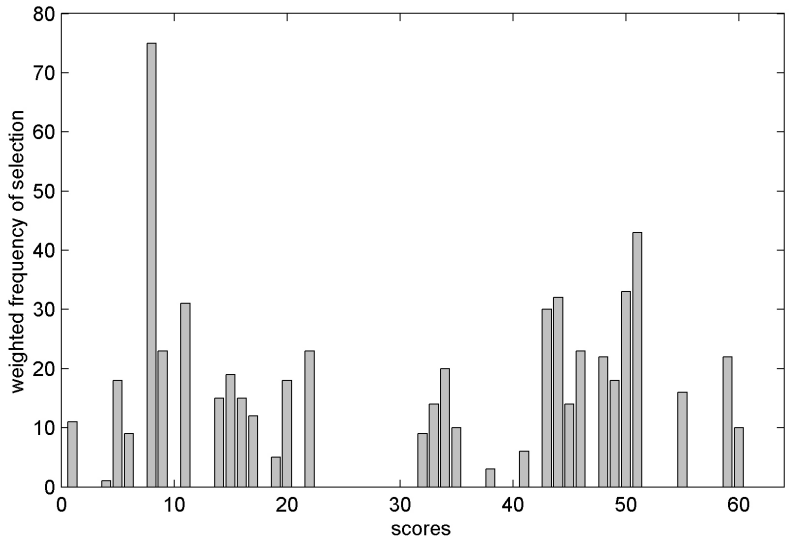

Figure 13.4: Forward Selection: histogram of weighted frequency of score selection.

chromatographic profile by means of PLS-DA and LDA coupled with PCA (Table 13.2), all the models obtained on the reduced score matrix give better predictions: the difference of the worst cross validated NER obtained on the reduced data (80%) and the best NER obtained on the entire profile (60%) is considerable, and the same conclusion can be observed on the test set (81% and 68% respectively). Only the ECVA model on the entire chromatographic profile gives comparable results.

Considerations on selected windows

Finally, in order to compare the scores retained by the three different approaches, the canonical weights (Figure 13.6) calculated in the ECVA model have been considered: the scores with absolute values of the relative canonical weights higher than one, i.e. with a relevant role in the model, are resumed in Table 13.5, together with the scores selected by GAs and Forward Selection. It is interesting to notice that several scores are common in the three different approaches (scores 8 and 44), while scores

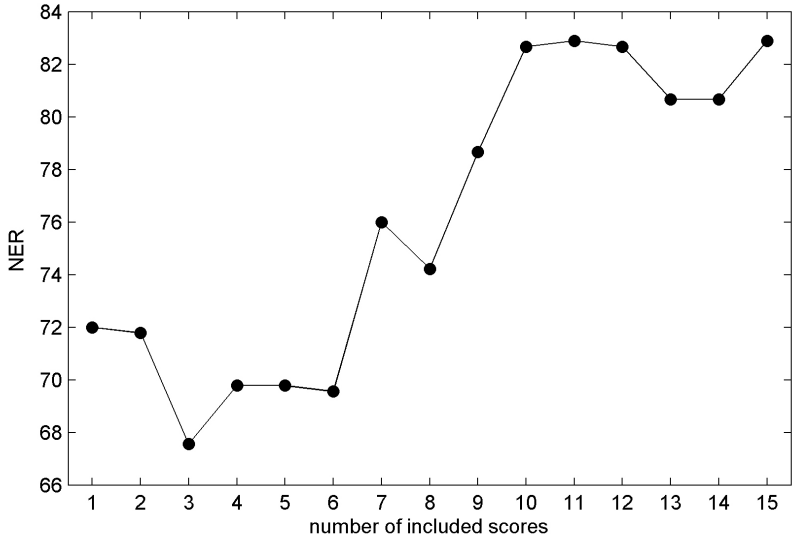

Figure 13.5: Forward Selection: cross validated NER as a function of included scores.

Table 13.5: Lists of the scores selected by GAs, Forward Selection and with a relevant weight (higher then 1) in the first and second ECVA variates. Common scores are marked in bold.

Classification methods	**Scores**
GA-LDA	**44**, **45**, 35, 20, 19, **8**
FORWARD LDA	**8**, 51, **50**, **44**, 11, 43, 9, 22, 46, 48
ECVA, variate 1	**8**, 10, 23, **44**, **50**, 59
ECVA, variate 2	6, **8**, 30, 33, 37, 38, **45**, 46, 49, **50**

with high loadings in ECVA are also selected by GAs (45) or Forward Selection (50). Despite the applied classification and selection methods are based on different approaches, the same scores are recognised as significant fingerprints for the class discrimination, i.e. they contain the relevant information for the classification task.

Moreover, considering that each score is extracted from a specific window and more scores can be extracted from the same window or score extracted from adjacent windows can highlight the same phenomena (chro-

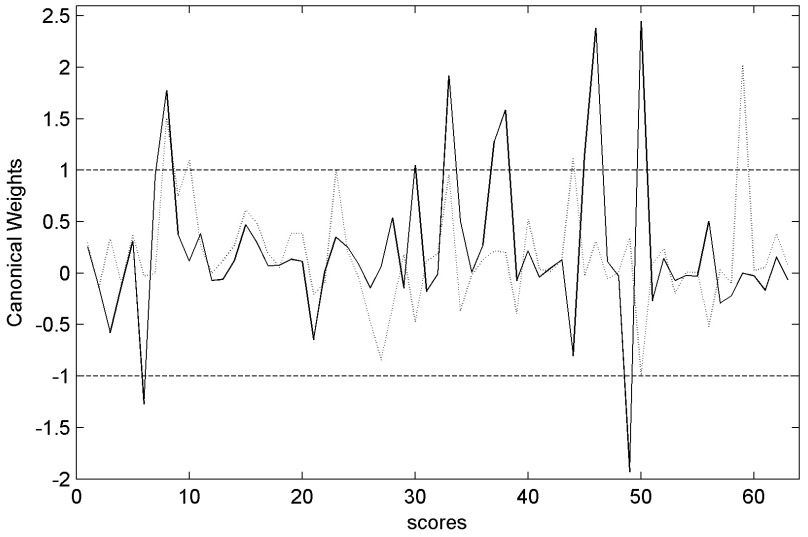

Figure 13.6: Extended Canonical Variates Analysis: canonical weights on the first (dotted line) and second (solid line) variates.

matographic peak), it is also important to highlight which are the relevant windows in the original chromatographic profile. The first two scores selected by GAs are number 44 and 45 and these refer to two different windows of the chromatographic profile, with ranges 1501-1510 and 1531-1540, respectively; Forward Selection includes scores no. 43 and 44, which refer to two adjacent windows (1491-1500 and 1501-1510) and to a unique chromatographic peak. Score no. 46 is selected by Forward Selection and refers to a window (1541-1550) which includes the descending part of the same peak described by score 45. Therefore, if windows are considered, there are more overlaps between the results achieved by the used classification methods.

On the other hand, the differences in the results can be due to the different approaches: the sequence of scores included by Forward Selection can depend on the first selected scores, since all the others are added to the model when these scores are still present and consequently the new scores can just contribute to solve marginal discriminations in the data. GAs

explore in a more complete way the available information and searches for the best solution with an higher number of possibilities. This can be confirmed by the better results achieved by GAs with respect to Forward Selection, even if this last method is less time-consuming and give in this study classification models with acceptable predictive performances.

Modelling on original windows

Finally, in order to check the consistency of the proposed approach for dimension reduction of data and in order to avoid final models based only on local principal components, PLS-DA and ECVA have been applied on the original data using the primarily selected windows, i.e. using all the windows where at least one score was extracted. Therefore, 56 windows have been considered (as a result of the score extraction explained in paragraph 4.3) and a data matrix with 560 columns used for the classification tasks, where each of this columns represents an original variable in the retained windows. PLS-DA (with 7 factors) gives NER equal to 76% and 75% on the cross validation and test groups, respectively. By comparing these results with the NER obtained on the entire data (Table 13.2), it is possible to see that now the model quality is improved, since the NER increase in a considerable way. On the other hand, ECVA gave the same results obtained on the original data (NER equal to 87% and 81% on the cross validation and test groups, respectively) and consequently proves again to be an optimal approach for the retention of useful information in classification tasks.

In conclusion, different classification methods, such as PLS-DA, LDA and ECVA, in combination with variable selection approaches (GAs and Forward Selection) have been compared, evaluating their performances in the geographical discrimination of chromatographic profiles of wine samples. Since variable selection techniques can not handle a huge number of variables, a new approach based on Principal Component Analysis has been tested, in order to extract significant features from high dimensional data. The local useful information extracted from specific windows of the

original spectral profile can be collected in a matrix and this solution can perform a significant reduction in collinear high dimensional data, which can be subsequently treated with the proposed variable selection methods.

The results achieved with the proposed data reduction are better then the ones obtained on the entire chromatographic profile, with the exception of ECVA, which gives acceptable classification models in both cases. Moreover, some windows of the original chromatographic profile have been selected by both the variable selection approaches, confirming that the extraction score method seems to be able to capture the significant local information that can be subsequently treated by different multivariate techniques. The results indicate that for high-dimensional data, the compression part plays a more significant role than the specific variable selection method in achieving reasonable results.

The achieved results permit also to plan a future extension of the score extraction method to multiway data, by applying PCA on unfolded three dimensional windows, and to subsequently perform variable selection.

chapter 14

Conclusions and perspectives

The presented chemometric applications and the literature reviews indicate that multivariate methods hold several possibilities and advantages in the characterisation of chemical and physical fingerprints of food products. The literature survey reveals an increasing number of research in the field within the last decades. Hopefully, the increasing research activities can address the problems of chemometric applications on food products and further explore the applications of the method.

The present study demonstrates that chemometrics is able to provide valuable information on the characterisation of food products. Despite the fact that multivariate analysis has been widely applied on food science, chemometric methods have been shown to be successfully applicable on innovative analytical methods, such as electronic sensors and mechanical and acoustical signals, in order to provide information about the quality of the food product including authenticity and influence of processing.

Technical trends and options

Three novel chemometric approaches have been proposed, in order to face with different tasks linked to the characterisation of food fingerprints.

The Classification And Influence Matrix Analysis (CAIMAN) deals with classification based on leverage-scaled functions (chapter 4): the proposed approach seems to show several advantages. First of all, it shows-on an average basis-good performance when compared to the other popular methods. CAIMAN is able to solve in a very simple and intuitive way classification problems related to tipicity, pathologies, and single class characterisation. Like QDA, UNEQ, KNN, SIMCA and CART, CAIMAN does not to suffer of nonlinear class separability. The a priori probability based on the object frequency is implicitly assumed in the leverage algorithm, while, as a disadvantage, a good representation of the classes is necessary, i.e. high enough objects/variables ratios.

Regarding the novel similarity measure for sequential data based on the Hasse matrix (chapter 5), the proposed approach shows several advantages: (a) it seems able to link each electronic nose time profile to a meaningful mathematical term (the Hasse matrix), which can be consequently treated and studied by multivariate analysis; (b) the Hasse matrices and the corresponding distances are calculated with a simple algorithm; (c) the Hasse distance is standardised, allowing a natural interpretation of the results; (d) the distances consider the whole time profile, i.e. no parameterisation is needed and the time information is preserved; (e) the distances can be obtained by a flexible strategy (the weights) depending on the aim of the analysis.

Finally, regarding the proposed reduction of sequential data dimension (chapter 6) coupled with variable selection, the local useful information extracted from specific windows of the sequential profiles performs a significant reduction in collinear high dimensional data. The results achieved with the proposed data reduction are better then the ones obtained on the entire chromatographic profile, confirming that the extraction score method seems to be able to capture the significant local information that can be subsequently treated by different multivariate techniques. Moreover, the results indicate that for high-dimensional data, the compression part plays a more significant role than the specific variable selection method in achieving reasonable results. The achieved results permit also

to plan a future extension of the score extraction method to multiway data, by applying PCA on unfolded three dimensional windows, and to subsequently perform variable selection.

In order to apply the proposed algorithms, some routines have been developed during the thesis by using the software MatLab 6.5 (MathWorks), in order to diffuse the proposed methodologies. The code sources of these routines and the web sites where it is possible to download them are presented in the appendix.

Prospective applications

The modern possibility to perform rapid and advanced analytical analyses makes chemometrics a needed choice as a screening method, both in food production and in regulatory affairs, in order to get a rapid evaluation of the amount of information provided by analytical methods. Food authenticity is a term which refers to whether food purchased by the consumer matches its description. Recently, the development of rapid methods for confirming the authenticity of food products has become an important research area. Branded and high-quality products with a geographic affiliation are in considerable demand and therefore such products are vulnerable to economic adulteration. Because of the high capacity of handling huge amount of data and extract useful information, chemometrics has the potential to reveal minor differences in food products that can be related to its authenticity.

For example, in 2002, the Office of Pharmaceutical Science of the U.S Food and Drug Administration (FDA) launched the Process Analytical Technology (PAT) initiative. FDAs definition of PAT is: A system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality. The PAT strategy calls for relevant tools for process control, and in this context multivariate data analysis is a candidate because of the ability to perform analyses on-line at site and to

handle complex chemical systems.

List of Figures

List of Tables

Bibliography

Alcazar, A., Pablos, F., Martin, M. J., and Gonzalez, A. G. (2002). Multivariate characterisation of beers according to their mineral content. *Talanta*, **57**, 45–52. data. [citation in 1.2]

Alves, M. R., Cunha, S. C., Amaral, J. S., Pereira, J. A., and Oliveira, M. B. (2005). Classification of pdo olive oils on the basis of their sterol composition by multivariate analysis. *Analytica Chimica Acta*, **549**, 166–178. [citations in 1.1 and 8.1]

Ampuero, S., Bogdanov, S., and Bosset, J. O. (2004). Classification of unifloral honeys with an ms-based electronic nose using different sampling modes: Shs, spme and index. *European Food Research and Technology*, **218**, 198–207. [citation in 1.2]

Andersson, C. and Bro, R. (2000). The n-way toolbox for matlab. *Chemometrics and Intelligent Laboratory Systems*, **52**, 1–4. [citation in 12.2]

Angerosa, F., Basti, C., and Vito, R. (1999). Virgin olive oil volatile compounds from lipoxygenase pathway and characterization of some italian cultivars. *Journal of Agricultural and Food Chemistry*, **47**, 836–839. [citation in 10.2.1]

Antonelli, A., Cocchi, M., Fava, P., Foca, G., Franchini, G. C., Manzini, D., and Ulrici, A. (2004). Automated evaluation of food colour by means

of multivariate image analysis coupled to a wavelet-based classification algorithm. *Analytica Chimica Acta*, **515**, 3–13. [citation in 1.2]

A.P. Worth, A. and Cronin, M. (1999). Embedded cluster modelling: a novel method for analysing embedded data sets. *Quantitative Structure-Activity Relationships*, **18**, 229–235. [citation in 4.5]

Aparicio, R. and Luna, G. (2002). Characterisation of monovarietal virgin olive oils. *European Journal of Lipid Science and Technology*, **104**, 614–627. [citation in 8.1]

Aparicio, R., Morales, M. T., and Alonso, V. (1997). Authentication of european virgin olive oils by their chemical compounds, sensory attributes, and consumers' attitudes. *Journal of Agricultural and Food Chemistry*, **45**, 1076–1083. [citation in 1.1]

Armanino, C., Leardi, R., Lanteri, S., and Modi, G. (1989a). Chemometric analysis of tuscan olive oils. *Chemometrics and Intelligent Laboratory Systems*, **5**, 343–354. [citations in 1.1, 4.5, 8.1, and 10.1.1]

Armanino, C., Lanteri, S., Forina, M., Balsamo, A., Migliardi, M., and Cenderelli, G. (1989b). Hirsutism: a multivariate approach of feature selection and classification. *Chemometrics and Intelligent Laboratory Systems*, **5**, 335–341. [citation in 4.5]

Arvanitoyannis, I. S., Katsota, M. N., Psarra, E., Soufleros, E. H., and Kallithraka, S. (1999). Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science & Technology*, **10**, 321–336. [citation in 1.1]

Aznar, M., Lopez, R., and Cacho, J.and Ferreira, V. (2003). Prediction of aged red wine aroma properties from aroma chemical composition. partial least squares regression models. *Journal of Agricultural and Food Chemistry*, **51**, 2700–2707. [citation in 7.2.1]

Ballabio, D., Mannino, S., and Regazzoni, C. (2004). Gli indici di desiderabilità e utilità nel processo. *ICP*, **11**, 38–40. [citations in 7.1 and 7.2.4]

Ballabio, D., Kokkinofta, R., Todeschini, R., and Theocharis, C. R. (2006a). Characterization of the traditional cypriot spirit zivania by means of counterpropagation artificial neural networks. *Chemometrics and Intelligent Laboratory Systems*, **in press**. [citations in 7.1 and 7.2.3]

Ballabio, D., Cosio, M. S., Mannino, S., and Todeschini, R. (2006b). A chemometric approach based on a novel similarity/diversity measure for the characterisation and selection of electronic nose sensors. *Analytica Chimica Acta*, **578**, 170–177. [citations in 5.4, 7.1, and 9.1]

Ballabio, D., Mauri, A., Todeschini, R., and Buratti, S. (2006c). Geographical classification of wine and olive oil by means of caiman (classification and influence matrix analysis). *Analytica Chimica Acta*, **570**, 249–258. [citations in 7.1 and 8.1]

Barker, M. and Rayens, W. S. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, **17**, 166–173. [citation in 3.2]

Bartlett, P. N., Elliot, J. M., and Gardner, J. W. (1997). Electronic nose and their application in the food industry. *Food Technology*, **51**, 44–48. [citations in 9.1 and 10.2.1]

Baumann, K. and Stiefl, N. (2004). Validation tools for variable subset regression. *Journal of Computer-Aided Molecular Design*, **18**, 549–562. [citations in 3.5, 8.1, and 11.2]

Benedetti, S., Mannino, S., Sabatini, A. G., and Marcazzan, G. L. (2004). Electronic nose and neural network use for the classification of honey. *Apidologie*, **35**, 397–402. [citation in 1.2]

Benito, M. J., Ortiz, M. C., Sanchez, M. S., Sarabia, L. A., and Iniguez, M. (1999). Typification of vinegars from jerez and rioja using classical chemometric techniques and neural network meth. *Analyst*, **124**, 547–552. [citations in 1.2, 4.5, and 8.1]

Bianchi, G., Giansante, L., Shaw, A., and Kell, D. B. (2001). Chemometric criteria for the characterisation of italian protected denomination of

origin (dop) olive oils from their metabolic profiles. *European Journal of Lipid Science and Technology*, **103**, 141–150. [citation in 8.1]

Boggia, R., Zunin, P., Lanteri, S., Rossi, N., and Evangelisti, F. (2002). Classification and class-modeling of riviera ligure extra-virgin olive oil using chemical-physical parameters. *Journal of Agricultural and Food Chemistry*, **50**, 2444–2449. [citations in 1.1 and 8.1]

Borges, A. and Peleg, M. (1996). Determination of the apparent fractal dimension of the force-displacement curves of brittle snacks by four different algorithms. *Journal of Texture Studies*, **27**, 243–255. [citation in 11.3]

Boselli, E., Boulton, R. B., Thorngate, J. H., and Frega, N. G. (2004). Chemical and sensory characterization of doc red wines from marche (italy) related to vintage and grape cultivars. *Journal of Agricultural and Food Chemistry*, **52**, 3843–3854. [citation in 7.2.1]

Breiman, L. J., Friedman, J. H., Olsen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Pacific Grove, CA. [citation in 4.1]

Brescia, M. A., Caldarola, V., De Giglio, A., Benedetti, D., Fanizzi, F. P., and Sacco, A. (2002). Characterization of the geographical origin of italian red wines based on traditional and nuclear magnetic resonance spectrometric determinations. *Analytica Chimica Acta*, **458**, 177–186. [citation in 8.1]

Brezmes, J., Llobet, E., Vilanova, X., Ortis, J., Saiz, G., and Correig, X. (2001). Correlation betweens electronic nose signals and fruit quality indicators on shelf-life measurements with pinklady apples. *Sensors and Actuators B*, **80**, 41–50. [citation in 9.1]

Bro, R. (1997). Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, **38**, 149–171. [citation in 3.4]

Bro, R. (1998). Multi-way analysis in the food industry. [citation in 3.4]

Bro, R. (1999). Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics and Intelligent Laboratory Systems*, **46**, 133–147. [citation in 1.2]

Bro, R., Andersson, C. A., and Kiers, H. A. L. (1999). Parafac 2 - part ii. modeling chromatographic data with retention time shifts. *Journal of Chemometrics*, **13**, 295–309. [citation in 3.4]

Brodnjak-Voncina, D., Kodba, Z. C., and Novic, M. (2005). Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. *Chemometrics and Intelligent Laboratory Systems*, **75**, 31–43. [citations in 1.1, 4.5, and 8.1]

Bruggemann, R. and Bartel, H. G. (1999). A theoretical concept to rank environmentally significant chemicals. *Journal of Chemical Information and Computer Science*, **39**, 211–217. [citations in 5.1 and 5.2]

Buratti, S., Benedetti, S., Scampicchio, M., and Pangerod, E. C. (2004). Characterization and classification of italian barbera wines by using an electronic nose and an amperometric electronic tongue. *Analytica Chimica Acta*, **525**, 133–139. [citation in 1.1]

Buratti, S., Ballabio, D., Benedetti, S., and Cosio, M. S. (2006). Prediction of italian red wine sensorial descriptors from electronic nose, electronic tongue and spectrophotometric measurements by means of genetic algorithm regression models. *Food Chemistry*, **100**, 211–218. [citations in 7.1, 7.2.1, 8.2, and 9.1]

Burden, F. R., Brereton, R. G., and Walsh, P. T. (1997). Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. *Analyst*, **122**, 1015–1022. [citations in 3.5 and 8.1]

Camean, A. M., Moreno, I., Lopez-Artiguez, M., Repetto, M., and Gonzalez, A. G. (2001). Differentiation of spanish brandies according to their metal content. *Talanta*, **54**, 53–59. data. [citation in 1.2]

Campanella, L., Favero, G., Pastorino, M., and Tomassetti, M. (1999). Monitoring the rancidification process in olive oils using a biosensor operating in organic solvents. *Biosensors & Bioelectronics*, **14**, 179–186. [citations in 10.1.1 and 10.2.1]

Carcione, F., Chiesa, L., Ballabio, D., Cattaneo, C., Cantoni, C., and Biondi, P. (2006). Fatty acid composition in capon meat as markes of feeding traceabilty. *Journal of Chromatografic*, **submitted**. [citations in 7.1 and 7.2.2]

Cardoso, D. R., Andrade-Sobrinho, L. G., Leite-Neto, A. F., Reche, R. V., Isique, W. D., Ferreira, M. M. C., Lima-Neto, B. S., and Franco, D. W. (2004). Comparison between cachacua and rum using pattern recognition methods. *Journal of Agricultural and Food Chemistry*, **52**, 3429–3433. data. [citation in 1.2]

Carpino, S., Acree, T. E., Barbano, D. M., Licitra, G., and Siebert, K. J. (2002). Chemometric analysis of ragusano cheese flavor. *Journal of Agricultural and Food Chemistry*, **50**, 1143–1149. [citation in 1.2]

Casanas, R., Gonzalez, M., Rodriguez, E., Marrero, A., and Diaz, C. (2002). Chemometric studies of chemical compounds in five cultivars of potatoes from tenerife. *Journal of Agricultural and Food Chemistry*, **50**, 2076–2082. data. [citation in 1.2]

Cerrato Oliveros, C., Perez Pavon, J. L., Garcia Pinto, C., Fernandez Laespada, E., Moreno Cordero, B., and Forina, M. (2002). Electronic nose based on metal oxide semiconductor sensors as a fast alternative for the detection of adulteration of virgin olive oils. *Analytica Chimica Acta*, **459**, 219–228. [citations in 1.1, 9.1, and 10.1.1]

Cerrato Oliveros, C., Boggia, R., Casale, M., Armanino, C., and Forina, M. (2005). Optimisation of a new headspace mass spectrometry instrument. discrimination of different geographical origin olive oils. *Journal of Chromatography A*, **1076**, 7–15. [citation in 8.1]

Charlton, A. J., Farrington, W. H. H., and Brereton, P. (2002). Application of 1h nmr and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee. *Journal of Agricultural and Food Chemistry*, **50**, 3098–3103. [citation in 1.2]

Christenses, J., Miquel Becker, E., and Frederiksen, C. S. (2005). Fluorescence spectroscopy and parafac in the analysis of yogurt. *Chemometrics and Intelligent Laboratory Systems*, **75**, 201–208. [citation in 1.2]

Christy, A. A., Kasemsumran, S., Du, Y., and Ozaki, Y. (2004). The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics. *Analytical Sciences*, **20**, 935–940. [citation in 1.1]

Cocchi, M., Corbellini, M., Foca, G., Lucisano, M., Pagani, M. A., Tassi, L., and Ulrici, A. (2005). Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on nir spectra. *Analytica Chimica Acta*, **544**, 100–107. [citation in 1.2]

Community, E. (1987). Regulation 823/1987. [citation in 8.1]

Community, E. (1991). Regulation 2568/91. [citation in 10.2.1]

Community, E. (1992). Regulation 2081/1992. [citations in 8.1 and 10.1.1]

Community, E. (2002). Regulation 796/02. [citation in 10.2.1]

Community, E. (2003). Regulation 1989/03. [citation in 10.2.5]

CompStat, p. o. C. S., editor (1974). *A stepwise discriminant analysis program using density estimation*. CompStat. [citation in 4.5]

Cook, R. D. and Weisberg, S. (1982). *Residual and Influence in Regression*. Chapman, New York. [citations in 4.1 and 4.2.1]

Coomans, D., Derde, M. P., Broeckaert, I., and Massart, D. L. (1981). Potential methods in pattern recognition. *Analytica Chimica Acta*, **133**, 241–250. [citation in 4.1]

Cosio, M. S., Ballabio, D., Benedetti, S., and Gigliotti, C. (2006). Geographical origin and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks. *Analytica Chimica Acta*, **567**, 202–210. [citations in 1.1, 7.1, 8.2, 9.4, and 10.1.3]

Cosio, M. S., Ballabio, D., Benedetti, S., and Gigliotti, C. (2007). Evaluation of different storage conditions of extra virgin olive oils with an innovative recognition tool built by means of electronic nose and electronic tongue. *Food Chemistry*, **101**, 485–491. [citations in 7.1 and 10.2.3]

Cozzolino, D., Smyth, H. E., and Gishen, M. (2003). Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins. *Journal of Agricultural and Food Chemistry*, **51**, 7703–7708. data. [citation in 9.1]

Cozzolino, D., Smyth, H. E., Cynkar, W., Dambergs, R. G., and Gishen, M. (2005). Usefulness of chemometrics and mass spectrometry-based electronic nose to classify australian white wines by their varietal origin. *Talanta*, **68**, 382–387. [citation in 1.1]

Derde, M. P. and Massart, D. L. (1986). Uneq: a disjoint modelling technique for pattern recognition based on normal distribution. *Analytica Chimica Acta*, **184**, 33–51. [citation in 4.1]

Distante, C., Leo, M., Siciliano, P., and Persaud, K. C. (2002). On the study of feature extraction methods for an electronic nose. *Sensors and Actuators B*, **87**, 274–288. [citation in 9.4]

Drake, B. (1963). Food crushing sounds. an introductory study. *Journal of Food Sceince*, **28**, 233–241. [citation in 11.1]

Duizer, L. (2001). A review of acoustic research for studying the sensory perception of crisp, crunchy and rackly textures. *Trends in Food Science & Technology*, **12**, 17–24. [citation in 11.1]

Dunn III, W. J. and Wold, S. (1980). Structure-activity analyzed by pattern recognition: the asymmetric case. *Journal of Medicinal Chemistry*, **23**, 595–599. [citation in 4.1]

Eddib, O. and Nickless, G. (1987). Elucidation of olive oil classification by chemometrics. *Analyst*, **112**, 391–395. [citations in 1.1, 8.1, and 10.1.1]

Eklov, T., Martensson, P., and Lundstrom, I. (1999). Selection of variables for interpreting multivariate gas sensor data. *Analytica Chimica Acta*, **381**, 221–232. [citation in 9.1]

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188. [citation in 4.5]

Flath, R. A., Forrey, R. R., and Guadagni, D. G. (1973). Aroma components of olive oil. *Journal of Agricultural and Food Chemistry*, **21**, 948–952. [citation in 10.2.1]

Forina, M. (1990). Artificial data produced by m. forina, university of genova. [citation in 4.5]

Forina, M. and Drava, G. (1997). Chemometrics for wine. applications. *Analusis*, **25**, 38–42. [citations in 1.1 and 8.1]

Forina, M., Armanino, C., and Lanteri, S. (1982). Acidi grassi degli animali acquatici: uno studio chemiometrico. *Rivista Italiana Scienze Alimentari*, **11**, 15–22. [citation in 4.5]

Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983). *Food research and Data Analysis*, chapter Classification of olive oils from their fatty acid composition, pages 189–214. Applied Science Publishers, London. [citation in 8.2]

Forina, M., Armanino, C., Castino, M., and Ubigli, M. (1986). Multivariate data analysis as discriminating method of the origin of wines. *Vitis*, **25**, 189–201. [citations in 4.5 and 8.1]

Forina, M., Armanino, C., Leardi, R., and Drava, G. (1991). A class-modelling technique based on potential functions. *Journal of Chemometrics*, **5**, 435–453. [citation in 4.1]

Frank, I. E. (1988). Dasco: a new classification method. *Chemometrics and Intelligent Laboratory Systems*, **4**, 215–222. [citation in 4.1]

Frank, I. E. and Friedman, J. H. (1989). Classification: oldtimers and newcomers. *Journal of Chemometrics*, **3**, 463–475. [citation in 4.1]

Frias, S., Conde, J. E., Rodriguez-Bencomo, J. J., Garcia-Montelongo, F. J., and Perez-Trujillo, J. P. (2003). Classification of commercial wines from the canary islands (spain) by chemometric techniques using metallic contents. *Talanta*, **59**, 335–344. [citation in 1.1]

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175. [citation in 4.1]

Gallardo, J., Alegret, S., and del Valle, M. (2005). Application of a potentiometric electronic tongue as a classification tool in food analysis. *Talanta*, **66**, 1303–1309. [citations in 9.1, 10.1.1, and 10.2.1]

Garcia-Gonzalez, D. L. and Aparicio, R. (2003a). Detection of defective virgin olive oils by metal-oxide sensors. *European Food Research and Technology*, **215**, 118–123. [citation in 8.1]

Garcia-Gonzalez, D. L. and Aparicio, R. (2003b). Virgin olive oil quality classification combining neural network and mos sensors. *Journal of Agricultural and Food Chemistry*, **51**, 3515–3519. [citation in 9.1]

Garcia-Gonzalez, D. L. and Aparicio, R. (2004). Classification of different quality virgin olive oils by metal-oxide sensors. *European Food Research and Technology*, **218**, 484–487. [citations in 1.1 and 8.1]

Gardner, J. W. and Bartlett, P. N. (1993). Brief history of electronic nose. *Sensors and Actuators B*, **18**, 211–217. [citations in 9.1 and 10.1.1]

Gerrild, G. and Lantz, R. (1990). Chemical analysis of 75 crude oil samples from pliocene sand units. Technical report, U. S. Geol. Surv. [citation in 4.5]

Gibson, L. and Ashby, M. (1988). *Cellular solids: Structure and Properties*. Pergamon Press. [citation in 11.1]

Golbraikh, A. and Tropsha, A. (2002). Beware of q2! *Journal of Molecular Graphics and Modelling*, **20**, 269–276. [citation in 3.5]

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley. [citation in 3.3]

Gonzalez, G. and Pena-Mendez, E. M. (2000). Multivariate data analysis in classification of must and wine from chemical measurements. *European Food Research and Technology*, **212**, 100–107. [citations in 1.1 and 8.1]

Gonzalez Martin, Y., Perez Pavon, J. L., Moreno Cordero, B., and Garcia Pinto, C. (1999). Classification of vegetable oils by linear discriminant analysis of electronic nose data. *Analytica Chimica Acta*, **384**, 83–94. [citations in 1.1, 9.1, and 10.2.1]

Goodacre, R., Kell, D. B., and Bianchi, G. (1992). Neural networks and olive oil. *Nature*, **359**, 594–594. [citation in 10.1.1]

Guadarrama, A., Rodriguez-Mendez, M. L., de Saja, J. A., Rios, J. L., and Olias, J. M. (2000). Array of sensors based on conducting polymers for the quality control of the aroma of the virgin olive oil. *Sensors and Actuators B*, **69**, 276–282. [citations in 1.1, 9.1, and 10.2.1]

Guadarrama, A., Rodriguez-Mendez, M. L., Sanz, C., Rios, J. L., and de Saja, J. A. (2001). Electronic nose based on conducting polymers for the quality control of the olive oil aroma. discrimination of quality, variety of olive and geographic origin. *Analytica Chimica Acta*, **432**, 283–292. [citations in 1.1 and 10.1.1]

Guimet, F., Boque, R., and Ferre, J. (2004). Cluster analysis applied to the exploratory analysis of commercial spanish olive oils by means of excitation-emission fluorescence spectroscopy. *Journal of Agricultural and Food Chemistry*, **52**, 6673–6679. [citations in 1.1 and 10.1.1]

Guimet, F., Ferre, J., and Boque, R. (2005). Rapid detection of olive-pomace oil adulteration in extra virgin olive oils from the protected denomination of origin siurana using excitation-emission fluorescence spectroscopy and three-way methods of analysis. *Analytica Chimica Acta*, **544**, 143–152. [citation in 10.1.1]

Halfon, E. and Reggiani, M. G. (1986). On ranking chemicals for environmental hazard. *Environmental Science & Technology*, **20**, 1173–1179. [citations in 5.1 and 5.2]

Hamilton, L. (2001). Cross-shelf colour zonation in northern great barrier reef lagoon surficial sediments. *Australian Journal of Earth Science*, **48**, 193–200. [citation in 4.5]

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester (UK). [citation in 4.1]

Hendriks, M. M. W. B., de Boer, J. H., Smilde, A., and Doornbos, D. A. (1992). Multicriteria decision making. *Chemometrics and Intelligent Laboratory Systems*, **16**, 175–191. [citation in 7.2.4]

Hines, E. L., Llobet, E., and Gardner, J. W. (1999). Neural netwrok based electronic nose for apple ripeness determination. *Electronics letters*, **35**, 821–823. [citation in 9.1]

James, M. (1985). *Classification Algorithms*. Collins, London (UK). [citation in 4.1]

Jennrich, R. J. (1977). *Stepwise discriminant analysis*. Wiley, NewYork (USA). [citations in 3.3, 4.1, and 8.1]

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, volume 3ed. Prentice-Hall. [citations in 4.5 and 8.1]

Juriaeskay, I. and Veress, G. E. (1985). Prima: a new pattern recognition method. *Analytica Chimica Acta*, **171**, 61–76. [citation in 4.1]

Kallithraka, S., Arvanitoyannis, I. S., Kefalas, P., El-Zajouli, A., Soufleros, E., and Psarra, E. (2001). Instrumental and sensory analysis of greek wines: implementation of principal component analysis (pca) for classification according to geographical origin. *Food Chemistry*, **73**, 501–514. [citations in 1.1 and 8.1]

Karoui, R., Mazerolles, G., and Dufour, E. (2003). Spectroscopic techniques coupledwith chemometric tools for structure and texture determinations in dairy products. *International Dairy journal*, **13**, 607–620. [citation in 1.2]

Kaufman, L. and Rousseau, P. J. (1990). *Finding groups in data. An Introduction to cluster analysis*. Wiley. [citations in 4.5 and 4.3.2]

Keller, H. R., Massart, D. L., and Brans, J. P. (1991). Multicriteria decision making: A case study. *Chemometrics and Intelligent Laboratory Systems*, **11**, 175–189. [citation in 7.2.4]

Kelly, J. F. D., Downey, G., and Fouratier, V. (2004). Initial study of honey adulteration by sugar solutions using midinfrared (mir) spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry*, **52**, 33–39. [citation in 1.2]

Kennard, R. W. and Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, **11**, 137–148. [citation in 13.3]

Kim, J., Mowat, A., Poole, P., and Kasabov, N. (2000). Linear and nonlinear pattern recognition models for classification of fruit from visible-near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **51**, 201–216. [citation in 1.2]

Kokkinofta, R. and Theocharis, C. R. (2005). Chemometric characterization of the cypriot spirit zivania. *Journal of Agricultural and Food Chemistry*, **53**, 5067–5073. [citations in 1.2 and 7.2.3]

Kokkinofta, R., Petrakis, P. V., Mavromoustakos, T., and Theocharis, C. R. (2003). Authenticity of the traditional cypriot spirit zivania on the basis of metal content using a combination of coupled plasma spectroscopy and statistical analysis. *Journal of Agricultural and Food Chemistry*, **51**, 6233–6239. [citation in 7.2.3]

Kosir, I. J., Kocjancic, M., Ogrinc, N., and Kidric, J. (2001). Use of snif-nmr and irms in combination with chemometric methods for the determination of chaptalisation and geographical origin of wines (the example of slovenian wines). *Analytica Chimica Acta*, **429**, 195–206. [citations in 1.1 and 8.1]

Kowalski, B. R. and Bender, C. F. (1972). The k-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Analytical Chemistry*, **44**, 1405–1411. [citation in 4.1]

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of t to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27. not in file. [citation in 3.1]

Kvalheim, O. M. (1987). Latent-structure decompositions (projections) of multivariate data. *Chemometrics and Intelligent Laboratory Systems*, **2**, 283–290. [citation in 3.1]

Lachenmeier, D. W., Frank, W., Humpfer, E., Schafer, H., Keller, S., Mortter, M., and Spraul, M. (2005). Quality control of beer using high-resolution nuclear magnetic resonance spectroscopy and multivariate analysis. *European Food Research and Technology*, **220**, 215–221. [citation in 1.2]

Lanteri, S., Armanino, C., Perri, E., and Palopoli, A. (2002). Study of oils from calabrian olive cultivars by chemometric methods. *Food Chemistry*, **76**, 501–507. [citations in 1.1 and 8.1]

Leardi, R. (2000). Application of genetic algorithm-pls for feature selection in spectral data sets. *Journal of Chemometrics*, **14**, 643–655. [citation in 6.1]

Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics*, **15**, 559–569. [citation in 3.3]

Leardi, R. and Lupianez, A. (1998). Genetic algorithms applied to feature selection in pls regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, **41**, 195–207. [citations in 13.2 and 13.3]

Leardi, R. and Norgaard, L. (2004). Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics*, **18**, 486–497. [citation in 11.2]

Leardi, R., Boggia, R., and Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, **6**, 267–281. [citations in 3.3 and 6.1]

Legin, A., Rudnitskaya, A., Lyova, L., Vlasov, Y., Di Natale, C., and D'Amico, A. (2003). Evaluation of italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. *Analytica Chimica Acta*, **484**, 33–44. [citations in 10.1.1 and 10.2.1]

Legin, A., Rudnitskaya, A., Seleznev, B., and Vlasov, Y. (2005). Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie. *Analytica Chimica Acta*, **534**, 129–135. [citation in 1.2]

Lewi, P. J. and Boey, J. V. H. (1992). Multicriteria decision making using pareto optimality and promethee preference ranking. *Chemometrics and Intelligent Laboratory Systems*, **16**, 139–144. [citation in 7.2.4]

Llobet, E., Hines, E. L., Gardner, J. W., and Franco, S. (1999). Non-destructive banana ripeness determination using a neural network-based

electronic nose. *Measurement Science and Technology*, **10**, 538–548. [citations in 1.2 and 9.1]

Lopez, B., Latorre, M. J., Fernandez, M. I., Garcia, M. A., Garcia, S., and Herrero, C. (1996). Chemometrics classification of honeys according to their type based on quality control data. *Food Chemistry*, **55**, 281–287. [citation in 1.2]

Maeztu, L., Andueza, S., Ibanez, C., Paz de Pena, M., Bello, J., and Cid, C. (2001). Multivariate methods for characterization and classification of espresso coffees from different botanical varieties and types of roast by foam, taste, and mouthfeel. *Journal of Agricultural and Food Chemistry*, **49**, 4743–4747. [citation in 1.2]

Mager, P. (1991). *Design Statistics in Pharmacochemistry*. Research Studies Press. [citation in 4.5]

Mannina, L., Patumi, M., Proietti, N., Bassi, D., and Segre, A. L. (2001). Geographical characterization of italian extra virgin olive oils using high-field 1h nmr spectroscopy. *Journal of Agricultural and Food Chemistry*, **49**, 2687–2696. [citations in 8.1 and 10.1.1]

Mannino, S., Buratti, S., Cosio, M. S., and Pellegrini, N. (1999). Evaluation of the 'antioxidant power' of olive oils based on a fia system with amperometric detection. *Analyst*, **124**, 1115–1118. [citations in 10.1.1 and 10.2.1]

Marini, F., Zupan, J., and Magri, A. L. (2004a). On the use of counterpropagation artificial neural networks to characterize italian rice varieties. *Analytica Chimica Acta*, **510**, 231–240. [citation in 1.2]

Marini, F., Magri, A. L., Balestrieri, F., Fabretti, F., and Marini, D. (2004b). Supervised pattern recognition applied to the discrimination of the floral origin of six types of italian honey samples. *Analytica Chimica Acta*, **515**, 117–125. [citation in 1.2]

Martens, H. and Naes, T. (1989). *Multivariate calibration*. Wiley. [citations in 1.3 and 6.1]

Martin, M. J., Pablos, F., and Gonzalez, A. G. (1999). Characterization of arabica and robusta roasted coffee varieties andmixture resolution according to their metal content. *Food Chemistry*, **66**, 365–370. [citation in 1.2]

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., de Jong, S., Lewi, P. J., and Smeyers Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam. [citations in 1.1 and 7.2.4]

Mattiello, S., Todeschini, R., Tripaldi, P., and Crimella, P. (1993). Influenza dellaltitudine e della specie su alcuni parametri ematici e valori biochimici del siero di sangue in camelidi sudamericani. *Rivista Agricoltura Subtropicale e Tropicale*, **87**, 231–244. [citation in 4.5]

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York. [citations in 3.2 and 4.1]

Miyashita, Y., Takahashi, Y., Takayama, C., Ohkubo, T., Fumatsu, K., and Sasaki, S. (1986). Computer-assisted structure/taste studies on sulfamates by pattern recognition methods. *Analytica Chimica Acta*, **184**, 143–149. [citation in 4.5]

Moller, J. K. S., Parolari, G., Gabba, L., Christenses, J., and Skibsted, L. H. (2003). Monitoring chemical changes of dry-cured parma ham during processing by surface autofluorescence spectroscopy. *Journal of Agricultural and Food Chemistry*, **51**, 1224–1230. [citation in 1.2]

Morales, M. T., Rios, J. J., and Aparicio, R. (1997). Changes in the volatile composition of virgin olive oil during oxidation: Flavors and off-flavors. *Journal of Agricultural and Food Chemistry*, **45**, 2666–2673. [citation in 10.2.1]

Norgaard, L., Bro, R., Westad, F., and Engelsen, S. B. (2006). A modification of canonical variates analysis to handle highly collinear multivariate data. *Journal of Chemometrics*, **in press**. [citation in 3.2]

Nozal Nalda, M. J., Bernal Yague, J. L., Diego Calva, J. C., and Martin Gomez, M. T. (2005). Classifying honeys from the soria province of spain via multivariate analysis. *Analytical and Bioanalytical Chemistry*, **382**, 311–319. [citation in 1.2]

Ortiz, M. C., Gutierrez, A. H., Sanchez Pastor, M. S., Sarabia, L. A., and Iniguez, M. (1995). The uneq, pls and mlf neural network methods in the modelling and prediction of the colour of young red wines from the denomination of origin 'rioja'. *Chemometrics and Intelligent Laboratory Systems*, **28**, 273–285. [citation in 1.1]

Ortiz, M. C., Sarabia, L. A., Symington, C., Santamaria, F., and Iniguez, M. (1996). Analysis of ageing and typification of vintage ports by partial least squares and soft independent modelling class analogy. *Analyst*, **121**, 1009–1013. [citation in 1.1]

Oshima, T., Hopia, A., German, B., and Frankel, E. N. (1996). Determination of hydroperoxides and structures by hplc with post-column detection with diphenyl-1- pyrenylphosphina. *Lipids*, **31**, 1091–1091. [citation in 10.2.1]

Ovejero-Lopez, I., Bro, R., and Bredie, W. L. P. (2005). Univariate and multivariate modelling of flavour release in chewing gum using time-intensity: a comparison of data analytical methods. *Food Quality and Preference*, **16**, 327–343. [citations in 1.2 and 12.2]

Pavan, M. and Todeschini, R. (2004). New indices for analysing partial ranking diagrams. *Analytica Chimica Acta*, **515**, 167–181. [citations in 5.1 and 5.2]

Perez-Magarino, S., Ortega-Heras, M., and Gonzalez-San Jose, M. L. (2002). Multivariate classification of rose wines from different spanish

protected designations of origin. *Analytica Chimica Acta*, **458**, 187–190. [citation in 8.1]

Petrakis, P. V., Touris, I., Liouni, M., Zervou, M., Kyrikou, Y., Kokkinofta, R., Theocharis, C. R., and Mavromoustakos, T. (2005). Authenticity of the traditional cypriot spirit zivania on the basis of 1h nmr spectroscopy diagnostic parameters and statistical analysis. *Journal of Agricultural and Food Chemistry*, **53**, 5293–5303. [citation in 7.2.3]

Peyvieux, C. and Dijksterhuis, G. (2001). Training a sensory panel for ti: a case study. *Food Quality and Preference*, **12**, 19–28. [citation in 12.2]

Piazza, L., Gigli, J., and Ballabio, D. (2006). On the application of chemometrics for the study of acoustic-mechanical properties of crispy bakery products. *Chemometrics and Intelligent Laboratory Systems*, **in press**. [citations in 7.1 and 11.1]

Piggott, J. R. (2000). Dynamism in flavour science and sensory methodology. *Food Research International*, **33**, 191–197. [citation in 12.1]

Pilar Marti, M., Busto, O., and Guasch, J. (2004). Application of a headspace mass spectrometry system to the differentiation and classification of wines according to their origin, variety and ageing. *Journal of Chromatography A*, **1057**, 211–217. [citation in 9.1]

Pinheiro, P. B. M. and Esteves da Silva, J. C. V. (2005). Chemometric classification of olives from three portuguese cultivars of olea europaea l. *Analytica Chimica Acta*, **544**, 229–235. [citations in 1.1 and 8.1]

Raatikainen, O., Reinikainen, V., Minkkinen, P., Ritvanen, T., Muje, P., Pursiainen, J., Hiltunen, T., Hyvonen, P., von Wright, A., and Reinikainen, S. (2005). Multivariate modelling of fish freshness index based on ion mobility spectrometry measurements. *Analytica Chimica Acta*, **544**, 128–134. [citation in 1.2]

Rajer-Kanduc, K., Zupan, J., and Majcen, N. (2003). Separation of data on the training and test set for modelling: a case study for modelling of

five colour properties of a white pigment. *Chemometrics and Intelligent Laboratory Systems*, **65**, 221–229. [citation in 13.3]

Reid, L. M., O'Donnell, C. P., Kelly, J. F. D., and Downey, G. (2004). Preliminary studies for the differentiation of apple juice samples by chemometric analysis of solid-phase microextraction-gas chromatographic data. *Journal of Agricultural and Food Chemistry*, **52**, 6891–6896. [citation in 1.2]

Reinbach, H. C., Meinert, L., Ballabio, D., Aaslyng, M., Bredie, W., Olsen, K., and Moller, P. (2006). Interactions between oral burn, meat flavour and texture in chili spiced pork patties evaluated by time-intensity. *Food Quality and Preference*, **submitted**. [citation in 7.1]

Rencher, A. C. (2002). *Methods Of Multivariate Analysis*, volume 2. Wiley, England. [citation in 8.1]

Resmini, R., Pellegrino, L., and Bertuccioli, M. (1986). Moderni criteri per la valutazione chimico-analitica della tipicita‘ di un formaggio: lesempio del parmigiano-reggiano. *Rivista Italiana Scienze Alimentari*, **15**, 315–326. [citation in 4.5]

Roudaut, G., Dacremont, C., Valles Pamies, B., Colas, B., and Le Meste, M. (2002). Crispness: a critical review on sensory and material science approaches. *Trends in Food Science & Technology*, **13**, 217–227. [citation in 11.1]

Salter, G. J., Lazzari, M., Giansante, L., Goodacre, R., Jones, A., Surricchio, G., Kell, D. B., and Bianchi, G. (1997). Determination of the geographical origin of italian extra virgin olive oil usyng pyrolysis mass spectrometry and artificial neural networks. *Journal of Analytical and Applied Pyrolysis*, **40-41**, 159–170. [citations in 8.1 and 10.1.1]

Saviozzi, A., Lotti, G., and Piacenti, D. (1986). La composizione amminoacidica delle farine di girasole. *Rivista Italiana di Scienze Alimentari*, **15**, 437–444. [citation in 4.5]

Sawyer, J., Wood, C., Shanahan, D., Gout, S., and McDowell, D. (2003). Real-time pcr for quantitative meat species testing. *Food Control*, **14**, 579–583. [citation in 1.2]

Seymour, S. and Hammann, D. (1988). Crispness and crunchiness of selected low moisture foods. *Journal of Texture Studies*, **19**, 79–95. [citation in 11.1]

Skov, T. and Bro, R. (2005). A new approach for modelling sensor based data. *Sensors and Actuators B*, **106**, 719–729. [citations in 9.2 and 9.4]

Skov, T., Van den Berg, F., Tomasi, G., and Bro, R. (2006). Automated alignment of chromatographic data. *Journal of Chemometrics*, **submitted**. [citation in 13.3]

Smilde, A., Bro, R., and Geladi, P. (2004). *Multi-way analysis with Applications in the Chemical Sciences*. Wiley, Chichester. paper. [citations in 3.4, 3.4, and 12.1]

Speaks, C. (1999). *Introduction to Sound*. Singular Publishing. [citation in 11.1]

Stahle, L. and Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: a monte carlo study. *Journal of Chemometrics*, **1**, 185–196. [citations in 3.2 and 4.1]

Stella, R., Barisci, J. N., Serra, G., Wallace, G. G., and De Rossi, D. (2000). Characterisation of olive oil by an electronic nose based on conducting polymer sensors. *Sensors and Actuators B*, **63**, 1–9. [citations in 8.1, 9.1, and 10.1.1]

Streuli, H. (1987). Mathematische modelle fur die chemische zusammensetzung von libensmitteln und ihre bedeutung fur deren beurteilung. *Lebensm.- Technol.*, **20**, 203–211. [citation in 4.5]

Tapp, H. S., Defernez, M., and Kemsley, E. K. (2003). Ftir spectroscopy and multivariate analysis can distinguish the geographic origin of extra

virgin olive oils. *Journal of Agricultural and Food Chemistry*, **51**, 6110–6115. [citation in 8.1]

Todeschini, R. (1990). Artificial data produced and based on the example proposed by breiman, friedman e olsen in classification and regression trees, wadswarth and brooks, ca (usa). [citation in 4.5]

Todeschini, R., Ballabio, D., Consonni, V., Mauri, A., and Pavan, M. (2005). Caiman (classification and influence matrix analysis): a new approach to the classification based on leverage-scaled functions. *Chemometrics and Intelligent Laboratory Systems*, **in press**. [citations in 4.1 and 8.1]

Todeschini, R., Consonni, V., Mauri, A., and Ballabio, D. (2006). On the characterization of dna primary sequences by a new similarity/diversity measure based on the partial ordering. *Journal of Chemical Information and Modeling*, **46**, 1905–1911. [citation in 5.4]

Todeschini, R., Ballabio, D., Consonni, V., and Mauri, A. (2007). A new similarity/diversity measure for sequential data. *MATCH communications in mathematical and in computer chemistry*, **57**, 51–67. [citation in 5.4]

Tsimidou, M. and Karakostas, K. (1993). Geographical classification of greek virgin olive oil by non-parametric multivariate evaluation of fatty acid composition. *Journal of the Science of Food and Agriculture*, **62**, 253–257. [citations in 8.1 and 10.1.1]

Tsimidou, M., Macrae, R., and Wilson, I. (1987). Authentication of virgin olive oils using principal components analysis of triglyceride and fatty acid profiles: part 1. classification of greek olive oils. *Food Chemistry*, **25**, 227–239. [citations in 1.1, 8.1, and 10.1.1]

Tsimidou, M., Papadopoulos, G., and Boskow, D. (1992). Determination of phenolic compounds in virgin olive oil by reversed-phase hplc with emphasis on uv detection. *Food Chemistry*, **44**, 53–60. [citation in 10.2.1]

Van der Voet, H. and Coenegracht, P. M. (1988). The evaluation of probabilistic classification methods. part 2. comparison of simca, alloc, classy and lda. *Analytica Chimica Acta*, **209**, 1–27. [citation in 4.1]

Vandeginste, B. G. M., Massart, D. L., Buydens, L. M. C., de Jong, S., Lewi, P. J., and Smeyers Verbeke, J. (1998). *Handbook of Chemometrics and Qualimetrics. Part B.* Elsevier, The Netherlands. [citation in 4.1]

Vichi, S., Castellote, A. I., Pizzale, L., Conte, L. S., Buxaderas, S., and Lopez-Tamames, E. (2003). Analysis of virgin olive oil volatile compounds by headspace solid-phase microextraction coupled to gas chromatography with mass spectrometric and flame ionization detection. *Journal of Chromatography A*, **983**, 19–33. [citation in 10.2.1]

Vickers, Z. and Bourne, M. (1976). A psychoacoustical theory of crispness. *Journal of Food Science*, **41**, 1158–1164. [citation in 11.1]

Vinaixa, M., Marin, S., Brezmes, J., Llobet, E., Vilanova, X., Correig, X., Ramos, A., and Sanchis, V. (2004). Early detection of fungal growth in bakery products by use of an electronic nose based on mass spectrometry. *Journal of Agricultural and Food Chemistry*, **52**, 6068–6074. [citations in 1.2 and 9.1]

Vinaixa, M., Vergara, A., Duran, C., Llobet, E., Badia, C., Brezmes, J., Vilanova, X., and Correig, X. (2005). Fast detection of rancidity in potato crisps using e-noses based on mass spectrometry or gas sensors. *Sensors and Actuators B*, **106**, 67–75. [citations in 1.2 and 9.1]

Winsberg, S. and Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended euclidean model. *Psychometrika*, **54**, 217–219. not in file. [citation in 3.1]

Wold, S. (1972). S. spline functions, a new tool in data-analysis. *Kemisk Tidskrift*, **84**, 34–37. [citation in 1.1]

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, **8**, 127–139. [citation in 4.1]

Wold, S. (1990). Chemometrics; what do we mean with it, and what do we want from it? *Chemometric and Intelligent Laboratory Systems*, **35**, 109–115. [citation in 1.1]

Wold, S., Esbensen, K. H., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**, 37–52. [citation in 3.1]

Wu, W., Walczak, B., Massart, D. L., Heuerding, S., Erni, F., Last, I. R., and Prebble, K. A. (1996). Artificial neural networks in classification of nir spectral data: Design of the training set. *Chemometrics and Intelligent Laboratory Systems*, **33**, 35–46. [citation in 13.3]

Zalacain, A., Ordoudi, S. A., Diaz-Plaza, E. M., Carmona, M., Blazquez, I., Tsimidou, M., and Alonso, G. L. (2005). Near-infrared spectroscopy in saffron quality control: Determination of chemical composition and geographical origin. *Journal of Agricultural and Food Chemistry*, **53**, 9337–9341. [citation in 1.2]

Zupan, J. (1994). Introduction to artificial neural network (ann) methods: What they are and how to use them. *Acta Chimica Slovenica*, **41**, 327–352. [citation in 3.2]

Zupan, J. and Gasteiger, J. (1999). *Neural Networks in Chemistry and Drug Design*, volume Second. VCH, Weinheim. [citation in 3.2]

Zupan, J., Novic, M., Li, X., and Gasteiger, J. (1994). Classification of multicomponent analytical data of olive oils using different neural networks. *Analytica Chimica Acta*, **292**, 219–234. [citations in 1.1, 8.1, and 10.1.1]

Zupan, J., Novic, M., and Ruisanchez, I. (1997). Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemometrics and Intelligent Laboratory Systems*, **38**, 1–23. [citation in 3.2]

List of publications

Food Science

Ballabio, D., Cosio, M.S., Mannino, S., Todeschini, R. (2006). A chemometric approach based on a novel similarity/diversity measure for the characterization and selection of electronic nose sensors. *Analytica Chimica Acta*, **578**, 170-177

Piazza, L., Gigli, J., Ballabio, D. (2006). On the Application of Chemometrics for the study of Acoustic-Mechanical properties of Crispy Bakery products. *Chemometrics and Intelligent Laboratory Systems*, in press, available on-line

Ballabio, D., Kokkinofta, R., Todeschini, R., Theocharis, C.R. (2006). Characterization of the traditional Cypriot spirit Zivania by means of Counterpropagation Artificial Neural Networks. *Chemometrics and Intelligent Laboratory Systems*, in press, available on-line

Ballabio, D., Mauri, A.,Todeschini, R., Buratti, S. (2006). Geographical classification of wine and olive oil by means of CAIMAN (Classification And Influence Matrix Analysis). *Analytica Chimica Acta*, **570**, 249-258

Cosio, M.S., Ballabio, D., Benedetti, S., Gigliotti, C. (2006). Geographical characterization and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks. *Analytica Chimica Acta*, **567**, 202-210

Cosio, M.S., Ballabio, D., Benedetti, S., Gigliotti, C. (2007). Evaluation of different storage conditions of extra virgin olive oils with a innovative recognition tool built by means of electronic nose and electronic tongue. *Food Chemistry*, **101**, 485-491

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

Carcione, F., Chiesa, L.M., Ballabio, D., Cattaneo, C., Cantoni, C., Biondi, P.A. (2006). Fatty acid composition in capon meat as markes of feeding traceabilty. *Journal of Chromatografic*, submitted

Reinbach, H.C., Meinert, L., Ballabio, D., Aaslyng, M.D., Bredie, W.L.P., Olsen, K., Mller, P. (2006). Interactions between oral burn, meat flavour and texture in chili spiced pork patties evaluated by Time-Intensity *Food Quality and Preference*, submitted

Chemometrics

Todeschini, R., Consonni,V., Mauri, A., Ballabio, D. (2006). On the characterization of DNA primary sequences by a new similarity/diversity measure based on the partial ordering. *Journal of Chemical Information and Modeling* , **46**, 1905-1911

Todeschini, R., Ballabio, D., Consonni,V., Mauri, A. (2007). A new similarity/diversity measure for sequential data. *MATCH Communications in Mathematical and in Computer Chemistry*, **57**, 51-67

Todeschini, R., Ballabio, D., Consonni,V., Mauri, A., Pavan, M. (2006). CAIMAN (Classification And Influence Matrix Analysis): a new approach to the classification based on leverage-scaled functions. *Chemometrics and Intelligent Laboratory Systems*, in press, available on-line

Fermo, P., Cariati, F., Ballabio, D., Consonni, V., Bagnasco, G. (2004). Classification of ancient Etruscan ceramics using statistical multivariate analysis of data. *Applied Physics A*, **79**, 299-307

Consonni,V., Mauri, A., Ballabio, D., Todeschini, R. (2006). A new similarity/diversity measure for the characterization of DNA sequences. *Croatica Chemica Acta*, submitted

Mauri, A.,Ballabio, D.,Manganaro, A.,Todeschini, R. (2006) Molecular descriptors relationships. Part 1. Searching for optimal molecular descriptor subsets by ranking methods applied on a large data set. *Journal of Chemical Information and Modeling* , submitted

Others

Buratti, S., Benedetti, S., Ballabio, D., Cosio, M.S. (2006). Tecniche innovative combinate "naso elettronico" e "lingua elettronica" per la predizione di descrittori sensoriali di vini rossi secchi mediante l'uso degli Algoritmi Genetici. *Ingredienti Alimentari*, in press

Cosio, M.S., Gigliotti, C., Ballabio, D., Benedetti, S., Buratti, S. (2005). Geographical characterization of extra virgin olive oils by electronic nose. *7th "congresso di scienza e tecnologia degli alimenti" (ciseta)*, congress proceedings.

Ballabio, D., Schiraldi, A. (2005). Multi-way analysis on sensory data: application on time-intensity evaluation of chilli spiced pork patties. *10th Workshop on the developments in the Italian PhD Research in Food Science Technology*, workshop proceedings.

Ballabio, D., Regazzoni, C., Mannino, S. (2004). Gli indici di Desiderabilit e Utilit nel processo. *ICP*, november, 38-40

Ballabio, D., Magni, M. (2004). Oltre il Plug and Play: uso e abuso dell'ICT a supporto delle decisioni manageriali. *Ticonzero*, web journal

Software and code

During the PhD thesis, some chemometric algorithms have been developed. In order to make them available, these algorithms have been written as Matlab (Mathworks) modules and can be downloaded, as explained in the following paragraphs.

CAIMAN

CAIMAN is a classification method exploiting the properties of the diagonal terms of the influence matrix, also called leverages (chapter 4). Different MATLAB modules for facing the three CAIMAN methods (Asymmetric, Discriminant and Modelling) have been built. Moreover, a graphical interface for MATLAB has been built, in order to facilitate the user. With these modules, an help file containing a full explanation about the CAIMAN classification approach, together with an overview on the classification and the most common methods is also provided. The provided help file contains also a full explanation about the MATLAB modules of CAIMAN. The modules have been developed in collaboration with the Milano Chemometrics and QSAR Research Group and are provided here: www.disat.unimib.it/chm/

Hasse Matlab modules

The **Hasse** Matlab modules calculate the Hasse distances between several kinds of data sequences. In particular, the hasse_enose module calculates Hasse distances between samples of each electronic nose sensor. If a class is defined for the samples, these distances can be also used to sort the sensors on the basis of their class discrimination capability. These modules have been developed during the application of this technique to electronic nose data (chapter 5). In order to run these modules, Matlab (Mathworks) is needed. The modules have been developed in collaboration with the Milano Chemometrics and QSAR Research Group and are provided here: www.disat.unimib.it/chm/

DAUS

DAUS, Desirability And Utility Software, is a software for the calculation of desirability and utility indices. This software has been built in order to face with the Multicriteria Decison Making for process monitoring (paragraph 7.2.4). Basically, the data can be loaded from a text file. After that, in the DAUS set parameters form, it is possible to define (save and load) the utility/desirability function settings. Some variable properties (mean, minimum and maximum values) are also provided for each variable. Finally, after Desirability and Utility indices have been calculated, it is possible to see and save numerical and graphical results. An help file is also provided together with the software. DAUS has been built in Visual Basic and the setup exe file will be available soon.