DEVELOPMENT OF GENETIC ALGORITHM BASED CLASSIFICATION AND CLUSTER ANALYSIS METHODS FOR ANALYTICAL DATA

A Thesis Submitted to

the Graduate School of Engineering and Sciences of

İzmir Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Betül ÖZTÜRK

January 2010

İZMİR

We approve the thesis of **Betül ÖZTÜRK**

Assoc.Prof.Dr. Durmuş ÖZDEMİR
Supervisor

Prof.Dr. F. Nil ERTAŞ
Committee Member

Prof.Dr. Ahmet E. EROĞLU
Committee Member

Prof.Dr. Cevdet DEMİR
Committee Member

Assoc.Prof.Dr. Figen TOKATLI
Committee Member

08 January 2010

Prof.Dr. Levent ARTOK
Head of the Department of Chemistry

Assoc.Prof.Dr. Talat YALÇIN
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGEMENT

It takes a long time to finish my PhD study. I would like to express my thanks to the people who have been very helpful to me during the time it took me to write this thesis.

First, I thank my advisor Assoc.Prof. Durmuş ÖZDEMİR, for his continuous support in the PhD program. He is responsible for involving me in the chemometrics in the first place. He thought me how to ask questions and express my ideas. I also thank the other two PhD thesis committee members; Prof. Ahmet E. EROĞLU and Assoc.Prof. Figen TOKATLI. All of three showed me different ways to approach a research problem and the need to be persistent to accomplish my goal. They were always there to listen and to give advice.

During my PhD studies, I collected a lot of data. Data are very essential for my study. I thank to TARİŞ Olive and Olive Oil Co-operatives for the olive oil samples. I am pleased to pay my thanks to Technological Research Council of Turkey (TÜBİTAK) for funding the project (107T037) and providing scholarship.

Furthermore I am deeply indebted to my colleagues İbrahim KARAMAN and Ayşegül YALÇIN that have provided the environment for sharing their experiences about the problem issues involved as well as participated in stimulating team exercises developing solutions to the identified problems. Also I would like to thank to Didem ŞEN for her support and kindness.

Finally I want to thank my lovely dad M. Sadi ÖZTÜRK and mom Sabire ÖZTÜRK for their encouragement and powerful source of inspiration and energy. I specially thank to my brother Cahit ÖZTÜRK for listening my complaints and frustrations and for believing in me. A special thought is devoted to my parents for a never ending support.

ABSTRACT

DEVELOPMENT OF GENETIC ALGORITHM BASED CLASSIFICATION AND CLUSTER ANALYSIS METHODS FOR ANALYTICAL DATA

In this study genetic algorithm based classification and clustering methods were aimed to develop for the spectral data. The developed methods were completely achieved hybridization of nature inspired algorithm (genetic algorithms, GAs) to other classification or clustering methods. The first method was genetic algorithm based principal component analysis (GAPCAD), and the second was genetic algorithm based discriminant analysis (GADA). Both methods were performed to achieve the best discrimination between the olive oil and vegetable oil samples. The classifications of samples were examined directly from their spectral data obtained from using near infrared spectrometry, Fourier transform infrared (FTIR) spectrometry, and spectrofluorometry. The GA was used to optimize the performance of classification or clustering techniques' on training set in order to maximize the correct classification of acceptable and unacceptable samples or samples of dissimilar properties and to reduce the spectral data by wavelength selection. After GA optimization the classification results of training set were controlled by validation set. Lastly, the success of both algorithms was compared to the results of PCA and SIMCA.

ÖZET

ANALİTİK VERİLER İÇİN GENETİK ALGORİTMA TEMELLİ SINIFLANDIRMA VE ÖBEK ANALİZİ METOTLARI GELİŞTİRİLMESİ

Bu çalışmada spektral veriler için genetik algoritmaya dayalı sınıflandırma ve öbekleme yöntemlerinin geliştirilmesi amaçlanmıştır. Geliştirilen yöntemler, genetik algoritma gibi doğayı model alarak, var olan çeşitli sınıflandırma ve kümeleme yöntemlerine uyarlanmıştır. İlk olarak genetik algoritmaya dayalı temel bileşen analiz yöntemi (Genetic Algorithm based Principle Component analysis, GAPCA), ikinci olarak da genetik algoritmaya dayalı diskriminant analizi yöntemi (Genetic Algorithm based Discriminant Analysis, GADA) geliştirilmiştir. Her iki yöntem de zeytin yağları ve bitkisel yağların sınıflandırılması için kullanılmıştır. Bu süreçte, yakın infrared (NIR), Fourier dönüşümlü infrared (FTIR) ve floresans spektroskopi verileri kullanılmıştır. Sınıflandırma ve kümeleme yöntemlerinin performansını iyileştirmek ve spektral verileri daraltarak dalga boyu seçimini gerçekleştirmek için genetik algoritma kullanılmış, böylelikle farklı özellikler üzerinden uyumlu ve uyumsuz örneklerin sınıflandırılması gerçekleştirilmiştir. Genetik algoritmanın optimizasyonunu takiben, sınıflandırma sonuçları farklı bir test setiyle karşılaştırılmıştır. Son olarak geliştirilen her iki algoritmanın da başarısı PCA ve SIMCA yöntemlerinden elde edilen sonuçlarla karşılaştırılmıştır.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figures **Page**

CHAPTER 1

INTRODUCTION

The use of spectroscopic techniques is continuously increasing rapidly and this increase lets more research on spectroscopic data. There are two main reasons in increasing the usage of spectroscopic techniques. The first one is mainly based on the development of computer techniques. By the help of the computer sciences, the mathematical techniques can be easily applied to the spectroscopic data and this application provides more precise results, better methods in both qualitative and quantitative analysis techniques and lastly more knowledge on samples interested. Secondly, spectroscopic techniques are fast, have to be non-invasive methods and provide more than hundreds of data among the other instrumental and traditional techniques. The advantages of spectroscopic techniques lead to usage of these techniques both in industry and academic research. As it is known, spectroscopy is a general term that concerns the interactions of various types of radiation with matter. These interactions between the radiation and the matter interested depend on the energy of radiation, the types and numbers of the atoms and molecules that are present in the matter, and lastly the influence of sample matrix. By the help of these effects, it was realized that spectroscopic signal was mainly proportional to the amount of sample. The spectroscopic signal of pure samples at a specific wavelength provides information both in quantitative and qualitative means. However, the amplitude of signal often changes on the complexity of sample interested. The contamination of other constituents presented in the sample cause the changes of amplitude of a spectroscopic signal and less accurate prediction results in the concerning component. The more constituents cause the more complex and overlapped spectroscopic signal and also a spectrum contains hundreds or thousands of wavelengths/wavenumbers with their corresponding instrumental signal. Traditionally, these complexities have been diminished using hyphenated systems such as gas chromatography–infrared spectroscopy. Firstly, gas chromatography separates the components than each component analyzed by infrared spectrometer. In last fifty years, to solve the complexity of a spectrum signal and predict the results accurately, chemometric techniques are used in both qualitative and

quantitative manner. In quantitative analysis, multivariate calibration methods such as principal component regression (PCR), partial least squares (PLS), and genetic regression (GR) are used to predict the concentrations or amounts of constituents interested. For the qualitative analysis, classification and clustering techniques are used to identify the class or cluster of samples.

In this thesis, classification methods will be investigated mainly in two parts. Generally classification methods are divided into two ways named as: supervised and unsupervised classification. In supervised classification methods, the chemical or physical properties of samples are known previously and the class of unknown sample is identified. Soft independent methodology of class analogies (SIMCA), K-nearest neighbor (KNN) are the mostly used supervised methods. On the other hand, in the unsupervised classification methods such as principal component analysis (PCA), hierarchical cluster analysis (HCA), there is no knowledge about the samples and these samples are firstly clustered in order to identify the boundaries of clusters then clusters of unknown samples are identified. As mentioned before in spectral analyses hundreds or thousands of variables are observed for each sample. To observe the desired information classical classification techniques will not be enough. Therefore, hyphenated classification methods will be preferred for the data reduction. Nowadays, the genetic algorithms are mostly used algorithms. In this thesis study, the development of genetic algorithm based supervised classification methods will be investigated. Two different algorithms were developed named as distance based genetic algorithm principal component analysis (GAPCAD), genetic algorithm based discriminant analysis (GADA) and these were examined using spectral data. Both classification methods were designed supervised classification methods. Three different spectroscopic techniques were used to examine the developed algorithms. Near infrared spectroscopy (NIR), middle infrared spectroscopy (MIR), and fluorescence spectroscopy were the studied techniques. The spectral analysis of food samples was preferred due to the non-invasive property of spectroscopy. Olive oil samples and vegetable oil samples were chosen as samples and spectral data matrix of each sample set was observed. These spectral data matrices were used in the examination of GAPCAD and GADA.

This thesis is divided into six parts. In the first part (Chapter 2) the basic theory of spectroscopic techniques will be investigated. Chapter 3 concerns the basic principles of supervised and unsupervised classification methods. Chapter 4 describes the working principles of two developed genetic algorithm based classification methods in details. In

the next chapter, the samples that are used in case studies will be explained and the importance of classification of these samples will be discussed. In chapter 6 the experimental procedure of used methods will be described. The major part of this thesis study consisting of the major results of the developed two new algorithms will be given in Chapter 7.

CHAPTER 2

SPECTROSCOPY

2.1. Spectroscopy

Spectroscopy is a general term that concerns the interaction of various types the electromagnetic radiation with matters or substances. Electromagnetic radiation is described by using electromagnetic spectrum. Figure 2.1 shows the wavelengths and frequencies of radiations in an enormous range.

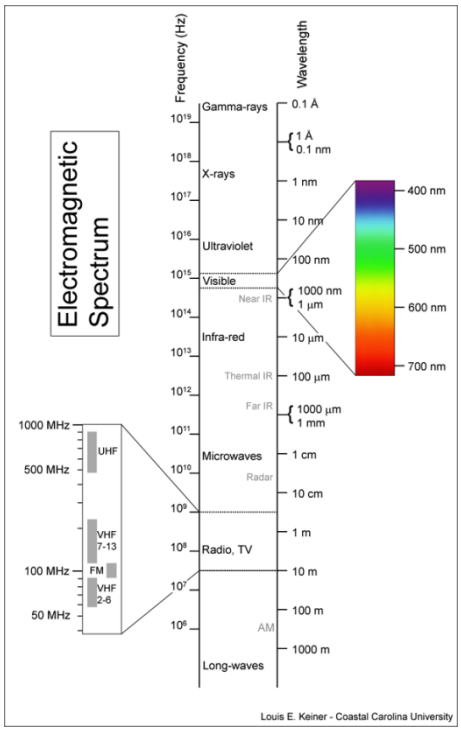


Figure 2.1 Regions of electromagnetic spectrum.

In spectroscopic techniques generally the energetic level of matter or substance is increased by the external enforced energy beams. Therefore spectroscopic techniques can be categorized into different groups depending on the wavelength range used for the energy beam. For instance, the technique concern the interaction of the electromagnetic

radiation exists in 200 to 1000 nm with matter is called as ultraviolet–visible (UV–Vis) spectroscopy. Most of the spectroscopic techniques measure the energy that absorbed or transmitted by the matter. UV–Visible spectroscopy, infrared spectroscopy are the examples of these types of techniques. On the other hand sometimes not only the absorbed energy but also the energy that is released from the matter is also measured. Fluorescence spectroscopy is an example of this type of technique.

In this study mainly two types of spectroscopic techniques were used in the analysis of samples. These are infrared and fluorescence spectroscopic techniques. In infrared region near infrared and middle infrared regions were examined using two different spectrometer called as near infrared (NIR) spectrometer and Fourier transform infrared (FTIR) spectrometer. And lastly fluorescence spectrofluorometer was used in two different measurement modes: excitation – emission fluorescence and synchronous fluorescence. The theory of each technique will be not given in detail since there are a lot of textbooks that explain these techniques (Skoog, et al. 1998, Ingo, et al. 1988, Lakowicz 1999, Valuer 2001, Stuart 2004, Burns, et al. 2001, Smith 2000).

2.2. Infrared Spectroscopy

Infrared spectroscopy deals with the absorption/transmission of light from any substance at vibrational or rotational levels. Infrared region (0.78 – 1000 μm) exists in electromagnetic spectrum divided into three different regions named as: near, middle, and far infrared region. These regions are classified according to the nature of the process in vibrational or rotational levels (Skoog, et al. 1998).

2.2.1. Near Infrared (NIR) Spectroscopy

2.2.1.1. Principles

Near infrared region (13,000–4000 cm^{-1}) is the higher energy section of the infrared region in electromagnetic spectrum. The absorptions observed in near infrared region are overtones and combinations of the fundamental stretching bands which occur in the 3000–1700 cm^{-1}. The bands are usually due to the CH, OH, NH stretching. Overtone bands are analogous and multiples of fundamental absorption frequency

(Stuart 2004). The energy levels of overtones are shown in Figure 2.2. Combination bands arise when two fundamental bands absorbing at ν_1 and ν_2 absorb energy simultaneously. The resulting band will appear at $(\nu_1 + \nu_2)$ wavenumbers.

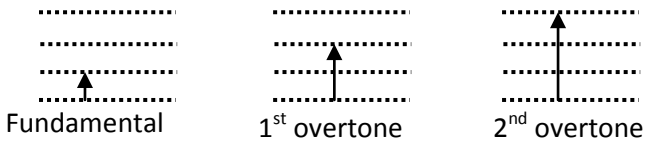


Figure 2.2. Energy levels for fundamental and overtone bands.

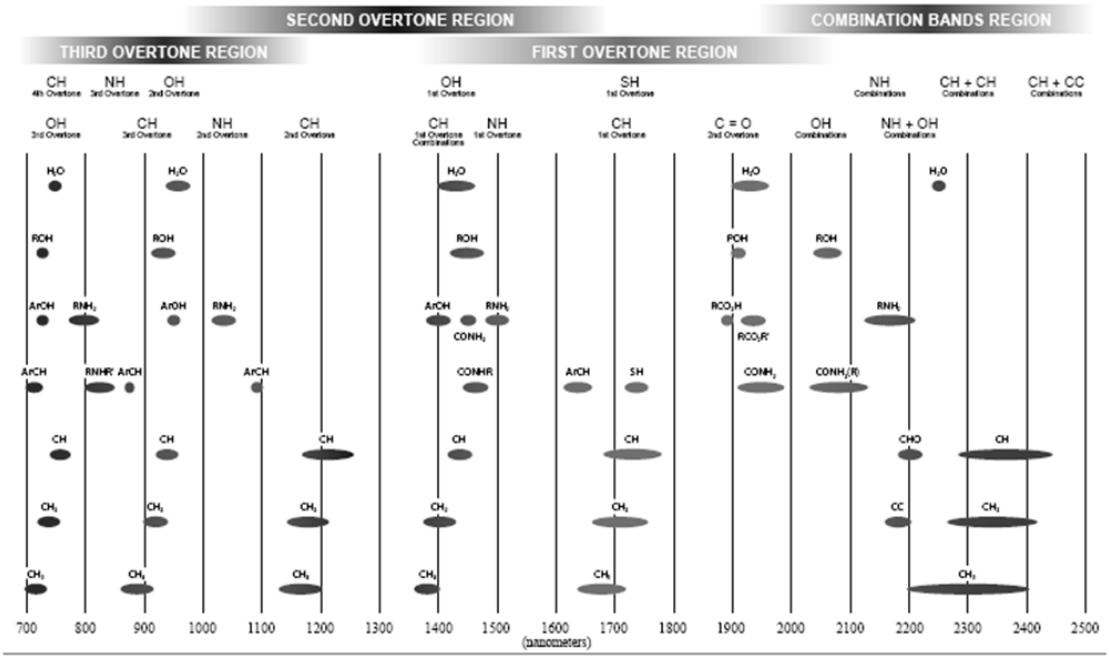


Figure 2.3. Near infrared absorption bands
(Source: Analytical Spectral Devices Inc. 2005)

The resulting bands in the near infrared are usually shows weak responses and the intensity of the response decreases one overtone to the next. The resulting spectrum obtained from measurement generally plotted against wavenumber to absorbance and each absorption bands refer the functional group that contains CH, OH, NH, and SH (Figure 2.3). The observed bands in near infrared region are generally overlapped (Figure 2.4), therefore using this region in qualitative analysis difficult compare to the middle infrared region. Chemometric methods will be needed to identify and

characterize the features of spectra. Fundamental of CH and OH stretching bands in middle infrared region are highlighted and their 1st, 2nd overtones are illustrated in near infrared region shown in different intensities.

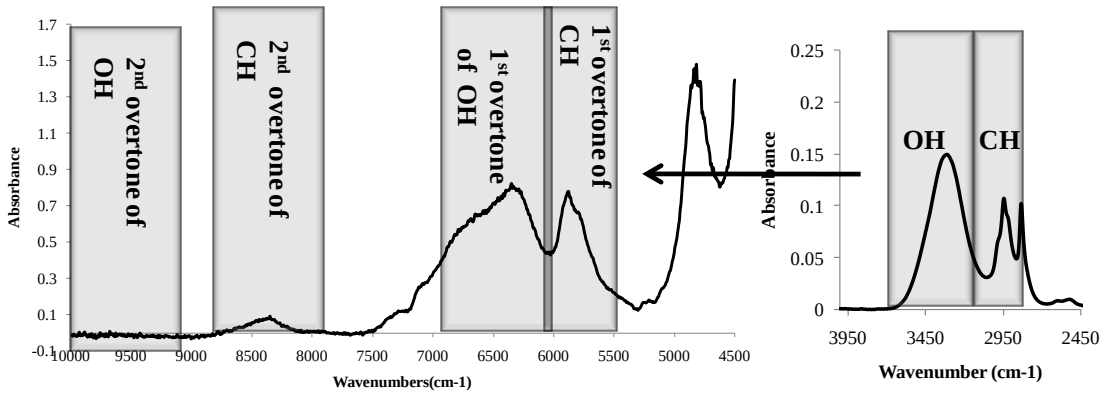


Figure 2.4. Infrared spectra of methanol in near and middle infrared regions.

2.2.1.2. Instrumental

The instrumentation in near infrared region is generally similar with UV-Visible absorption spectroscopy. Tungsten-halogen lamps are usually used as a source. Cells are used as sample holders are generally quartz or fused silica cells that are transparent up to 3000 nm. Detectors are generally lead sulfide photoconductors (Skoog, et al. 1998).

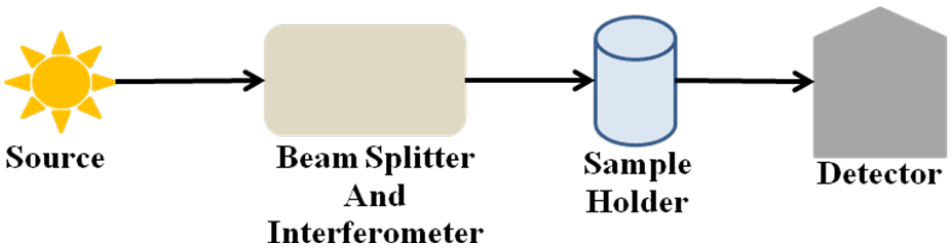


Figure 2.5. Optical diagram of typical near infrared spectroscopy instrument.

2.2.2. Fourier Transform Infrared (FTIR) Spectroscopy

2.2.2.1. Principles

The region starts from 4000 cm^{-1} and ends at 400 cm^{-1} in the electromagnetic spectrum assigns the middle infrared region. Infrared radiation is not sufficient to cause the transitions between the electronic states. The transitions occur only between the vibrational or rotational states. Vibrations include either a change in bond length (stretching) or bond angle (bending) (Figure 2.6). Some bands can stretch in-phase (symmetric stretching) or out-of-phase (asymmetric stretching). Bending also has different contributions in the infrared spectrum. There four types of bending vibrations named as: rocking, scissoring, twisting, and wagging.

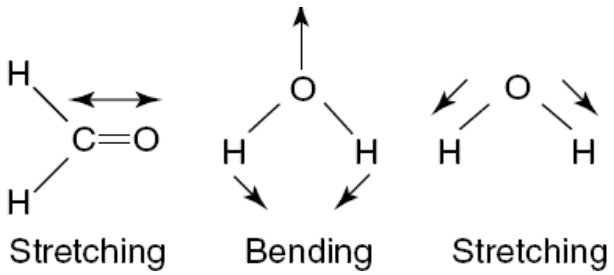


Figure 2.6. Stretching and bending vibrations in middle infrared region (Source: Stuart 2004).

Molecular bonds vibrate at various frequencies depending on the element and type of the bonds. According to the quantum mechanics these frequencies correspond to the ground state (lowest frequency) to the several excited states (higher frequencies).

2.2.3. Instrumental

Fourier transform infrared spectrometer is the preferred instrument in middle infrared region. It was developed to overcome the limitations of dispersive instruments. The main difficulty using dispersive instruments is the slow scanning process. In order to achieve the simultaneous measurement process, an interferometer is developed. Most interferometers employ a beamsplitter which takes the incoming infrared beam and

divides it into two optical beams. The result of these two beams is “interfering” with each other. The resulting signal is called an interferogram which has the unique property that every data point (a function of the moving mirror position) which makes up the signal has information about every infrared frequency which comes from the source. Because the analyst requires a frequency spectrum (a plot of the intensity at each individual frequency) in order to make identification, the measured interferogram signal cannot be interpreted directly. A means of “decoding” the individual frequencies is required. This can be accomplished via a well-known mathematical technique called the Fourier transformation (Figure 2.7). This transformation is performed by the computer which then presents the user with the desired spectral information for analysis (Thermo Nicolet Co. 2001).

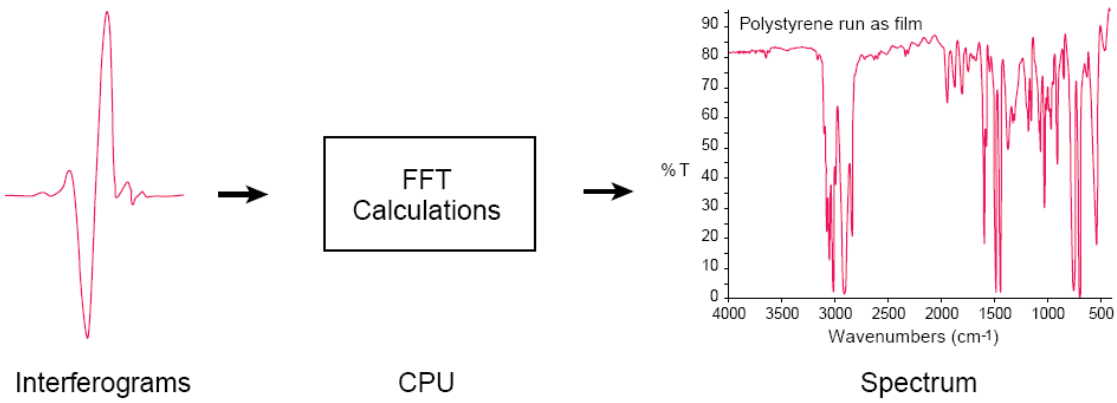


Figure 2.7. Schematic representation of Fourier transforms
(Source: Thermo Nicolet Co. 2001).

In infrared instruments Nernst glower, globar, tungsten filament, mercury arc or CO_2 laser are used as a source. Due to the heat property of sources, the detectors should be resistant to the heat. Thermocouples, bolometer, photoconducting tubes or pyroelectrics are generally used detectors in infrared spectrometers and also the mostly used one as an interferometer is the Michelson interferometer. Figure 2.8 shows the optical diagram of an infrared instrument.

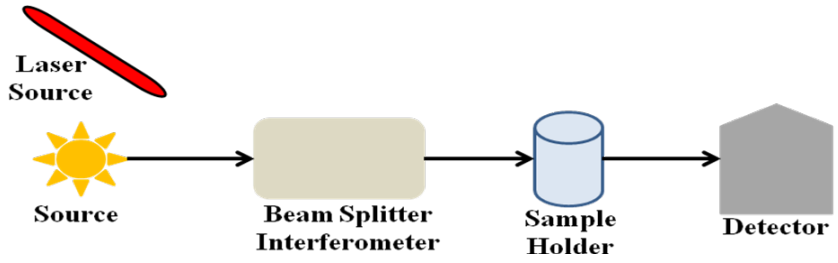


Figure 2.8. Optical diagram of Fourier transform infrared spectroscopy.

In infrared spectroscopy not only transmittance but also reflection can be measured by the help of the reflection accessories. Attenuated total reflectance (ATR) is the mostly used reflectance accessory in the measurements. In reflection techniques the infrared beam is bounced off the sample instead of passing through the sample (Smith 2000). In attenuated total reflectance, at the heart of the accessory is a crystal of infrared transparent material of high refractive index. Generally zinc selenide, thallium iodide/ thallium bromide, (KRS5) and germanium are used as crystal. Figure 2.9 shows the schematic diagram of an attenuated total reflectance accessory. As it is seen from the diagram, an evanescent wave is attenuated by the sample's absorbance; therefore this technique is called as attenuated total reflectance accessory (ATR).

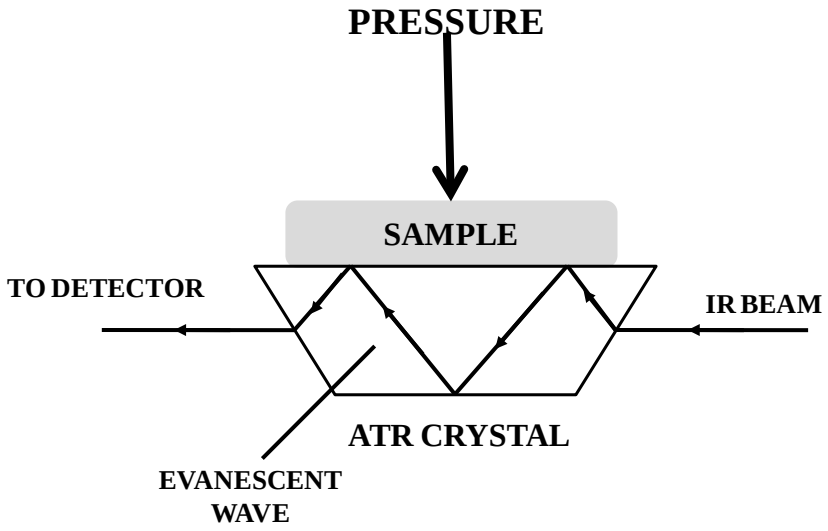


Figure 2.9. A schematic diagram of an attenuated total reflection accessory.

2.3. Fluorescence Spectroscopy

Molecular luminescence spectrometry deals with emission of light from any substance and occurs in electronically excited states (Lakowicz 1999). It is initially divided into two different categories named as fluorescence and phosphorescence. Chemiluminescence is also categorized in the luminescence spectrometry as a third type (Skoog, et al. 1998). These three types of luminescence are categorized according to the nature of the excited state. Fluorescence and phosphorescence are alike in which the excitation is brought by absorption of photons. The third type of luminescence; chemiluminescence is based on the emission spectrum of an excited species that are formed in a chemical reaction (Skoog, et al. 1998). In this thesis, only fluorescence will be considered.

2.3.1. Principles

Fluorescence and phosphorescence are different in terms of electronic energy transitions. In a singlet state all electron spins are paired, when one of a pair electrons are excited to the high energy level, either a singlet or a triplet state is formed. In fluorescence, the electron in the excited orbital is paired to the second electron in the ground state orbital. The electron spin is rapidly returned to the ground state and occurs in nanoseconds. Phosphorescence is the emission of light from the triplet excited states which has same spin orientation as the ground state. In this configuration, transition to the ground state are said to be forbidden, therefore the emission rates are slowly. Figure 2.10 shows the singlet/triplet excited states.

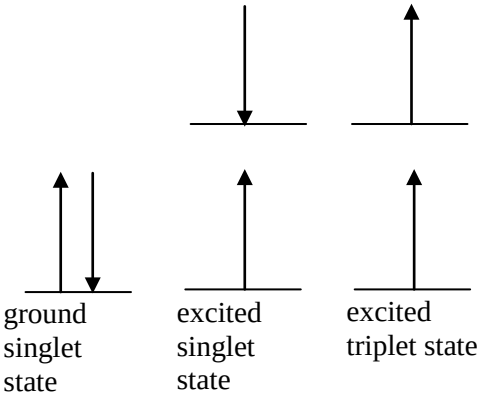


Figure 2.10. Singlet/Triplet excited states

The process which occurs between the absorption and emission of light are usually illustrated by a Jabloński diagram (Figure 2.11). Jabloński diagram typically explains the fluorescence and phosphorescence in terms of energy level. Three non-radiational processes are also explained here. These are internal conversion (IC), intersystem crossing (ISC), and vibrational relaxation. Internal conversion is the transition between energy states of the same spin state. The transition between the different spin states called as intersystem crossing. The last non-radiational process vibrational relaxation occurs in a molecule which is in excited vibrational and rotational sates. Molecules dissipate their excess vibrational energy and relax to the ground vibrational level in a given electronic state.

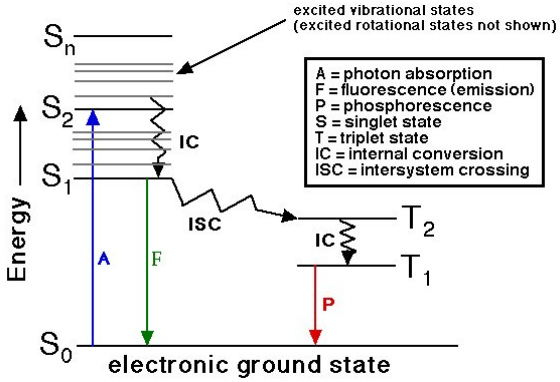


Figure 2.11. Jabloński diagram.

In many molecules which have aromatic ring fluorescence is observed, since fluorescence usually involves π-π^* transitions. Fluorescence efficiency is also not only depended on molecular structure but also the temperature, pH, solvent of the sample are important. Fluorescence spectral data are generally presented by emission spectra of fluorophores. The main advantage of studying with fluorescence spectrometry is the high sensitivity among the other spectroscopic techniques. Even in low concentrations there is linearity between the intensity and the concentrations of fluorophores. However at high concentrations, there is a deviation from the linearity. Since the high amount of fluorophores itself causes inner filter effect and/or by quenching (Ingle, et al. 1988). Quenching is the most seen process causes the decreasing in the fluorescence intensity. There are different phenomena which can be seen together in quenching and inner filter effect. *Dynamic quenching* requires contact between the excited fluorophores and the quencher. In *static quenching*, the fluorophores and the quencher form a stable compound at ground state. If dipole-dipole coupling occurs between the fluorophores and quencher, the *long-range quenching* is formed. Lastly, in *inner filter effect* the fluorophores can absorb the emitted light itself (Ingle, et al. 1988) (Figure 2.12).

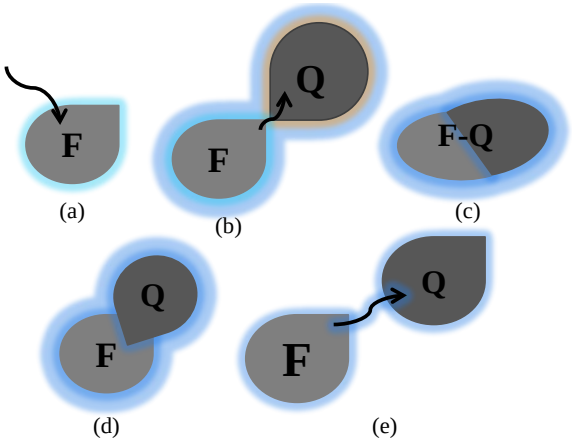


Figure 2.12. a) Excited fluorophore, four ways of a decrease in fluorescence intensity b) dynamic quenching, c) static quenching, d) long-range quenching, e) inner filter effect (Source: Rinnan 2004).

2.3.2. Instrumental

There are several types of fluorescence instruments. Figure 2.13 shows the general representation of fluorescence instruments. Fluorescence instruments are named as according to the wavelength selectors. If both are filter, the instrument is called as fluorometer. If monochromators are used as wavelength selectors, these types of instruments are called as spectrofluorometer (Skoog, et al. 1998). Generally in fluorometer, low-pressure mercury vapor lamp equipped with fused silica window is used whereas xenon-arc lamps are generally used as a continuum source in spectrofluorometer. Generally the fluorescence signal is a low intensity, therefore photomultiplier tubes are the most common transducer in fluorescence instruments. Diode-array and charge transfer detectors have been also proposed for spectrofluorometers (Skoog, et al. 1998).

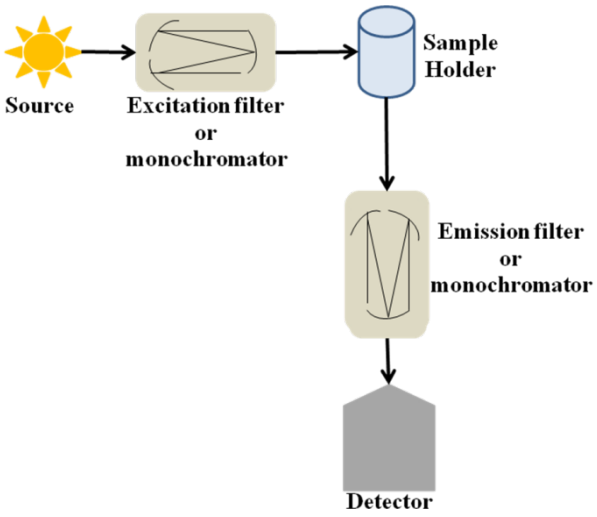


Figure 2.13. Optical diagram of typical fluorescence instrument.

In the fluorescence instruments, the beam passes through the samples. Two monochromators or filters allow the scanning of excitation spectra at a fixed emission wavelength or emission spectra at a fixed excitation wavelength or synchronous (both wavelengths scanned with a fixed wavelength offset between two monochromators or filters). At the end of the measurements fluorescence spectrum is obtained. If an emission spectrum is recorded at a fixed excitation wavelength region, three-dimension excitation – emission fluorescence spectrum will be obtained (Figure 2.14).

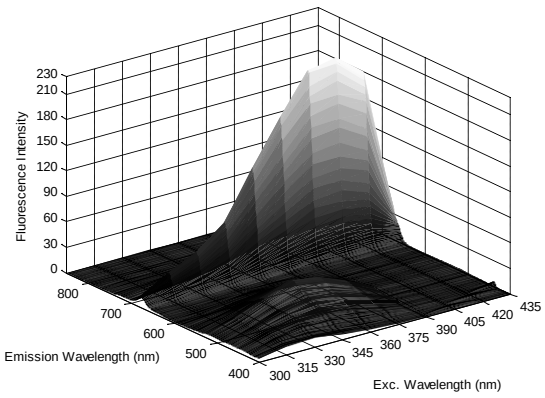


Figure 2.14. A typical excitation–emission fluorescence spectrum.

In excitation–emission fluorescence spectrum, generally includes Raman and Rayleigh scatters (Lakowicz 1988, Skoog, et al. 1998). Rayleigh scattering generally

occurs from the solute and sometimes fluorophores themselves. The dimensions of scatterer are much smaller than the incident beam wavelength. In the fluorescence measurements, the absorbing and emission wavelength are same and due the multiple of the absorbing wavelength scatter lines sometimes will occur (Figure 2.15). Therefore Rayleigh scattering is denoted by number as 1st, 2nd,…. While Rayleigh scattering is elastic, Raman scattering is inelastic. Raman scattering generally caused by the vibrational and rotational levels of molecules and sometimes can occur with a relatively small frequency shift that varies with the scattering angle (Ingle, et al. 1988).

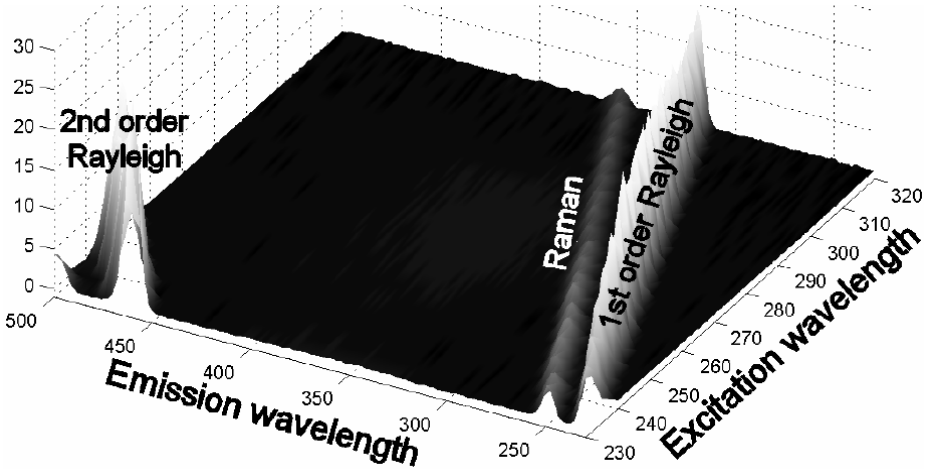


Figure 2.15. An EEF of de-ionized water showing three diagonal peaks: two Rayleigh (1st and 2nd order) and one Raman peaks (Source: Rinnan 2004).

CHAPTER 3

CLASSIFICATION METHODS

Chemometrics has a broad definition since it contains the application of mathematical and statistical techniques to the chemical data. Calibration and classification methods have a large application area in chemometrics. The development in computer technology causes to obtain large amounts of data after chemical analysis which is done using computerized instrumental methods. In order to evaluate results and extract the necessary information from the huge amount of data, mathematical and statistical techniques were begun to use from other disciplines. Chemometrics mainly contains the signal processing, time-series analysis, optimization and experimental designs, pattern recognition, classification and calibration modeling methods. Each of these techniques can be found in chemometrics textbooks (Otto 1999, Brereton 2002)

In this chapter not only the classification and clustering techniques will be investigated but also preprocessing techniques in data analysis will be given. Since the raw spectroscopic data sometimes contain irrelevant sources such as random or systematic errors.

3.1. Preprocessing Techniques

Preprocessing is a very important part of chemometrics and it is any mathematical manipulation of the data prior to the primary data analysis (Beebe, et al. 1998). Choosing the most appropriate preprocessing techniques will affect the results either positively or negatively. Preprocessing techniques can be investigated in to different types depending on whether they operate on samples or variables. Normalizing, weighting, smoothing and baseline corrections are the preprocessing techniques that operate on samples. On the other side preprocessing of variables includes the mean centering and variable weighting. Normalization of a sample vector is observed by dividing each variable by a constant. At the end of the process, normalization puts all the samples on the same scale. For example removing the variable injection volume in chromatography or reducing the pathlength variation in

near–infrared reflectance spectrum of a sample. Weighting is similar to the normalization but differs in the defining of the criteria. It is accomplished by multiplying each element in a sample vector. Sample weighting gives some samples more influence on the analysis than the others. For instance weight of zero eliminates a sample. In instrumentation, the instrumental signal contains the true signal together with random noise. The amount and type of noise depends on the experimentation. Smoothing procedures are used to reduce the noise and maximize signal-to-noise ratio. The last preprocessing technique on samples is the baseline correction. Baseline correction is applied to the low-frequency variations of a sample vector, in order to reduce the systematic variations (Beebe, et al. 1998).

Mean centering of a variable is obtained by subtracting the mean of that variable vector from all of its elements. Mean centering is the most used one in data analysis, since it generally helps to improve the results. The second one variable weighting emphasizes some variables over others, and increases their influence on the primary analysis. Variance scaling, autoscaling and the variable selection are the types of the variable weighting (Beebe, et al. 1998).

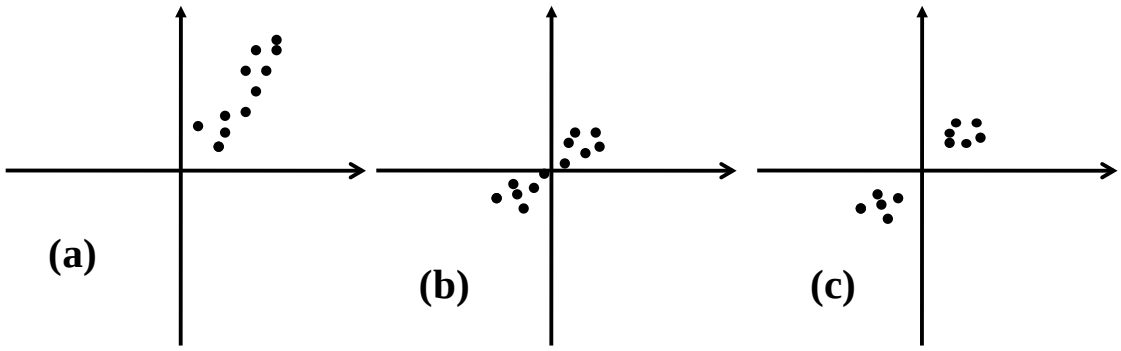


Figure 3.1. Schematic representation of some preprocessing techniques (a) original data, (b) centered data, (c) autoscaled data (Source: Otto 1999).

Figure 3.1 shows the demonstration of mostly used preprocessing that operates on variables. It should be reminded here; the spectroscopic data also contains variables that are the wavelengths or wavenumbers with their corresponding instrumental responses and also the chemical profiles of the samples such as their metal contents refers the variables of the data. The better ways to choose the most appropriate

preprocessing technique, both non preprocessed and preprocessed data are tried and the results are compared.

3.2. Classification Techniques

Human skills are enough to identify or recognize the differences between the samples depending on the shapes. But it is generally acceptable for the sample set which includes several numbers of samples. In analytical chemistry the data matrices including a large number of rows and columns and human skills cannot enough to recognize the properties of the data. On account of this, the classification methods have been developed to present the chemical data into the pictures instead of matrices. Mathematical relationship is built by the help of the computer and the results are interpreted using plots in order to observe differences or similarities between the samples. The rule of the distinction is observed by the mathematical equations. The samples which have similarities are called as class membership. According to the class membership term, a class is defined as a collection of samples are defined as being similar. Classification methods can be categorized mainly two groups named as unsupervised and supervised (Otto 1999, Beebe, et al. 1998, Brereton 2002).

Figure 3.2 shows the diagram decision tree for the classification methods. Choosing the right technique depends on the prior knowledge of the samples. If there is no information about unknown generally unsupervised classification methods are preferred. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) are the unsupervised techniques. If the aim is to develop a predictive model, a training set with known class membership is required to construct the model and the techniques are called as supervised classification methods (Beebe, et al. 1998). Soft independent modeling of class analogy (SIMCA) and K-nearest neighbor (KNN) are typically used as supervised techniques.

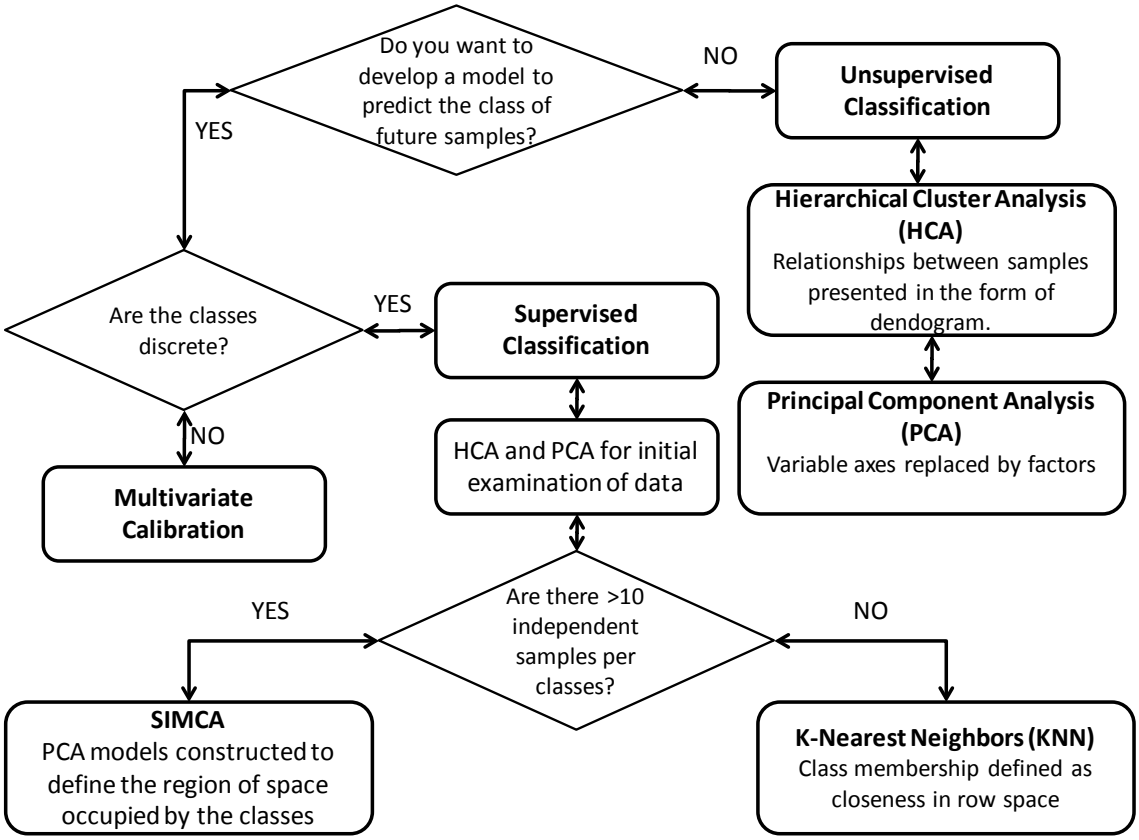


Figure 3.2. Classification Decision Tree
(Source: Beebe, et al. 1998).

3.2.1. Unsupervised Classification Techniques

In the examination of unsupervised techniques, the evaluation of whether the clustering is existed in a data set without any prior class membership information in the calculations is investigated. As it understood from the theory behind the unsupervised techniques, the presence or absence of the clustering in data is displayed. Mainly two different techniques are mostly used in unsupervised methods. Hierarchical cluster analysis (HCA) based on cluster analysis which is based on the aggregation of samples according to their similarities. Principal component analysis (PCA) is a factor analysis method that is aimed projecting the original data from high dimensional space on to a line, a plane, or a 3D-coordinate system.

3.2.1.1. Hierarchical Cluster Analysis (HCA)

HCA examines the interpoint distances between all the samples and represents that information in the form of a two-dimensional plot called a dendrogram (Figure 3.3). To generate the dendrogram, HCA forms clusters of samples based on their similarities in space. Different approaches are used to measure distances between the clusters (Beebe, et al. 1998). Firstly the distances are calculated and linkage methods are used to form the dendrogram.

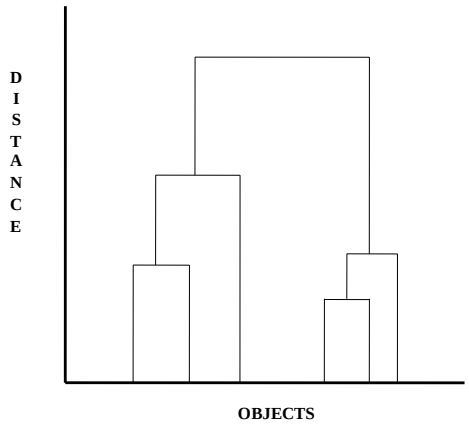


Figure 3.3. Schematic representation of a dendrogram for 6 objects.

A general distance measure is the Minkowski distance or L_p-metric.

$$d_{ij} = \left[\sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^{p} \right]^{1/p} \tag{3.1}$$

where K is the number of variables and *i, j* indices the object *i* and *j*. The mostly used distance is the Euclidean distance for which *p*=2. If the distances refer to travel around corner, this distance is called as Manhattan or city-block distance. In this case *p* is equal to the 1. These distances are generally based on metrics that are used in the process. Therefore the scaling is mostly unavoidable in these distance measurements. The variables which have different scales need to use another distance measurement

technique. This measure is calculated by use of the following formula and it is called as Mahalanobis distance.

$$D_{ij}^2 = \left(x_i - x_j\right)^T C^{-1}\left(x_i - x_j\right) \tag{3.2}$$

where C is the covariance matrix and x_i, x_j are the column vectors for objects *i* and *j*. After calculation of the distances between the objects, a distance data matrix is observed. Reducing of the distance matrix is performed by aggregation of objects.

Table 3.1. The summary of linkage methods that are used HCA (Source: Otto 2002).

Linkage Method	**Mathematical Representation**	**Principle**
Weighted Average	$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2}$	The sizes of the clusters and their weights are assumed to be equal.
Single	$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2} - \frac{\left\|d_{Ai} - d_{Bi}\right\|}{2}$	The shortest distance between the opposite clusters is calculated.
Complete	$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2} + \frac{\left\|d_{Ai} - d_{Bi}\right\|}{2}$	The largest distance between the opposite clusters is calculated
Unweighted average	$d_{ki} = \frac{n_A}{n} d_{Ai} + \frac{n_B}{n} d_{Bi}$ n_A, n_B – numbers of objects $n=n_A+n_B$	The number of objects in a cluster in a cluster is used for weighting the cluster distances. Outliers can be detected.
Centroid	$d_{ki} = \frac{n_A}{n} d_{Ai} + \frac{n_B}{n} d_{Bi} - \frac{n_A n_B}{n^2} d_{AB}$	The centroid calculated as the average of a cluster and it is applied as the basis of the aggregation.
Median	$d_{ki} = \frac{d_{Ai}}{2} + \frac{d_{Bi}}{2} - \frac{d_{AB}}{4}$	It is used for the determination of the centroid
Ward's Method	$d_{ki} = \frac{n_A + n_i}{n + n_i} d_{Ai} + \frac{n_B + n_i}{n + n_i} d_{Bi} - \frac{n_i}{n + n_i} d_{AB}$	The clusters are aggregated such that a minimum increase in the within group error sum of squares results.

The mathematical process in the aggregation of the objects is called as linkage method. In general the distance to a new object or cluster *k* is computed by calculating the average distances from the objects A and B to object *i*. Table 3.1 summarizes the principles and mathematical expression linkage methods that are used to form the dendrogram.

In a summary, dendrogram shows the closeness of samples in row space in the form of two-dimensional graph. The samples are plotted against the distances and the similarities/differences between samples without imposing prior information regarding the class membership. Samples existed at large distances can indicate as outliers, on the other side the samples with small distances have similar properties.

3.2.1.2. Principal Component Analysis (PCA)

Principal component analysis is the oldest and the mostly used algorithms in the unsupervised classification techniques. It is used both in multivariate calibration and classification techniques (Lavine 2000). PCA is a mathematical manipulation of a data matrix where the goal is to represent the variation present in many variables using a small number of "factors" or "principal components (PC)" (Beebe, et al. 1998). Simply, PCA decomposes the data matrix into smaller matrices named as scores and loadings. In order to predefine the axes using factors instead of original variables. These new axes are called as principal components. Equation (3.3 shows the mathematical formula of principal component analysis.

$$X_{nxp} = T_{nxd} L_{dxp}^{T} \tag{3.3}$$

where $\boldsymbol{X}$ is the original data matrix with *n* rows and *p* columns; $\boldsymbol{T}$ is the scores matrix with *n* rows and *d* columns (number of principal components); $\boldsymbol{L}$ is the loading matrix with *d* columns and *p* rows; T indicates the transpose of the matrix (Otto 2002). When two variable systems are considered, the relationships between the samples are defined by the help of the distances between the samples which have similarities or differences. PCA describes the spread or variation of the distances in a few axes or dimensions, since in analytical chemistry more than two varieties are studied. As it is seen from the

Figure 3.4., the original variables are not enough to explain the variation in the data set. However the first principal component describes more than the original variables.

The principal components calculated in the PCA have some properties. The following entries are explained these properties (Beebe, et al. 1998).

- The first principal component explains the maximum amount of variation possible in the dataset in one direction.
- Sample has coordinates in the original space; it has also coordinates with respect to the new principal components. The coordinates of the samples relative to the principal components are typically named as “scores”.
- Each PC is constructed from the combinations of the original measurement variables. The contribution of original measurement variable is proportional to the principal components. This contribution of each axis to the principal component is the cosine of the angle between the variables axes and the principle component axis. These cosine values are often called “loadings” and can range -1 to 1.
- Excluding nonsignificant principal components can be used filter the noise from the data set.
- The maximum number of PCs can be calculated is the smaller of the number of samples or variables.
- Principal components are orthogonal (perpendicular) to each other.

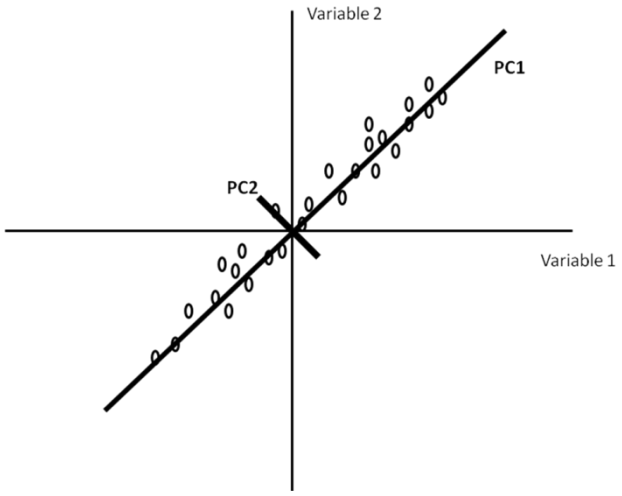


Figure 3.4. A row plot data in a two-measurement system, with the first two principal components.

PCA has some different mathematical operation in the computation. The simplest one is the NIPALS (nonlinear iterative partial least squares). More powerful methods are based on the matrix diagonalization such as singular value decomposition. The detailed information about NIPALS algorithm can be found in chemometrics textbook (Otto 2002). Here only singular value decomposition (SVD) based principal component analysis will be discussed.

SVD decomposes the data matrix, ***X***, into the matrices coded as ***U***, ***W***, and ***V***.

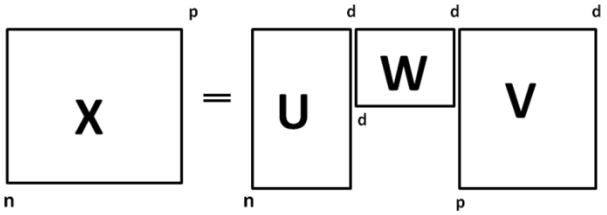


Figure 3.5. Schematic representation of decomposition in SVD-PCA

Equation (3.4) shows the mathematical representation of the SVD.

$$X = UWV^T \tag{3.4}$$

The matrix U contains the same column vectors as does the scores matrix T in Equation (3.3) but it is normalized to one. W is the diagonal matrix containing the square roots of the eigenvalues or singular values. The number of principal components after decomposition of the original data matrix is important for the success of the model. There are some methods to determine the number of PCs. Percentage explained variance, eigenvalues-one criterion, Scree-test and cross validation are the examples of heuristic and statistical criterion tests. The percent explained variance is the simplest heuristic criterion. The fraction of the explained (cumulative) variance s_e^2, is calculated from the ratio of the sum of *d* important eigenvalues and the sum of all *p* eigenvalues. The eigenvalues-one criterion is based on the average eigenvalue of autoscaled data is just one. In the Scree-test, the residual variance off when the appropriate number of PCs is obtained. The eigenvalues are often plotted against the PCs and leveling-off point is

used to determine the number of components. Lastly, in cross validation, every object of the X-matrix is removed from the data set once, a model is computed from the remaining data, and then the removed data are predicted by PCA model. The sum of the square of the roots of the residuals overall removed object is calculated. The number of significant PCs is determined by minimum residual error. Also for the large data sets, leave-one-out method can be used. The matrix V^T is identical to matrix L^T Equation (3.3). The matrices U and V are also denoted as left and right vector of singular values, respectively (Otto 2002).

3.2.2. Supervised Classification Techniques

In supervised classification techniques, the main question is as followed: use the learning or training objects to derive a classification rule which allows classifying new objects with unknown origin in one of three known classes, based on the values of the features of the new object (Massart 1997). Supervised classification techniques are generally based on:

Selection of training or learning set which contains the objects of known classification for which a certain number of variables are measured.

Feature selection, that is the selection of variables that are meaningful for the classification and elimination f those that have no discriminating.

Derivation of a classification rule using the training set and modeling techniques.

Validation of the classification rule, using an independent test set.

There are a lot of supervised classification methods which are linear learning machine (LLM), discriminant analysis (DA), k-nearest neighbor (KNN) and soft independent modelling of class analogy (SIMCA). Here only SIMCA will be discussed. The detailed information can be found in textbooks (Beebe, et al. 1998, Otto 1999, Brereton 2002, Massart 1997).

3.2.2.1. SIMCA

The SIMCA method, first presented by the S. Wold in the early 1970s and it is regarded as a form of soft modeling used in classification methods. The idea of soft

modeling comes from two overlapping classes and there is no problem with an object belonging to both (or neither) classes simultaneously. In most cases, an object belongs to discrete. This is the expected result; therefore these types of classification are named as hard modeling (Brereton 2002). Figure 3.6 illustrates the phenomena of soft modeling by two overlapping classes.

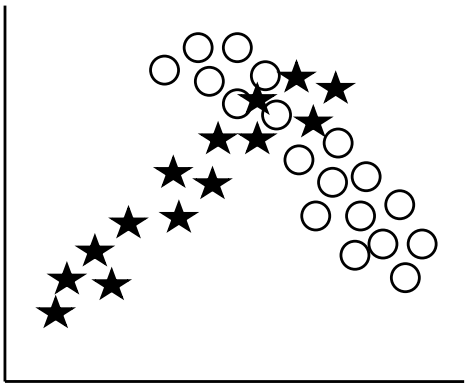


Figure 3.6. Illustration of two overlapping classes in concept of soft modeling (Source: Brereton 2002).

SIMCA uses the PCA model to determine the principal components or eigenvectors needed to builds a distinct confidence region around each class of the training class (Maesschalck, et al. 1999). If the numbers of observed eigenvectors coded as r^*, the value of r^* presents the dimensionality of the model shape. As it is seen from Figure 3.7, if r^* is equal to the 1, all data are considered to be modeled by a one-dimensional model, a line (Figure 3.7. a). For r^*=2 two-dimensional model, a plane (Figure 3.7.b) is observed. The residuals of the training class towards such a model are assumed to follow a normal distribution with a residual standard deviation:

$$s = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij}^2 / \left[\left(r - r^*\right)\left(n - r - r^*\right)\right]} \qquad (3.5)$$

The residuals from the model can be computed from the scores on the non-retained eigenvectors, i.e. the scores t_{ij} on the eigenvectors r^* + 1 to r (r = min {{n -1), m)). Then:

$$s = \sqrt{\sum_{i=1}^{n} \sum_{j=r^*+1}^{m} t_{ij}^2 / \left[(r - r^*)(n - r - r^*) \right]} \tag{3.6}$$

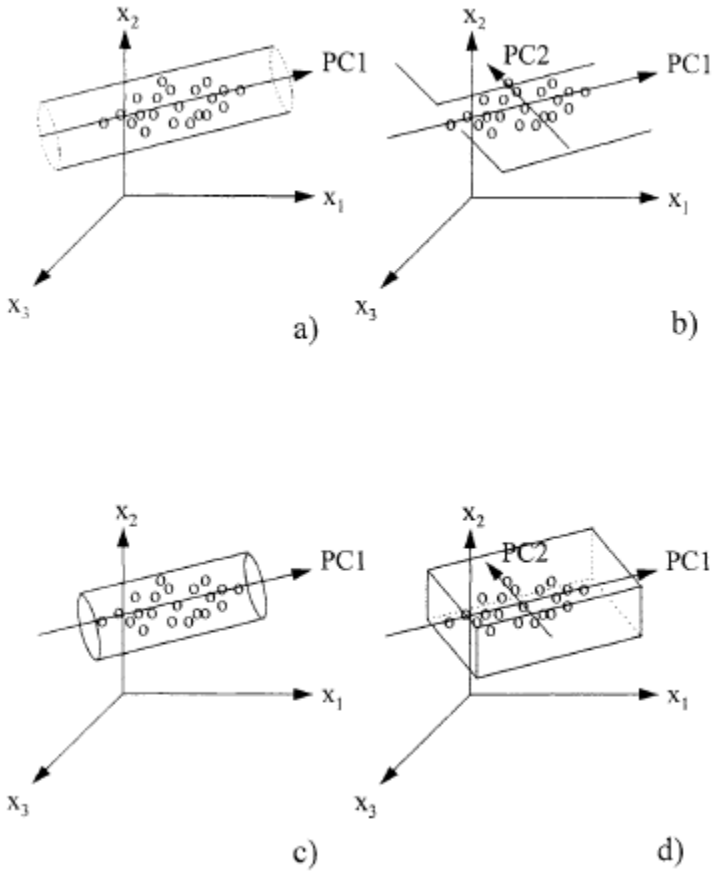


Figure 3.7. SIMCA: a) step 1 in a 1 PC model, b) step 1 in a 2 PC model, c) step 2 in a 1 PC model, d) step 2 in a 2 PC model (Source: Massart 1997)

If care is not taken about the way *s* is obtained, SIMCA has a tendency to exclude more objects from the training class than necessary. The *s*-value should be determined by cross-validation. Each object in the training set is then predicted, using the r*-dimensional PCA model obtained, for the other *(n*-1*)* training set objects. The (residual) scores obtained in this way for each object are used in Equation (3.4).

A confidence limit is obtained by defining a critical value of the (Euclidean) distance towards the model. This is given by:

$$s_{crit} = \sqrt{F_{crit}}\, s \tag{3.7}$$

F_{crit} is the tabulated one-sided value for (r - r*) and (r - r*) (n - r* - 1) degrees of freedom. The s_{crit} is used to determine the boundary (the cylinder) around the PC1 line in Figure 3.7.c and the planes around the PC1, PC2 plane in Figure 3.7.d. Objects with $s<s_{crit}$ belong to class K, otherwise they do not. To predict whether a new object x_{new} belongs to class K one verifies whether it falls within the cylinder (for a one-dimensional model), between the limiting planes (for a two-dimensional model, etc.). Suppose the following r* dimensional PC model was obtained.

$$X_K = T_K L_K^T + E_K \tag{3.8}$$

with X_K the centred X-matrix for class K, T_K the (un-normed) score matrix (nx$r$*), ($T_K = U_K \Lambda_K$, where U_K is the normed score matrix and Λ_K is the singular value matrix). L_K is the loading matrix *(mx$r$*) and* E_K. is the matrix of residuals *(nxm)*

For a new object x_{new} one first determines the scores using the equation below.

$$t_{new}^T = \left(x_{new} - \bar{x}_K\right)^T L_K \tag{3.9}$$

The Euclidean distance from the model is then obtained as:

$$s_{new} = \sqrt{\sum_{j=r^*+1}^{r} t_{new,j}^2 / \left(r - r^*\right)} \tag{3.10}$$

If $s_{new}<s_{crit.}$ then the new object belongs to class K, otherwise it does not.

A useful tool in the interpretation of SIMCA is the so-called Coomans plot. It is applied to the discrimination of two classes (Figure 3.8).

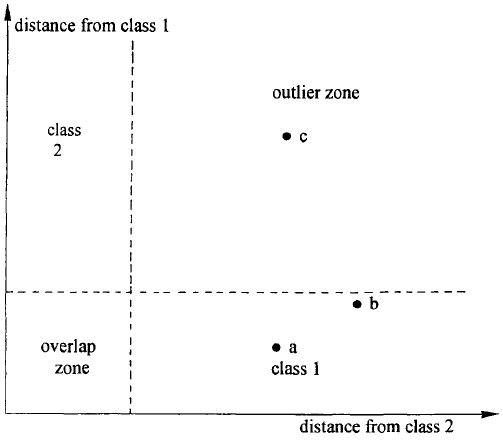


Figure 3.8. The Coomans plot.
(Source: Massart 1997)

The distance from the model for class 1 is plotted against that from model 2. On both axes, one indicates the critical distances. In this way, one defines four zones: class 1, class 2, overlap of class 1 and 2 and neither class 1 nor class 2. By plotting objects in this plot, their classification is immediately clear. It is also easy to visualize how certain a classification is. In Figure 3.8, object a is very clearly within class 1, object b is on the border of that class but is not close to class 2 and object c clearly belongs to neither class.

All the equations and the definitions of each step were taken from Handbook of Chemometrics and Qualimetrics Part B textbook edited by Massart.

CHAPTER 4

GENETIC ALGORITHMS AND GENETIC ALGORITHM BASED CLASSIFICATION METHODS

4.1. Genetic Algorithms (GAs)

Genetic algorithms are the evolution and optimization methods which were firstly represented by Holland in the early 1960s (Lucasius 1991). From the middle of 1980s onwards genetic algorithms have received increasing usage in problem solving. And now they have been providing a powerful general purpose search strategy in the area of image processing, pattern recognition, modeling and system identification, adaptive filtering etc. (Fontain 1992, Cong and Li 1994, Wienke, et al. 1993, Hibbert 1993, Lucasius and Kateman 1991). In terms of calibration and classification, there have been several applications of GAs to wavelength selection (Lucasius, et al. 1994, Lucasius and Kateman 1992, Paradkar and Williams 1997, Ozdemir, et al. 1998a, Ozdemir, et al. 1998b, Ozdemir and Williams 1999). They were developed to understand the adaptive process of natural systems and to design artificial systems software that retains the robustness of natural systems. They are basically based on the Darwinian's classical rules about natural evolution which is "struggle for life (competition rule) and survival of the fittest (selection rule)" (Lucasius 1993).

As it is mentioned above, GAs use the biological systems that exist in the nature. Therefore the definitions of terms that will be used in the explanation of GAs should be given. The ability of a living simple organism is to produce an enzyme which contains a code that is stored in its chromosome. This chromosome contains a string of deoxyribonucleic acids (DNA) and it can contain four different nucleic acids (A, C, G, or T). These four nucleic acids are the genetic alphabet of the chromosome. A region or substring of the chromosome that contains the code of the enzyme is called as gene. In GAs, the solution of a given problem is called a "gene" and the vector of genes is called a "chromosome". A set of chromosomes collected from the generation is used to describe a "population". The population is the genotype of this organism or assumed solution and the ability of the set of selected features to solve the problem. The criterion

which solves the problem is also known as its “fitness”. In the production of offspring, two parents are selected and the chromosome of the offspring is constructed by combining sections of the parents’ chromosomes. There is a small probability that a mutation can occur in the offspring before its genotype is established. This genotype produces a phenotype for the offspring. Darwin’s ‘survival of the fittest’ generally determines which parents are used for mating and whether or not the offspring is viable enough to become a parent (Luke 2003).

The operation principle of GAs can be summarized as evolutionary process in nature. They start with initial population of potential solutions to prescribed problems. Darwinian principle is used to generate new, modified populations. This process is repeated until converging on a near optimal solution for fitness criterion. Figure 4.1 summarized the operation steps of GAs. They include five basic steps followed by initialization of the population, evaluation and ranking the population, selection of parents for breeding, crossover and mutation, replacing the parents with their offspring.

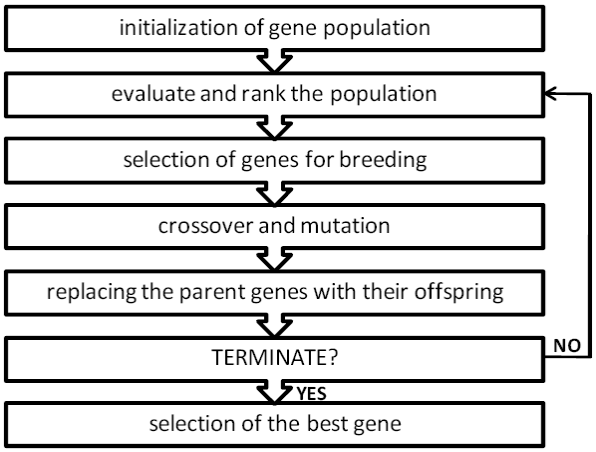


Figure 4.1. Block diagram of basic genetic algorithms
(Source: Karaman 2008)

GAs can handle large number of variables and they can use for discrete and continuous variables, and they can also be accepted as versatile nature evolution processes. However they have a few constraints. They are based on randomness and for each run of algorithm different results will be obtained.

4.2. Genetic Algorithm Based Classification Methods

Natural evolution based methods allow themselves to hybridization since they are flexible and also their strengths are to complement and assist other methods. As it is mentioned before GAs require only a specific fitness function that depends on a given problem. Thus a genetic algorithm can operate with other optimization methods in a number of ways. It can be used to optimize another method, provide a method of generating search conditions, or be totally integrated with other methods (Hibbert 2003).

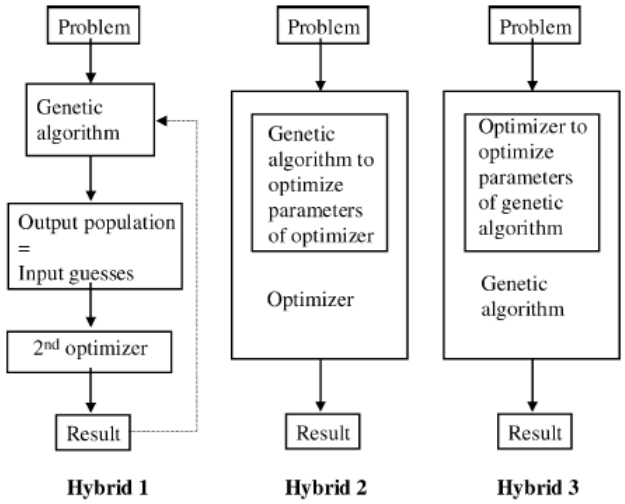


Figure 4.2. Three configurations of genetic algorithms hybrids.
(Source: Hibbert 2003)

Generally there are three different configurations for the genetic algorithms as summarized in Figure 4.2. Genetic algorithms can be used in conjunction with another method. The level of interaction can be changeable configuration to configuration. If the problem is very large, genetic algorithms can be used before or after another program without any interaction between them. Genetic algorithm can be a precursor to a second optimizer with or without interaction (Hybrid 1). There are some studies on this type of hybridization (Handschuh, et al. 1998, Hibbert 1993, Kemsle 2001, Liu 1999, Vivo-Truyols 2001a, Vivo-Truyals 2001b, de Wejver 1994). The most prevalent one in the hybridization is the genetic algorithm that provides input to a second optimization

method. The results may or may not be cycled back into the genetic algorithm in an iterative manner. They have great ability to search a parameter space makes it ideal as a formulator of input data for a more directed optimizer (Hibbert 2003). This type of hybrid is shown as Hybrid 2 in Figure 4.2. Raymer, et al. and Yoshida, et al. in 1997 work in this type of algorithm. Last type of hybridization (Hybrid 3) genetic algorithm with an optimizer determines an aspect of the genetic algorithm. Hanagandi and Nikolaou in 1998 present a study about Hybrid 3 type study.

4.2.1. Distance Based Genetic Algorithm Principal Component Analysis (GAPCAD)

In this study, the main goal is to perform a discrimination using spectral data. To observe efficient discrimination of the samples, partition on spectral data is commonly used in the process. Since spectral data contain more than thousands of response signals with corresponding wavelengths, using whole spectrum can cause overfitting in the training sample set (Reynes, et al. 2006). Instead of optimizing full spectra, local optimization can reduce the classification errors of the samples. The searching capability of genetic algorithms (GAs) seems to more adaptable to the classification and clustering techniques among the other nature inspired algorithms (Leardi 2003). Since they are randomized global search and optimization techniques based on the principles of natural evolution and selection (Lucasius and Kateman 1991, 1993; Hibbert 1993). The samples' spectral data matrix is arranged into two different sample set named as training and test sets. GAs is applied to the training set which includes whole spectra of each training sample. Initial population includes chromosomes and the best of these chromosomes are specified, tested and ranked according to their fitness criterion. The chromosomes having the best fitness have the possibility of surviving and producing offspring chromosomes. The principal component analysis (PCA) is firstly performed to initial population; score and loading matrices of data matrix are obtained. After that, score matrix is used to determine the distance values between groups in order to observe the classes. Those distances are designed as fitness criterion in the algorithm. The observed classes are tested by independent sample set. The whole process is called as distance based genetic algorithm principal component analysis classification analysis (GAPCAD). This method will be discussed in detail.

4.2.1.1. Population initialization

Spectral data is used directly in the classification process. Neither binary coding nor string representation is used in the presentation of genes. In this process, each wavelength with its corresponding instrumental response value refers the genes. The genes in each chromosome which are the solution of a given problem are initialized randomly from the data matrix set. Random selection minimizes the bias whereas maximizes the possibilities of recombination and the diversity of genes. Genetic algorithm based principal component analysis is designed to select initial genes in somewhat biased random fashion in order to start with genes better suited to the problem than those which would be randomly selected. This becomes more important when the search space contains large regions, which do not contain any useful information. The size of the initial population is predefined by the user.

4.2.1.2. Evaluation of Selected Genes and Ranking

The evaluation of the genes is done with a fitness function that measures the success of the population based on their ability to solve the given problem. Once a chromosome is selected, it is used to form the reduced spectral data matrix at the points determined by the elements in that chromosome. Then fitness calculation is done in a two-step process: Singular value decomposition based principal component analysis and calculation of Manhattan distance between the sample groups. The reduced data matrix is used in singular value decomposition PCA analysis where score and loading matrix of all principal components (PC's) are determined as it is shown in Equation (4.1).

$$X_{nxp} = U_{nxn} W_{nxp} V^{T}_{pxp} = T_{nxp} V^{T}_{pxp} \tag{4.1}$$

where n is the number of samples, p is the number of spectra, X represents the spectral data matrix, U represents the unweighted score matrix of principal components, whereas W is the diagonal matrix containing the square roots of eigenvalues, V is the loading matrix of principal components and T is the weighted score matrix (Maesschalck, et al. 1999). The summation of cumulative value of these first two eigenvalues generally explains most of the variability in the data. The total cumulative value of a few

eigenvalues is expected as near as 100%. According to the GAPCA algorithms, the first two eigenvalues are forced to be very significant in explaining the variability of the system. These first two principal components are used to calculate the distance between the groups, since the rows of first two PCs indicates the coordinates of objects in 2-dimensional illustration. Manhattan or city block distance is used to calculate the distances between the groups as shown in Equation (4.2).

$$d_{ij} = \sum_{n=1}^{n} \left| t_{in} - t_{jn} \right| \tag{4.2}$$

where n is the number of groups, *i* and *j* are the objects are existing in the same sample groups, d_{ij} is the value of calculated Manhattan distance, *t* is the elements of score matrix obtained from PCA. After calculating the distances, the classes will be constructed. In order to get accurate results a "check value" is assigned for the system. It is defined depending on the investigation and predefined by the user. Equation (4.3) represents the check value.

$$check = \sum_{\substack{i=1 \\ j=1}}^{n} d_{ij} \tag{4.3}$$

where d_{ij} is the distance between the groups. After determination of check values, the chromosomes are ranked according to their check values, the best one has the highest ranking value. These check values are used to define the selection probability of parents for the next generation.

4.2.1.3. Selection of Parents

This step is used to select the chromosomes from the mating pool. In this step roulette wheel selection method is used. The chromosomes with highest fitness value have a larger area on the wheel and highest probability of selection whereas the chromosomes with lowest fitness value have a smaller area and lowest selection

probability. The chromosomes with a larger fitness value have greater chance for further step of the genetic algorithm.

4.2.1.4. Crossover and Replacing the Parents by Their Offspring

Crossover makes combinations of selected parent chromosomes in order to get offspring. In this algorithm, single point crossover is used in a random way. The parent chromosomes are divided into two parts with a random length. Indeed, crossover can generate offspring which may be better or worse than their parents. Therefore the evaluation process is again applied to the offspring generation to obtain the best individuals and to eliminate the worst ones.

4.2.2. Genetic Algorithm Based Discriminant Analysis (GADA)

GADA has same steps just told in GAPCA, however, it shows some differences in the step of evaluation of the genes and the whole method is based on the full cross validation, in other words leave-one-out sample. In the evaluation step, after obtaining score matrix with variance larger than explained value in the PCA part, the T matrix is separated into predefined groups. Variance-covariance matrix of each group (Equation (4.4) is found in order to use in the calculation of Mahalanobis distances (MD) shown in Equation (4.5).

$$C_{Ti} = \frac{1}{n-1}(T_i)^T(T_i) \tag{4.4}$$

$$MD = \sqrt{(t_{ij} - \bar{t}_i) * C_{T_i}^{-1} * (t_{ij} - \bar{t}_i)^T} \tag{4.5}$$

where $\mathbf{C}_{Ti}$ is the variance-covariance matrix of each group, T_i is the score matrix of each group, t_{ij} is the score value of the each group in each column. All MD's are calculated and used to classify samples. In order to obtain the most valid classification, there should be a value which will be the criterion of validity. As it is shown in Figure 4.3, calculated MD's refer to the distance between the central point of a group and its corresponding sample (MD_{AA}), and the distance between the central point of the other

group and the sample of a different group (MD_{AB}). Same calculations are done for the MD_{BB} and MD_{BA}. In this study, the criterion value is calculated as shown in Equation (4.6) and named as “check value”:

$$\text{CheckValue} = \frac{\sum_{n=1}^{i} MD_{AA_i} + MD_{BBi}}{\sum_{n=1}^{i} MD_{AB_i} + MD_{BAi}} \tag{4.6}$$

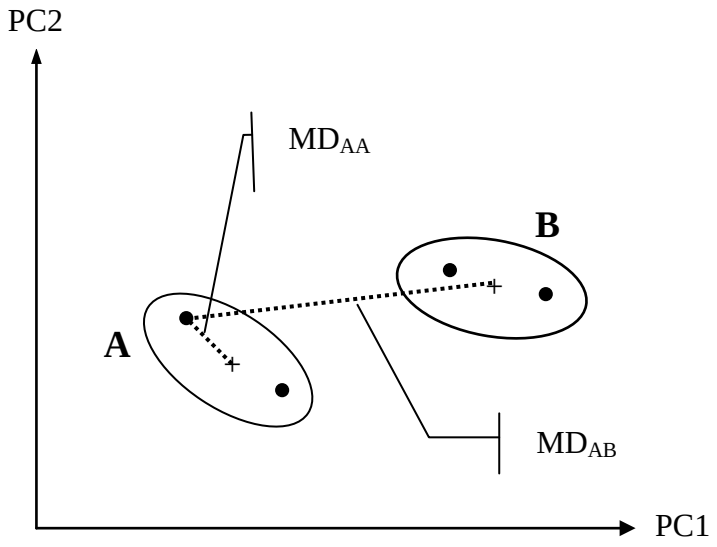


Figure 4.3. Schematic representation of Mahalanobis distances in PC space.

All the other genes in initial generation are determined by this way and a check value is assigned to each of them. After the calculation of the check values for all the genes in the population, the genes are sorted from largest to smallest fitness and the best one is reserved for the comparison with the best of the next generation. Then best parent genes are selected with roulette wheel selection method for meting and breeding, following by crossover and replacing the parent genes with their offspring. This whole cycle is based on leave-one-out cross validation is continued until a predefined number of iteration is reached. At the end the gene that has the highest classification power is selected to analyze the data at the final step. Because the GA based supervised pattern recognition is based on a lot of random processes, it expected that whenever the algorithm is rerun, it will generate a somewhat different result. For this reason, the

algorithm is designed to run multiple times for a given classification problem and it is possible to make a comparison among these runs in terms of the similarities and dissimilarities of the best genes of each run. At the end of the whole method Coomans plot is used for the interpretation of the results. The distance from the model for class 1 is plotted against that form model 2. On both axes, critical distances are shown in order to define the classes, overlap zone, and outlier zone. By plotting objects in this plot, the classification results can be clearly visualized. Critical distances are calculated as shown in Equation (4.7) and (4.8).

$$s_0 = \sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(d_{ji})^2}{m - PC - 1}} \tag{4.7}$$

$$s_{crit.} = \sqrt{F_{crit.} * s_0^2} \tag{4.8}$$

where d is the value of distances of samples are existed in each class, m is the number of samples, PC is the number of principal components that are used in the modeling of classes, $F_{crit.}$ is the tabulated value for (m-PC-1) degrees of freedom for each classes at a significance level of 95%.

CHAPTER 5

VEGETABLE OILS AND OLIVE OILS

In analytical chemistry the identification of chemical and physical properties of food and beverages have an important role. Since the amount of the components of food and beverage samples are used to define the quality of foods for the consumers. In this study the classifications of vegetable oils and olive oils were examined using developed methods GAPCAD and GADA, separately. The reason of choosing the vegetable oils and olive oils is the economical importance of these foods in the world. In this chapter the definitions of vegetable oils and olive oils will be given and the reasons of the importance of classification will be discussed.

5.1. Vegetable Oils

All most all pants contain oils in their seeds. For a plant to be suitable for oil production has two criteria. The first one is announced as that the plant must be suitable for high acreage cultivation. On the other side plants contain in a different amount of oil in their seed. Therefore the second one is stated as the oil content must reach the minimum commercially viable exploitation. Of course the only exceptions are plants that contain oils unique in their composition or with properties that cannot be found elsewhere (Bockisch 1998). There are mainly two clusters of vegetable oils according to their source namely pulp and seed oils. Within these two groups further categorization is possible, usually based on fatty acid composition. Table 5.1 shows the classes of pulp and seed oils. Generally the fatty acid composition of oils has an important role in the identification of economic value and the nutritional value or oxidation stability.

The list of natural fatty acids exceeds 1000, but commercial interest is limited to a smaller number perhaps around 20. Ignoring the lipid membrane, rich in a-linolenic acid and present in all green tissue, the three dominant acids in the plant kingdom are palmitic, oleic, and linoleic, sometimes accompanied by stearic acid and by linolenic acid. Others, occurring in specialty oils, include myristic, lauric, erucic, hexadecenoic, petroselinic, γ-linolenic acid, eleostearic and isomers, ricinoleic, and vernolic (Gunstone

2005). Table 5.2 assigns the more common fatty acids existing in vegetable oils. The classification of oils according to their fatty acid composition is also given in

Table 5.1. Classification of Oils
(Source: Bockisch 1998).

Pulp Oils	Seed Oils
Olive Oil	Sunflower Oil,
Palm Oil	Cottonseed Oil
Avocado Oil	Corn Oil
	Pumpkin Oil
	Sesame Oil
	Linseed Oil

Table 5.2. Structures of the More Common Acids in Vegetable Oils
(Source: Gunstone 2005)

Trivial Name	Structure	Unsaturations (If Any)
Saturated		
Lauric	12:0	---
Myristic	14:0	---
Palmitic	16:0	---
Stearic	18:0	---
Monounsaturated		
Oleic	18:1	9c
Petroselinic	18:1	6c
Erucic	22:1	13c
Polyunsaturated (non-conjugated)		
Linolenic	18:2	9c12c
Linolenic (α)	18:3	9c12c15c
Linolenic (γ)	18:3	6c9c12c
Polyunsaturated (conjugated)		
Eleostearic	18:3	9c11t13t
Calendic	18:3	8t10t12c
Oxygenated		
Ricinoleic	18:1	12-OH 9c
Vernolic	18:1	12,13epoxy 9c

Table 5.3. Vegetable oils by fatty acid type
(Source: Gunstone 2005)

Acids	Vegetable Oil
Lauric	coconut, palm kermel
Palmitic	palm, cottonseed
Oleic/Linoleic	groundnut, safflower, sesame, sunflower, cottonseed, canola, soybean
High oleic	olive, safflower, sunflower, canola, groundnut, soybean
Linoleic	linseed, canola, soybean
Vegetbale butters	cocoa butter
Erucic acid	HEAR (high erucic acid rapeseed oil), crambe
Conjugated acid	tung, calendula
Oxygenated acids	castor, vernolic

In this study, sunflower oil, corn oil and olive oil samples were used as vegetable oils. These vegetable oils will be defined and explain in terms of chemical and physical properties. Only olive oil is investigated in detail since olive oil samples were classified according to both their geographic origin and fatty acid ester composition.

5.1.1. Sunflower Oil

5.1.1.1. Definition of Sunflower Oil

Sunflower (*Helianthus annuus L.),* one of the most ancient oilseed species in North America, belongs to the family Compositae (*Asteraceae*) and the genus Helianthus. Sunflower was introduced into Europe by the Spanish explorers returning to the continent at the beginning of the 1500s A.D. Starting from Spain, sunflower crops spread rapidly through France and Italy, and toward the north and east of Europe. In several regions, it was a source of smoking leaves, flowers for consumption in salads, or for the manufacture of paint, edible, and medicinal seed, and cooking oil. But it was, perhaps, the beauty in the inflorescence of sunflowers that interested the first growers, large and bright yellow, always facing the sun. Hence, the name of the genus, Helianthus, derived from the Greek *helios* meaning sun and *anthos* meaning flower; and its Spanish, English, French, and German words: *girasol*, *sunflower*, *tournesol* and *Sonnenblumen* (Grompone 2005).

5.1.1.2. Chemical and Physical Properties of Sunflower Oil

Sunflower oil just like most vegetable oils is composed mainly of triacylglycerols (98–99%), and a small fraction of phospholipids, tocopherols, sterols, and waxes (all of the latter are commonly referred to as the ‘‘unsaponifiable fraction’’). Table 5.4 shows the variation range for major fatty acids in percentage of sunflower oil. Grompone was indicated the results according to the research results of Merrien in 1998 and American Oil Chemists' Society (AOCS) that is presented in 1997. Two important criteria regarding the composition of sunflower oil make worth than the other vegetable oils

- It provides an essential fatty acid which is linoleic acid.
- It has low amount of palmitic acid compared to the other vegetable oils. Since it is believed that palmitic acid is not good for the human health.

As expected from its high linoleic acid content, the main triacylglycerol is trilinolein (36.3%), followed by oleo-dilinolein (29.1%); triolein being practically nonexistent (0.6%). Thus, the percentage of triacylglycerols (TAG) with four or more double bonds is higher than 80%. This TAG distribution is responsible for the low solidification point of regular sunflower oil (-16^0C to -19^0C), allowing, for example, storage of mayonnaise manufactured with regular sunflower oil in a refrigerator without breakage of the emulsion (Grompone 2005). Lastly it is widely used as a cooking oil and is valued an important components of as soft spreads (Gunstone 2005).

Table 5.4. Variation range for major fatty acids (%) of sunflower oil
(Source: Grompone 2005)

Fatty Acid	AOCS	Merrien
16:0	5 – 8	5 – 7
18:0	2.5 – 7	4 – 6
18:1	13 – 40	15 – 25
18:2	40 – 74	62 – 70
18:3	<0.3	<0.2

5.1.2. Corn Oil

Corn oil is a major vegetable oil with an annual production of around 2 million tons obtained from corn or maize (*Zea mays*) by wet milling, particularly in the United States. The major acids are palmitic (9–17%), oleic (20–42%), and linoleic (39–63%), and the major triacylglycerols are typically LLL (15%), LLO (21%), LLS (17%), LOO% (14%), LOS (17%), LSS (5%), OOO (6%), and OOS (4%). Despite its high unsaturation, the oil has good oxidative stability. The refined oil is used as frying oil, salad oil, and in the production of spreads after partial hydrogenation (Gunstone 2005)

5.2. Olive Oil

In first aspect olive oil seems the most consumed and produced vegetable oil in the world. Olive oil production is mainly concentrated on Mediterranean countries such as Spain, Italy, Portugal, Turkey, Tunisia, and Morocco. These seven countries alone account for 90% of world production. The evolution of production and consumption shows a slight growth from the 1970s to the early nineties. In the mid 1990s there was a strong increase both in production and consumption. Despite the production fall that came afterwards, consumption did not decrease (Oliveoillife.com 2009)

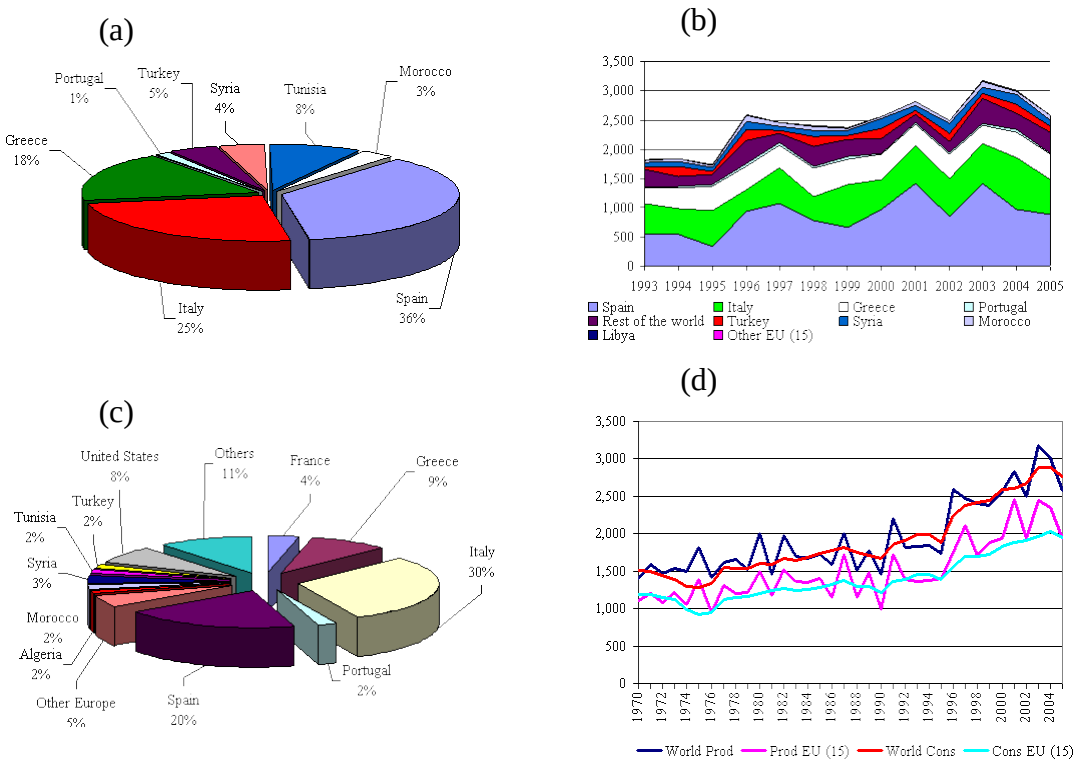


Figure 5.1. Statistical graphs of consumption and production of olive oils in the world. a) Main producing countries in 2005, b) Production of olive oil, 1993-2005 (1,000 tones), c) Main consuming countries in 2005, d) Production and consumption of olive oil in the world and in the European Union, 1970-2005 (1,000 tones) (Source: UNCTAD based on data from the Report on the proceedings of the 86th session of the International Oil Council June 2002).

5.2.1. Olive Oil Definitions

According to the International Olive Oil Council (IOOC) the international definition of olive oil is that oil produced by extraction of the fruit of the olive tree (Olea Europaea Sativa Hoffman et Link) to the exclusion of oils obtained using solvents or reesterification processes and of any mixture with oils of other kinds (Firestone 2005). There two main clusters of olive oils. The first one is the virgin olive oil, and the other is olive-pomace oil.

1. Virgin olive oil: It is the oil obtained from the fruit of the olive tree solely by mechanical or other physical means under conditions, particularly thermal conditions, that do not lead to alterations in the oil, and which has not undergone any treatment other than washing, decantation, centrifugation, and filtration. Virgin olive oil is

designated as nature oil is categorized into four different groups. Table 5.5. shows the categories of virgin olive oils.

Table 5.5. Categories of virgin olive oil
(Source: Internationaloliveoil.org 2009)

Olive Oil	Amount of Free Acidity (expressed as oleic acid) in 100 grams Olive Oil.
Extra Virgin Olive Oil	0.8 grams
Virgin Olive Oil	2.0 grams
Ordinary Virgin Olive Oil	3.3 grams
Lampante Virgin Olive Oil	3.3 grams

On the other side there are some different definitions for the different types of olive that are not fit into virgin olive oil. These types of olive oils have different amount of oleic acid. *Refined olive oil* is the olive oil obtained from virgin olive oils by refining methods which do not lead to alterations in the initial glyceridic structure. It has a free acidity, expressed as oleic acid, of not more than 0.3 grams per 100 grams (0.3%) and its other characteristics correspond to those fixed for this category in this standard. This is obtained by refining virgin olive oils which have a high acidity level and/or organoleptic defects which are eliminated after refining. Over 50% of the oil produced in the Mediterranean area is of such poor quality that it must be refined to produce an edible product. Note that no solvents have been used to extract the oil but it has been refined with the use of charcoal and other chemical and physical filters. An obsolete equivalent is “pure olive oil”. *Olive oil* is the oil consisting of a blend of refined olive oil and virgin olive oils fit for consumption as they are. It has a free acidity, expressed as oleic acid, of not more than 1 gram per 100 grams (1.0%). The cheap refined oil is mixed with flavorful virgin oil.

2. Olive-pomace oil: Pomace is the ground flesh and pits after pressing. Olive-pomace oil is the oil obtained by treating olive pomace with solvents or other physical treatments, to the exclusion of oils obtained by re-esterification processes and of any mixture with oils of other kinds. There are mainly three types of olive-pomace oil: *Olive-pomace oil* is the oil comprising the blend of refined olive-pomace oil and virgin

olive oils fit for consumption as they are. It has a free acidity of not more than 1 gram per 100 grams and its other characteristics correspond to those fixed for this category in this standard. In no case shall this blend be called “olive oil”. *Crude olive-pomace oil* is olive -pomace oil whose characteristics correspond to those fixed for this category in this standard. It is intended for refining for use for human consumption, or it is intended for technical use and refined olive-pomace oil is the oil obtained from crude olive -pomace oil by refining methods which do not lead to alterations in the initial glyceridic structure. It has a free acidity, expressed as oleic acid, of not more than 0.3 grams per 100 grams. All of these definitions are directly taken from the sources of IOOC.

CHAPTER 6

MATERIALS AND METHODS

6.1. Samples

6.1.1. Olive Oil Samples

Totally 108 of olive oil samples were obtained from TARİŞ, the Union of TARIS Olive and Olive Oil Co-operatives. These olive oil samples can be categorized in two different groups. The first categorization is based on the origin of olive oil samples. All of the olive oil samples were collected from Aegean region of Turkey. Forty nine of them belong to the North Aegean and thirty one of them are from the South Aegean region. The remaining olive oil samples are the mixtures of natural and refined olive oils and their origin are not known. The second categorization is organized according to the value of free acidity of olive oil samples. The existed types of olive oil in the sample set are extra virgin, natural, refined, lampante and virgin olive oil with the number of twenty six, fifteen, thirty five, twenty six, and five, respectively. The acronyms used to indicate the olive oil samples for different categorization are given in the Table 6.1.

The samples stored in deep freezer at $+4^0$C until they were analyzed. There was no any prior treatment before the analysis of samples.

During the examination of developed genetic algorithm based classification methods, GAPCAD and GADA, all of the samples were not used. Due to the working principle of these algorithms, training and validation set must contain the same number of samples. Therefore training and validation set were organized as equal number of samples.

Table 6.1. Acronyms that used to identify the olive oil samples.

Categorization according to the geographical origin	
Sample Name	Acronym
North Aegean Olive Oil	NA
South Aegean Olive Oil	SA

Categorization according to the level of free acidity	
Sample Name	Acronym
Extra Virgin Olive Oil	EVOO
Lampante Olive Oil	LOO
Refined Olive Oil	ROO
Natural Olive Oil	NOO
Virgin Olive Oil	VOO

6.1.2. Vegetable Oils

Totally 34 samples of vegetable oils with different brands were purchased from the global markets. Three different vegetable oils were chosen for the classification studies. The studied vegetable oil samples were olive, sunflower, and corn oil. These vegetable oils were categorized as pulp and seed oils. Olive oil is accepted as pulp oil whereas the remaining are seed oils. Table 6.2 shows the abbreviations of each type of vegetable oils.

Table 6.2. Acronyms of vegetable oils.

Vegetable Oil	Acronym
Olive Oil	OO
Sunflower Oil	SFO
Corn Oil	CO

The samples stored in deep freezer at $+4^0$C until they were analyzed. There was no any prior treatment before the analysis of samples.

6.2. Near Infrared Measurements

FTS-3000 NIR spectrometer (Bio-Rad, Excalibur, Cambridge, MA) near Infrared (NIR) spectrometer system was used to measure the olive oil samples at room temperature in 2 mm cuvette. The working range was set to 10,000 – 4000 cm^{-1} wavenumber with 8 cm^{-1} resolution by averaging 64 scan numbers. Both blank and sample spectra were collected in absorbance method.

Background was spectrum was obtained empty and dry quartz cell. The same number of scans, resolution, and scan numbers were using in the observation of background spectrum as the sample spectra. Before and after each sample analyses background was collected to reduce the contaminations that might come from sample cuvette. Quartz cuvette was cleaned with pure acetone and allowed to dry.

6.3. Middle Infrared Measurements

Spectrum 100 (Perkin Elmer, Waltham, MA, USA) Fourier Transform Infrared (FTIR) spectrometer system coupled to attenuated total reflectance (ATR) accessory was used to measure the olive oil samples. ATR (Miracle ATR, Pike Technology, Madison, WI, USA) equipped with ZnSe crystal plate was used to analyze the olive oil samples at room temperature. The working range was set to 630 – 4000 cm^{-1} wavenumber with 4 cm^{-1} resolution (data point interval of 1 cm^{-1}) by averaging 64 scan numbers. Both blank and sample spectra were collected in absorbance method.

Background was spectrum was obtained empty and dry ATR cell. The same number of scans, resolution, and wavenumbers interval were using in the observation of background spectrum as the sample spectra. Before and after each sample analyses background was collected to reduce the contaminations that were come from the ATR crystal. ATR crystal was cleaned with pure ethanol and allowed to dry.

6.4. Fluorescence Measurements

Excitation–emission fluorescence (EEF) and total synchronous fluorescence (TSyF) spectra were obtained using Varian Cary Elipse spectrofluorometer (Varian, Inc. Hansen Way, Palo Alto, CA) equipped with a xenon flash lamp. In both measurement modes, the detector was set to 600 V. The slid with of both monochromators were set to 5 nm. The acquisition interval was maintained as 1 nm and the right angle geometry was used to analyze olive oil samples in a 10 mm quartz cuvette. Three dimensional EEF spectra were obtained by measuring emission spectra in the range of excitation 400–850 nm, repeatedly, at excitation wavelengths from 300–435 nm with the 15 nm intervals in the excitation domain. Three-dimensional TSyF spectra were acquired by measuring the synchronous fluorescence spectra in the wavelength range of 250–800 nm, repeatedly, at offset values ($\Delta\lambda$) from 50–100 nm with the 10 nm intervals.

6.5. Data Analysis

Four different classification techniques were used in the data analysis which were singular value decomposition based principal component analysis (SVD-PCA), genetic algorithm based discriminant analysis (GADA), distance based genetic algorithm principal component analysis (GAPCAD), and soft independent modeling of class analogy (SIMCA). Developed genetic algorithm based and principal component analysis statistical processes were implemented in Matlab R2009a (MathWorks Inc, Natick, MA). SIMCA was implemented in SIMCA-P 10.5 (Umetrics, Umeå,Sweeden)

CHAPTER 7

RESULTS AND DISCUSSION

In this study, two different supervised classification methods which are based on genetic algorithm were developed. These two methods are named as distance based genetic algorithm principal component analysis (GAPCAD) and genetic algorithm discriminant analysis (GADA). Both methods were examined in various spectral data. These are near infrared (NIR) spectral, Fourier transform infrared (FTIR) spectral, fluorescence spectral (in two different modes: there way array of excitation–emission fluorescence and total synchronous fluorescence) data matrices. The spectral data were obtained from the measurement of different types of vegetable oil samples and olive oil samples and the classification of vegetable oils and olive oils were studied. Secondly the classification of olive oil samples was performed based on their chemical and physical composition. In this chapter, all the classification results will be discussed in detail for both supervised classification methods and all the sample types.

7.1. Classification of Olive Oils Based on Chemical Properties

Olive oil quality is largely controlled by the grower and processor. During the harvest and storage, bruising or damage to the fruit will result in reduced quality in the oil. Olive growers therefore need to have basic understanding of what oil quality is and how it is preserved. This prime fact covers a range of topics related to quality in olive oil including basic composition of olive oil and composition of virgin olive oils.

As it mentioned before, the main components of almost 97–98% of whole olive oil are substances of a glyceride nature concentrated in the pulp and seed. The remaining nonacylglycerol lipid fraction is a mixture of compound classes, including alkanes, squalene, wax esters, aliphatic alcohols and aldehydes, tetracyclic (sterols) and pentacyclic triterpenes (acids, alcohols and esters), free fatty acids, vitamins, phospholipids, polyphenols and glycosides, distributed in the various parts of the fruit The free fatty acid proportions are in the narrow 1–3% range for olive oils obtained by simply pressing the fruit, whereas these can be as high as 10–15% for solvent-extracted

oils. It is the minor compound classes, their concentrations and relative percentages that are the determinants of olive oil characterization and commercial grading. Furthermore, some of these minor components are determinant factors of oil stability, as well as being relevant from the hedonistic and salutary points of view (Table 7.1). The natural concentration of these minor components in an oil can vary greatly, being related mainly to the cultivar, the stage of maturity of the fruits, the soil, the climate and also to the extraction technique adopted (Bianchi 2002).

Table 7.1. Significance of the olive oil parameters
(Source: Bianchi 2002)

Parameters	Significance
Free Fatty Acids (%)	Deterioration of olive oils
Peroxide Value (meq O_2/kg)	Presence of hydroperoxides
Halogenated solvents	Detection of harmful contamination
Phenols (mg/kg)	Antioxidants
Induction time (h)	Stability, resistance to oxidation
Chlorophyll Pigments (mg/kg)	Influence on oil acceptability by consumer
K_{232}, K_{270}	Double bond conjugation: (i) drastic thermal and chemical treatment of oil; (ii) oxidation
Panel Test	Organoleptic analysis

The chemical composition of olive oil samples were purchased from the TARİŞ was identified by the manufacturer and all the analysis results were also taken together. Totally twenty six parameters of olive oil samples were organized to classify of these samples into two different ways. The first one was based on the geographical origin. The training set was designed as including both the North Aegean (NA) and South Aegean (SA) olive oils with totally forty samples. An independent test set was constructed using totally twenty olive oil samples. Both training and test sample sets were containing twenty of NA, twenty SA, ten of NA, ten of SA olive oil samples, respectively.

Before investigation of the results of the classification methods, the acronyms of each variable are given in Table 7.2.

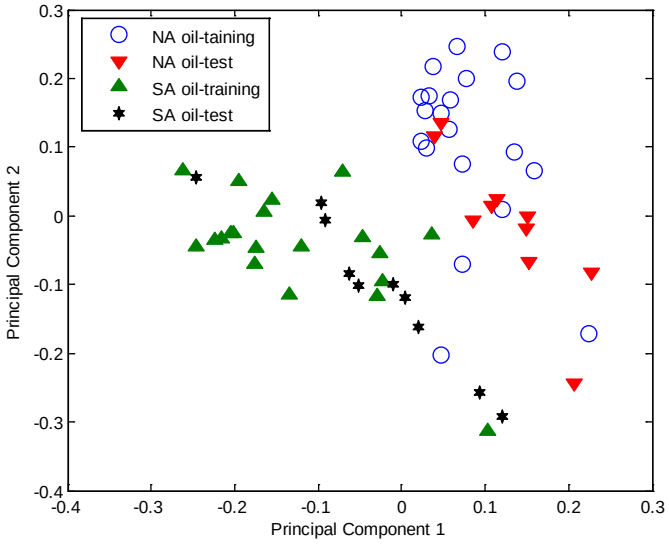
Table 7.2. The corresponding acronyms of each variable.

Variables	Acronyms	Variables	Acronyms
Free Fatty Acid (Oleic acid %)	FFA	Sterol Composition	
Peroxide Value (meq O_2 / kg oil)	PERO	Cholesterol	CHOL
UV Absorbance Value	UVA	Brassicasterol	BRA
K (λ=232 nm)	K232	Campesterol	CAM
K (λ=270 nm)	K270	Stigmasterol	STIG
ΔK	ΔK	Δ7 Stigmastanol	D7
Trilinolein %	TRI	β- Sitosterol	BSTE
Difference between the real and theoretical value of ECN 42 triglyceride	TRG	Total Sterol (mg/kg)	TOSTE
		Eritrodiol+Uvaol	ERIUVA
Fatty Acid Composition			
Myristic acid	MYRA		
Palmitic acid	PALA		
Palmitoleic acid	PALMA		
Heptadecanoic acid	HEP		
Stearic acid	STEA		
Oleic acid	OLEA		
Linoleic acid	LINOA		
Linolenic acid	LIA		
Arachidik acid	ARA		
Gadoleic acid	GADA		
Behenic acid	BEA		
Lignoceric acid	LIGA		

7.1.1. Classification Olive Oils Based Geographical Origin Using GAPCAD and SVD-PCA

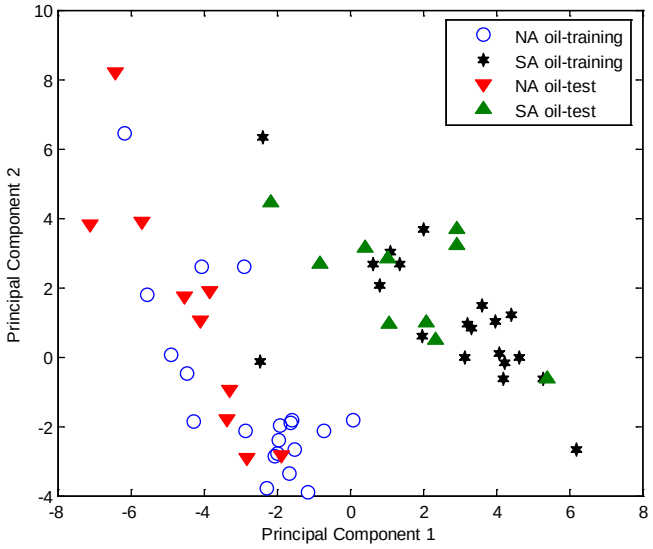
GAPCAD was performed to classify the olive oil samples according to their geographical origin. The initiated data matrix has 40x26 (samples x variables) dimensions. The algorithm was initiated 100 genes and 100 iterations. Also the training and test set was combined and examined using SVD-PCA. Both results were compared in order to prove the success of developed genetic algorithm based classification methods. Before starting the examination of both algorithms autoscaling was performed to the data matrix in order to organize the variables in one scale.

SVD-PCA



(a)

GAPCAD



(b)

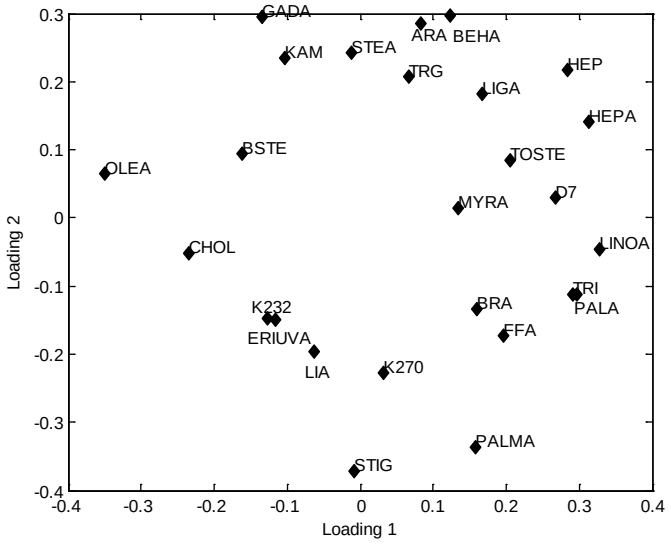
Figure 7.1. Score plots of principal components calculated from the chemical variables of olive oil samples a) SVD-PCA, b) GAPCAD.

Totally twenty six principal components (the data matrix contained twenty six variables.) were calculated in the examination of SVD-PCA. The first two principal components (25.24% of explained variance) were plotted against each other in order to

visualize the classes of the North and South Aegean olive oil samples in the space (Figure 7.1.a). As it seen from the score plot that obtained from the SVD-PCA, olive oil samples are laid on principal component 1 (PC1). Generally the positive scores of PC1 refer the distribution of SA olive oil samples whereas the NA olive oil samples are scattered on negative scores of PC1. Only a few samples of SA olive oil samples cannot be distinguished from the class of NA olive oil samples.

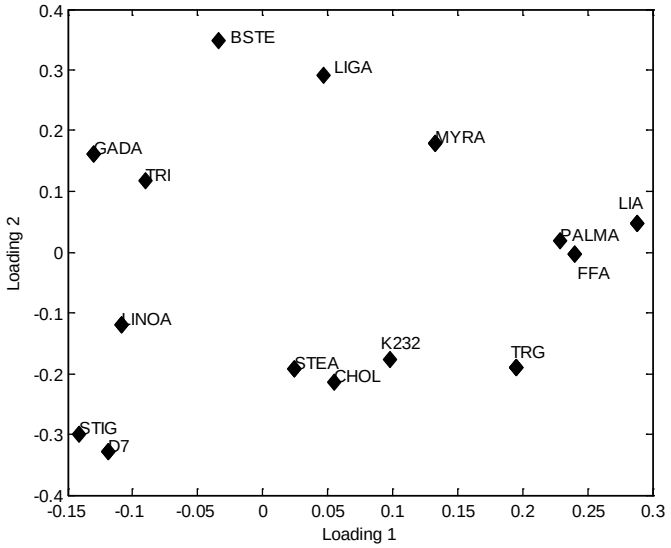
The resulting plots (score and loading) of GAPCAD are shown in Figure 7.1.b and Figure 7.2.b. The first one is the score plot of the first two principal components with 33.77% of explained variance. The South Aegean olive oil samples are scattered on the positive values of PC1 whereas the North Aegean olive oil samples are on negative scores of PC1. This distribution emphasize that the principal component 1 is the most explained component. The results of the GAPCAD also show a better distribution of two classes in the space. Due to the working principle of GA which is based on the selection of features, all the variables that are defined at the beginning of the algorithm will not be used in the classification of samples. Figure 7.2.b shows the distribution of selected variables which have the most influenced information about olive oil samples. The same path is followed in the explanation of the selected variables. The variables distributed on the positive values of loadings take a role in the identification of SA olive oil samples and for the NA olive oil samples vice versa.

SVD-PCA



(a)

GAPCAD



(b)

Figure 7.2. Loading plots of principal components calculated from the chemical variables of olive oil samples a) SVD-PCA, b) GAPCAD.

As a result the GAPCAD gives more efficient way in the identification olive oil samples, since genetic algorithm is selected the variables which have the most influence on the classification of samples. Knowing only these variables will be enough in the

classification of olive oil samples based on the geographical origin. These reduced the analysis time, cost of the experiments for the analyzers.

7.1.2. Classification Olive Oils Based Geographical Origin Using SIMCA and GADA

The chemical composition of olive oil samples were purchased from the TARİŞ was identified by the manufacturer and all the analysis results were also taken together. Totally twenty six parameters of olive oil samples were organized to classify of these samples into two different ways. The first one was based on the geographical origin. The training set was designed as including both the North Aegean (NA) and South Aegean (SA) olive oils with totally forty samples. An independent test set was constructed using totally twenty olive oil samples. Both training and test sample sets were containing twenty of NA, twenty SA, ten of NA, ten of SA olive oil samples, respectively.

SIMCA and GADA were performed to classify the olive oil samples according to their geographical origin. The two classes of olive oil were predefined as NA (North Aegean Olive Oil) and SA (South Aegean Olive Oil). The initiated training data matrix has 40x26 (samples x variables) dimensions, whereas the test data matrix has 20 x 26 dimensions. The GADA analysis was initiated 4 genes and 10 iterations. Both results were compared in order to prove the success of developed genetic algorithm based discriminant analysis method. Before starting the examination of both algorithms autoscaling was performed to the data matrix in order to organize the variables in one scale.

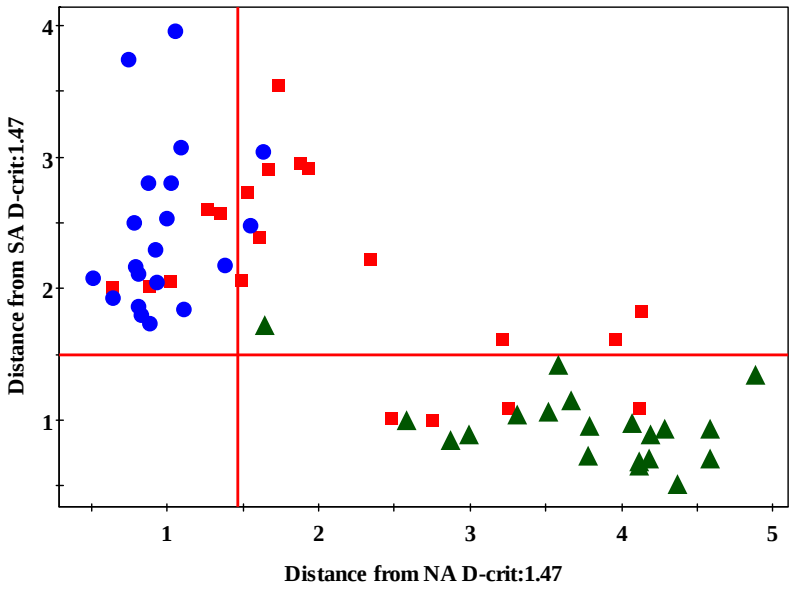


Figure 7.3. Cooman's plot of olive oil samples obtained from SIMCA analysis of NIR spectra (triangle: SA oils-training, circle: NA oils-training, box: test set).

Two principal components (54.50% of explained variance) for class of NA oil and three principal components (66.70% of explained variance) for class of SA oils were calculated in the examination of SIMCA. The Cooman's plot was drawn to visualize the classes of the North and South Aegean olive oil samples in the space (Figure 7.3). The critical limits that are used to define the boundaries between the classes were calculated as 1.47 at 95% confidence level. As it seen from the Cooman's plot that obtained from the SIMCA, olive oil samples mostly classified as North and South Aegean olive oil samples. However, the olive oil samples that existing in the training set coded as NA-5, NA-20 and SA-18 were classified as not belonging to any classes. In the results of test step, the 55% (11/20) of olive oil samples were found as not belonging to either NA oils or SA oils.

At the end of the GADA analysis, seven significant principal components were found out with a 92.53% of explained variance. The critical limits were calculated as 4.5 at 95% confidence level and the Cooman's plot was plotted to observe the distribution of olive oil samples (Figure 7.4). According to the critical limits for NA and SA classes, the training set was classified without any outlier samples or these classes do not contain any samples that have similar properties. However the test results of SA oil samples are not good as NA oil samples. The 50% of SA olive oil samples and NA-2

oil sample are classified as not belonging to any classes. In order to explain the reasons of classification results, the selected variables or chemical properties should be investigated.

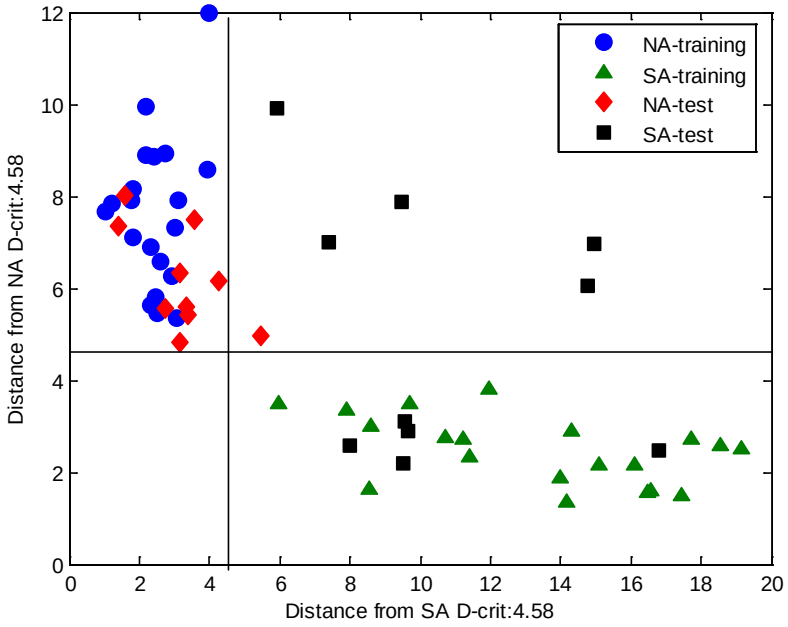


Figure 7.4. Cooman's plot of olive oil samples obtained from GADA analysis of NIR spectra.

As it is seen from the plot of loading vectors, there are only nine chemical variables that are used in the classification of olive oils according to their geographical origin. Generally the sterol composition of olive oils seems to be more effective in the classification of olive oil samples. The amount of sterol composition is very important in the determination of olive oil adulteration with high oleic acid content oil. In our country, the olive oils that are obtained from south region generally have high content of Δ7 Stigmastanol (D7) and this high level of D7 causes problem in the exportation of olive oils (Oliveoillife.com 2009). GADA analysis was classified the olive oil samples especially based on the sterol constituents. It should be noticed here, sterols does not affect the taste of olive oils. However it has very high influence on the chemical characterization of olive oils.

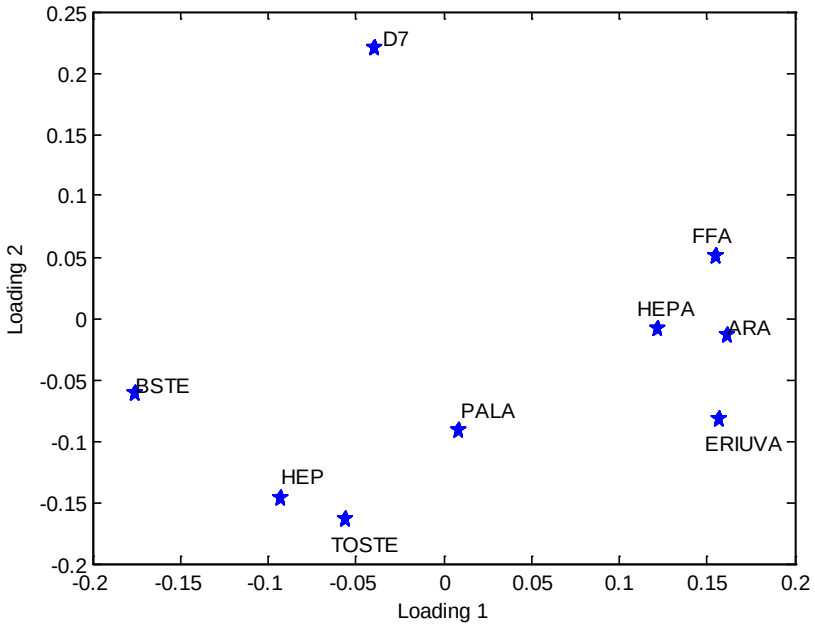


Figure 7.5. Loading plots of principal components calculated from the chemical variables of olive oil samples obtained using GADA.

7.2. Near Infrared Results

7.2.1. NIR Measurements of Olive Oil Samples

All the olive oil samples named as EVOO, ROO and LOO were analyzed using near infrared spectroscopy (NIR) in the range of 10,000 – 4000 cm^{-1}. Figure 7.6 shows the spectra of all olive oil samples in the range of 8900 – 4500 cm^{-1}. The region between the 4500 – 4000 cm^{-1} and 10,000 – 8900 cm^{-1} were discarded due to the noise and no relevant NIR peaks. Absorption maxima are clearly seen at 4590, 4656, 4700, 5675, 5786, 7074, 7170, 8240, 8560 cm^{-1} and smaller absorption bands at 5149 and 5250 cm^{-1}.

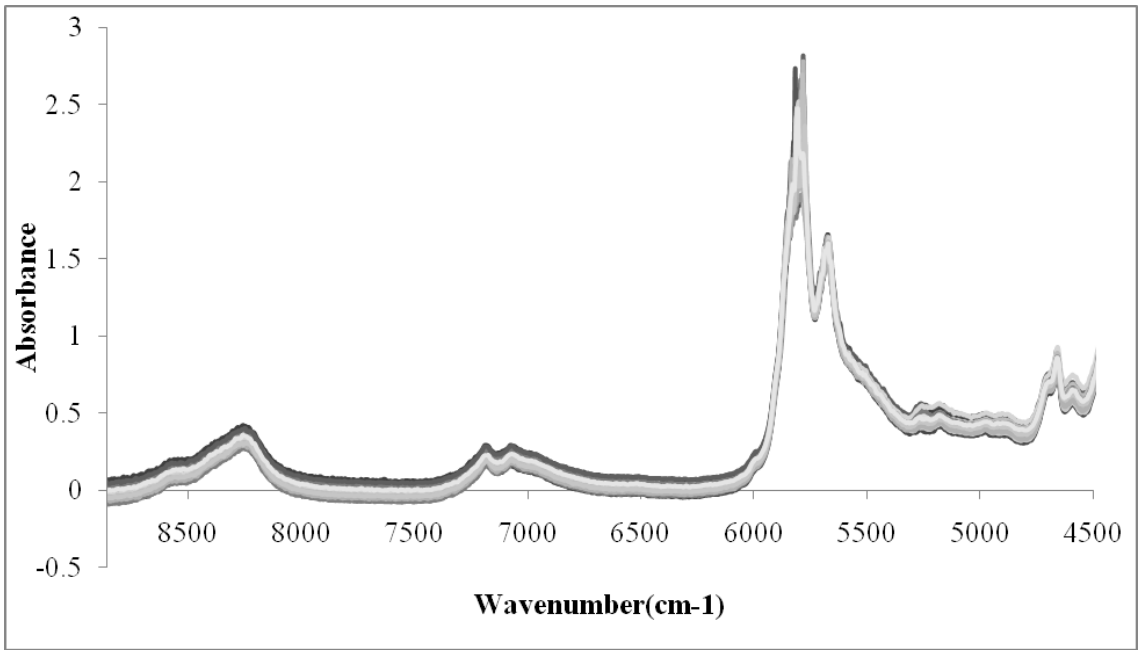


Figure 7.6. NIR spectra of olive oil samples measured in the range of 8900 – 4500 cm^{-1}.

The assignments of absorption maxima are given in Table 7.3. The overtones of and the combinations of the CH, OH, NH functional groups present in food samples give absorption maxima in NIR region. The low absorbance values of overtones and combinations allow using the NIR for the non destructive analysis of food samples. As it is seen from the NIR spectra plot of olive oil samples, NIR spectra generally contains numbers of broad and overlapping absorption bands. The nature of these overlapping bands makes qualitative and quantitative analyses difficult.

Table 7.3. Evaluation of NIR spectrum
(Source: Ozaki, et al. 2004).

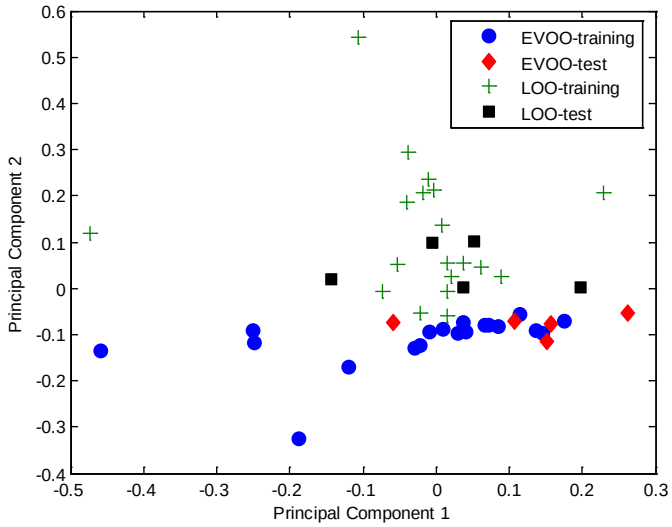
Wavenumber cm^{-1}	Functional Group
8560	CH_3, C-H strecthing 2^{nd} overtone
8240	CH_2, C-H strecthing 2^{nd} overtone
7170	CH_3, 2C-H strecthing
7074	CH_2, 2C-H strecthing
6000	*cis* $R_1CH=CHR_2CH_3$, *cis* CH
5786	CH_3, C-H strecthing 1^{st} overtone
5675	CH_2, C-H strecthing 1^{st} overtone
5250	C=O strecthing 2^{nd} overtone
5149	C=O strecthing 2^{nd} overtone
4700	COOR, C-H strecthing, C=O streching
4656	HC=CH, =CH strecthing, C=C strecthing
4590	HC=CH, CH asymmetric strecthing C=C streching

7.2.2. Classification Results of SVD-PCA and GAPCAD

7.2.2.1. Classification Results of Extra Virgin Olive Oils and Lampante Olive Oils

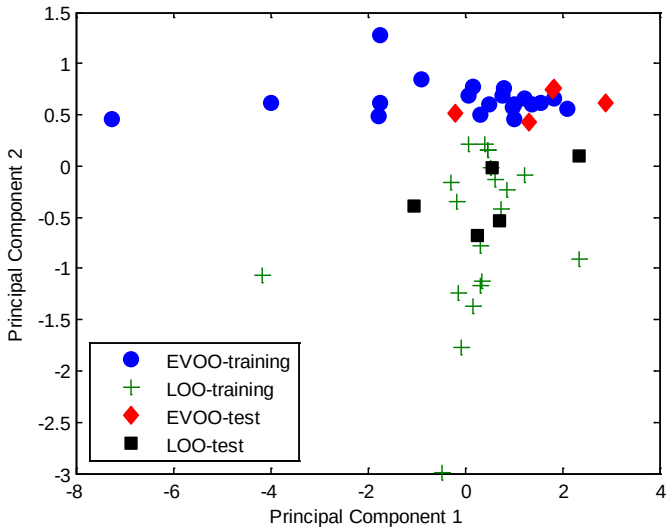
All NIR spectra of olive oil samples were designed into two different ways. In the examination of GAPCAD algorithm, NIR spectral data matrix into two different samples set: training set contains 38 olive oil samples and test set includes 10 olive oil samples. Both sample set were prepared independent from each other and randomly chosen olive oil samples. Spectral data matrix has 38 x 1079 (sample number x wavenumbers) dimensions for training set and 10 x 1079 dimensions for test set. Both training and test set were combined in order to observe the results of SVD-PCA algorithms. Both results were compared to each other. Before starting performing the classification methods, spectral data matrices were autoscaled to reduce the metric unit in one scale.

SVD-PCA



(a)

GAPCAD



(b)

Figure 7.7. Score plot of principal components calculated from NIR spectral data matrix of olive oil samples using a) SVD-PCA and b) GAPCAD for selected spectral data (ν = 8700 – 4500 cm^{-1}).

In SVD-PCA calculation, seven significant principal components were found out with 57.52% of explained variance. Principal component 1 (PC1) and principal component 2 (PC2) which have totally 44.50% of explained variance were plotted against to observe the distribution of olive oil samples. As it is seen from the Figure 7.7,

there are no definite olive oil classes. However we can say that EVOO samples are generally lay on the negative scores of PC2, whereas LOO samples are distributed on the positive scores of PC1. Same spectral data matrices were examined in GAPCAD calculations. GAPCAD algorithm was initiated with 100 genes and 100 iterations. At the end of the calculation 3 significant principal components (99.0% of total explained variance) were found out. The first two principal components with a 95.7% of explained variance were used to represent the score plot of olive oil samples.

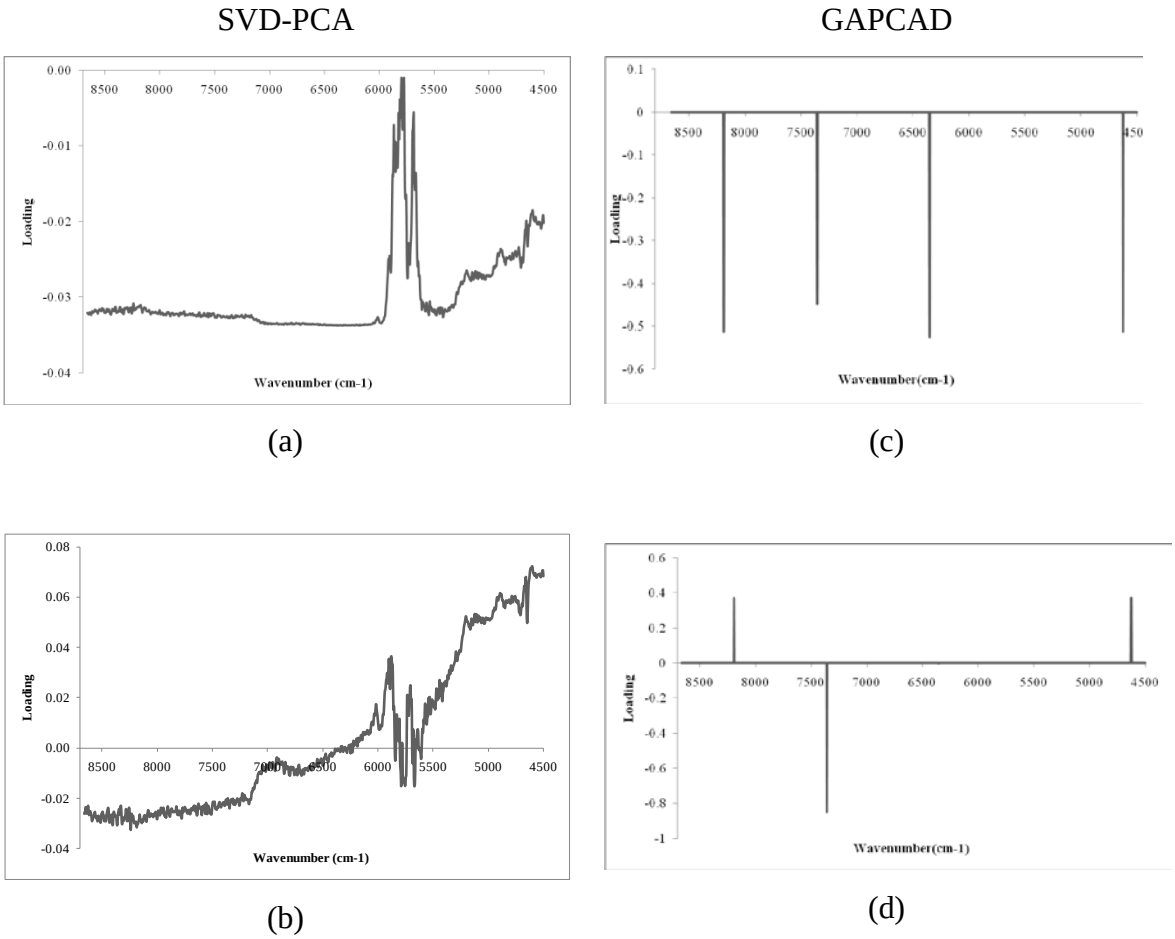


Figure 7.8. Loading plots of NIR spectral data of the olive oil samples (autoscaled data) obtained from the calculation of SVD-PCA a) loading of PC1 b) loading of PC2, and GAPCAD c) loading of PC1, d) loading of PC2.

Loading plots were obtained from both SVD-PCA and GAPCAD are given in Figure 7.8. and these plots refer the scores for PC1 and PC2 are given Figure 7.7. Loading plot of PC1 obtained from SVD-PCA shows that the overtones and

combinations of CH_3 and CH_2 stretching mostly have contribution on the classification of olive oils in the region of 6000 – 5500 cm^{-1}. The acidity of olive oils varies with different amount of fatty acids. Fatty acids have a general formula as CH_3-$(CH_2)_n$-*COOH* where n is typically an even number between 12 and 22. The acidity value of olive oils is generally expressed according to the amount of oleic acid. This fatty acid has 18 of C atoms with a single double bond. The loading plot PC2 (Figure 7.8.b.) proves the contributions of overtones and combinations exist in the region of 5000 – 4500 cm^{-1} have the most contribution in the classification of EVOO and LOO samples. In this region not only the stretching of C=C and =CH are assigned but also the contribution of oleic acid is emphasized.

On the other side, GAPCAD was selected only 4 wavenumbers that have the most contribution in the classification of olive oil samples. These wavenumbers are the 8190, 7357, 6355, 4625 cm^{-1} and the weight of these wavenumbers changes according to the scores and loadings matrices of PCA. Figure 7.8.c. and Figure 7.8.d. show the weight of loading for each wavenumbers. According to the scores of GAPCAD, the classification of olive oil samples generally is along on PC2. These contributions come from the overtones and combinations of CH_2, CH stretching and =CH, C=C stretching. In order to investigate the frequency of selected wavenumbers, GAPCA was repeated 100 times and the distribution of selected wavenumbers was plotted against the wavenumbers in the region of 8700 – 4500 cm^{-1}. Figure 7.9 shows the plot of frequency of selected wavenumbers.

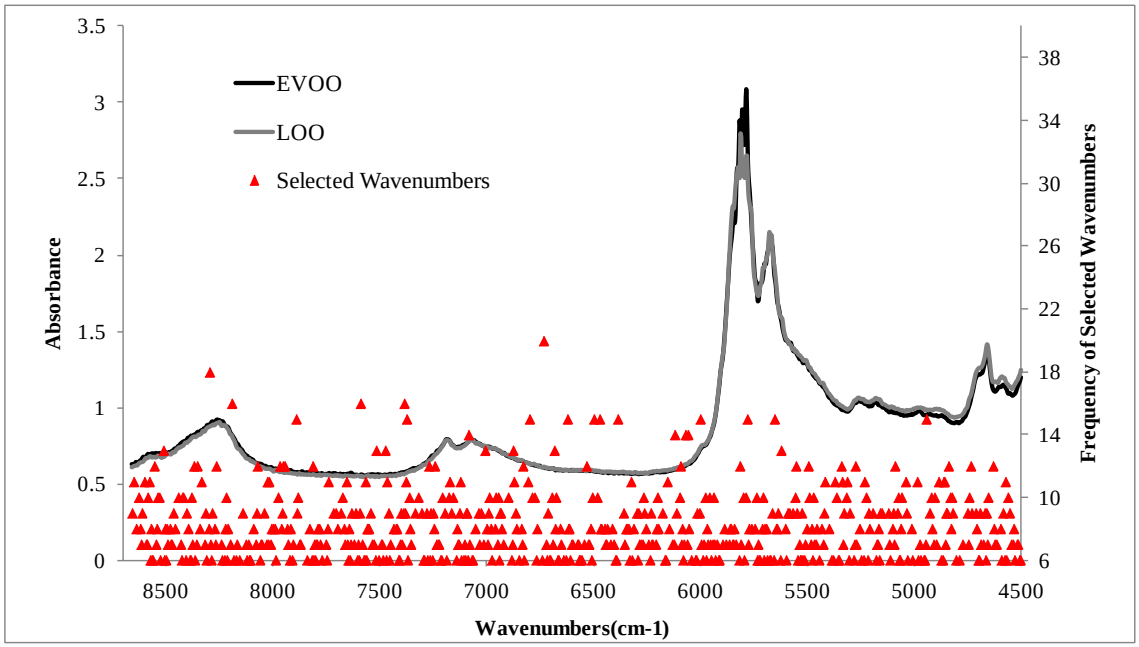


Figure 7.9. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of NIR spectral data.

As it is seen from the frequency of selected wavenumbers, the contribution of 2^{nd} overtones of CH_3 and CH_2 groups of fatty acids are larger than the other groups. The same result is also obtained from the loading values of corresponding scores.

7.2.3. Classification Results of SIMCA and GADA

7.2.3.1. Classification Results of Extra Virgin Olive Oils and Lampante Olive Oils

The training set with thirty eight samples, and test set with twelve samples were predefined at the beginning of the classification studies. The extra virgin olive oil samples containing 18 samples and lampante olive oil samples including 18 samples with their corresponding NIR spectral data were predefined as class 1 and class2, respectively. Before starting classification procedure, autoscaling was applied to the spectral data matrix as a preprocessing technique.

SIMCA analysis was firstly examined to the NIR spectral data matrix. Totally nine principal components were found out for both classes with a 92.40% and a 95.10% of explained variances, respectively. The critical limits were calculated as 1.38 at 95%

confidence level. The Cooman's plot was constructed and the boundaries of each class were identified by the help of the critical limits (Figure 7.10). As it is seen from the Figure 7.10, most of the extra virgin olive oil samples existing in the training set are classified as belonging to both classes, and also LOO-14 is classified in this region. On the other side, the samples of test set coded as LOO-1, LOO-2, LOO-3 and EVOO-1 belong to the neither class of EVOO nor class of LOO samples. Each row of spectral data matrix contains 1079 of wavenumbers with their corresponding absorbance values, therefore the identification of reasons of this classification results is difficult by naked eyes. Since all the wavenumbers were used in the classification.

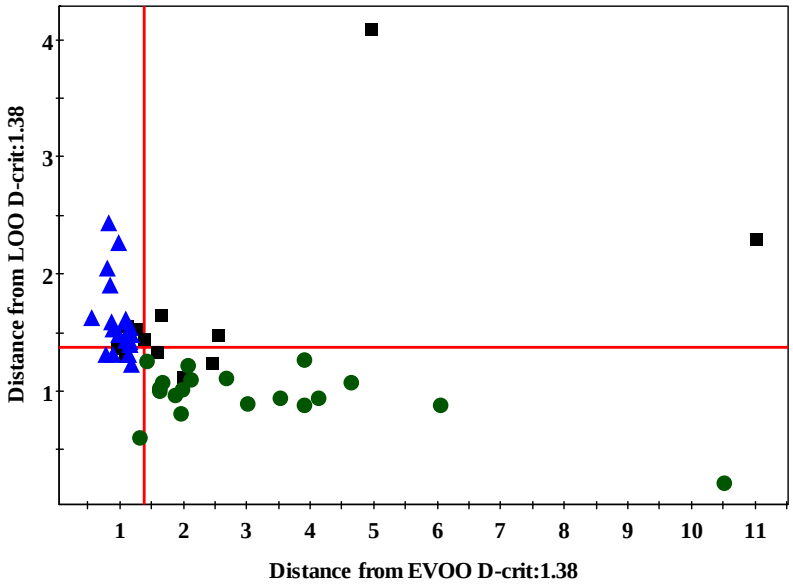


Figure 7.10. Cooman's plot of olive oil samples obtained from SIMCA analysis of TSyF spectra (triangle: EVOO-training, circle: LOO-training, box: test set).

Genetic algorithm based discriminant analysis was examined using the spectral data matrix. GADA analysis was initiated with 8 genes and 10 iteration numbers. Totally fifteen significant principal component analysis were found out with a 90.11% of explained variance. The critical limits that were used to identify the classes of olive oil samples were calculated as 8.32 at 95% confidence level. Cooman's plot of extra virgin olive oil and lampante olive oil samples was drawn to see the distribution of samples in the space (Figure 7.11). According to the results, the extra virgin olive oil samples coded as EVOO-2 and EVOO-6, and also the lampante olive oil sample coded

as LOO-12 existing in the training set show similar properties. Therefore they are assigned in the region of intersection of both classes. In the test step, the samples coded as EVOO-5, EVOO-4 belongs to the both classes, whereas the samples coded as EVOO-2, EVOO-3, LOO-3, and LOO-5 are classified as not belonging to none of classes. In order to see which wavenumbers have the most contribution on the classification of olive oil samples, selected wavelengths were plotted against the wavenumbers of NIR region.

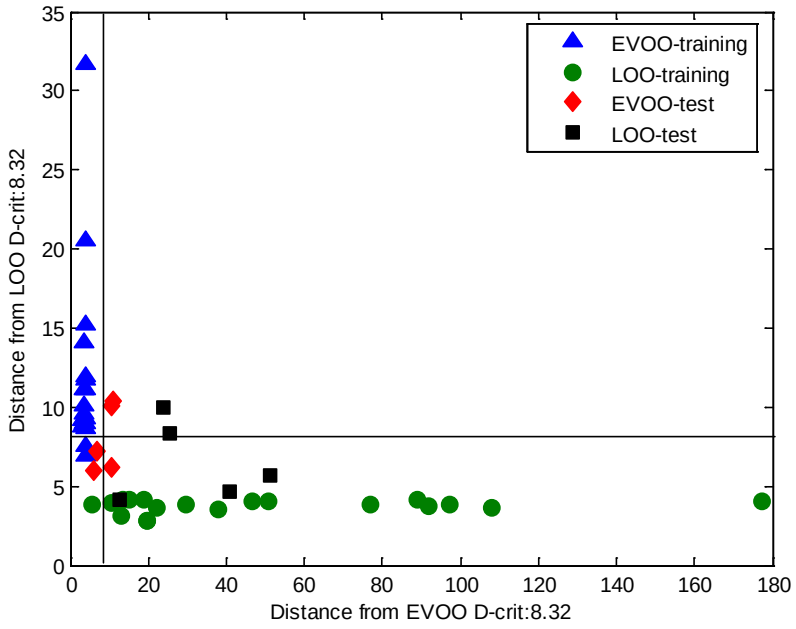


Figure 7.11. Cooman's plot of olive oil samples obtained from GADA analysis of TSyF spectra

As it is seen from the plot of selected wavenumbers, the wavenumbers which assign the differences in the absorbance values were chosen by the genetic algorithm. The overtones and the combinations of CH, CC, and OH groups have the contribution on the classification of olive oil samples. It can be concluded as GADA is selectable analysis techniques compare to the SIMCA analysis when the spectral information is sued as data matrix in the classification procedure.

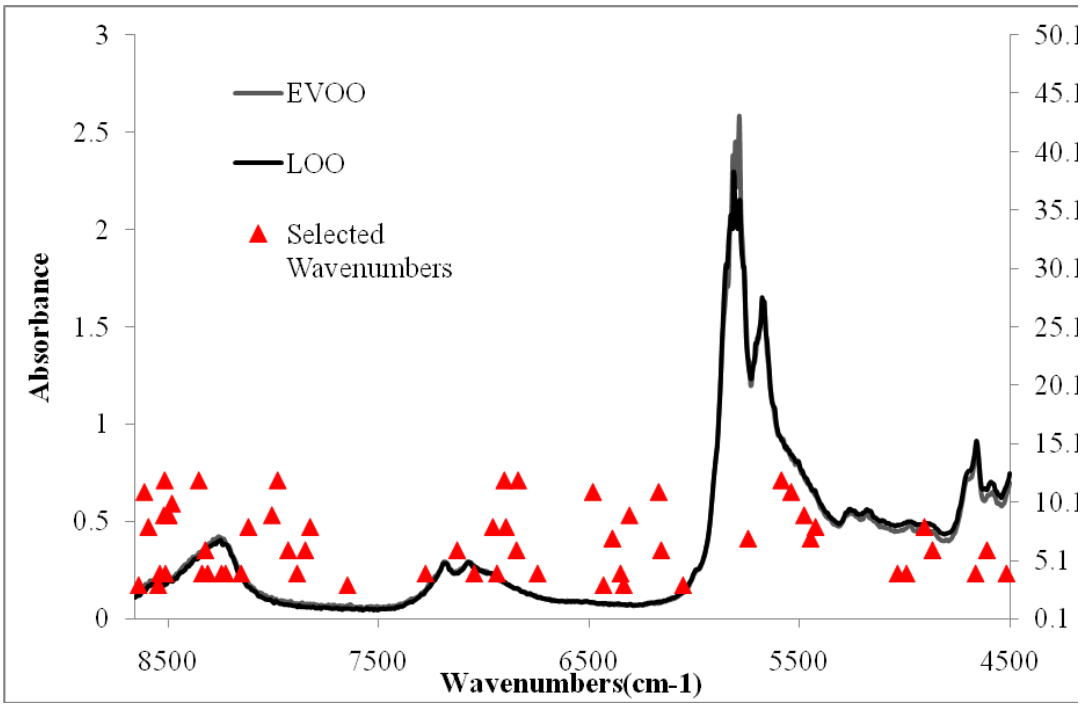


Figure 7.12. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis

7.2.3.2. Classification Results of Refined Olive Oils and Lampante Olive Oils

The spectral data matrix was constructed as training set and test set that contains totally 38 and 12 olive oil samples, respectively. Each sample has 3371 of wavenumbers in the spectrum. Before starting the classification procedure, autoscaling was applied to the spectral data matrices as a preprocessing technique. SIMCA analysis firstly was examined and then the results of the both methods were compared to achieve the success of GADA analysis.

In SIMCA analysis, refined and lampante olive oil samples was predefined as class 1 and class2, respectively. Nine principal components were found out for both classes with 96.1% and 98.00% of explained variance. In order to construct the Cooman's plot, the vertical and horizontal limits were calculated as 1.38 for each class at 95% confidence level. Figure 7.13 shows the distribution of olive oil in the space. The refined olive oil samples of training set which are ROO-5, 9, 15, 19, are classified as belonging to the both classes. Also the samples of test set coded as ROO-2 and ROO-

5 belong to the same class. Only LOO-5 of test set is classified as not belonging to the any classes.

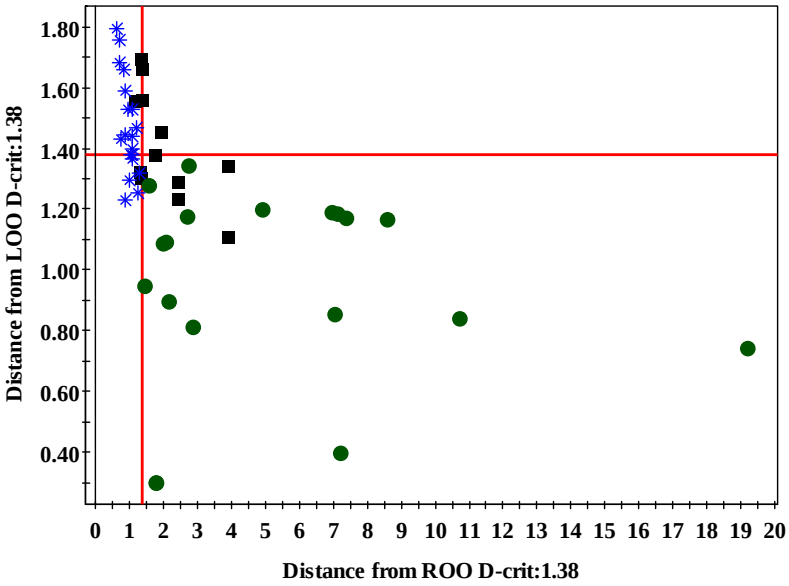


Figure 7.13. Cooman's plot of olive oil samples obtained from SIMCA analysis of TSyF spectra (star: ROO-training, circle: LOO-training, box: test set).

GADA analysis was initiated with 6 genes and 10 iteration numbers. At the end of the analysis, sixteen principal components were found out with a 90.85% of explained variance. The Cooman's plot was drawn to visualize the distribution of the refined and lampante olive oil samples (Figure 7.14). The critical limits for each class were calculated as 10.18 at 95% confidence level. These limits show that there are two separate classes for olive oil samples. Only LOO-3 existing in the test set is classified as not belonging any of the classes and also LOO-5 can be expected as not belonging to any classes.

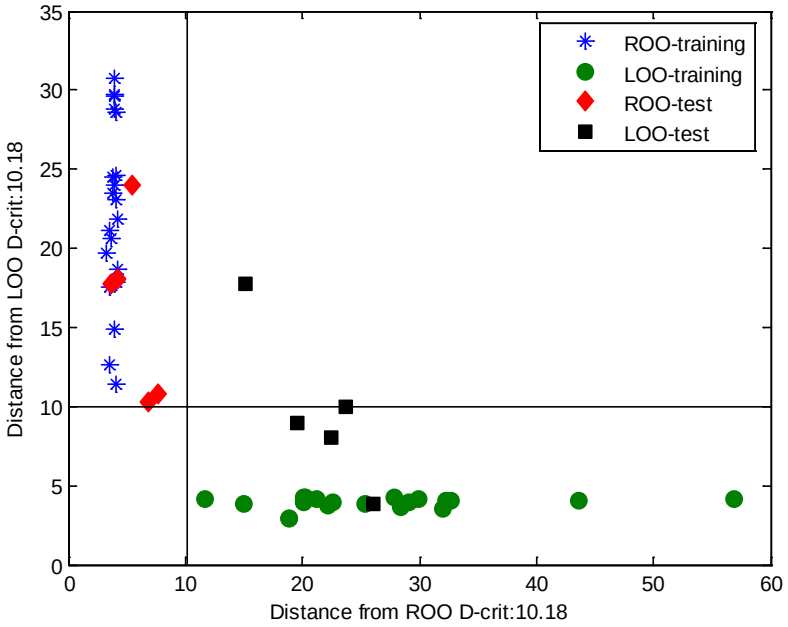


Figure 7.14. Cooman's plot of olive oil samples obtained from GADA analysis of TSyF spectra

GADA analysis is based on the wavelengths selection that contains the necessary information for the classification. Totally forty six wavenumbers with their corresponding absorbance values were selected for the classification of olive oil samples. These wavenumbers were plotted against the wavenumber region of NIR in order to see the selected wavenumbers visually. As it is seen from the Figure 7.15 the wavenumbers which shows different intensity in the absorbance values are the most selected ones. And also these means that these have the most contribution on the classification of olive oil samples.

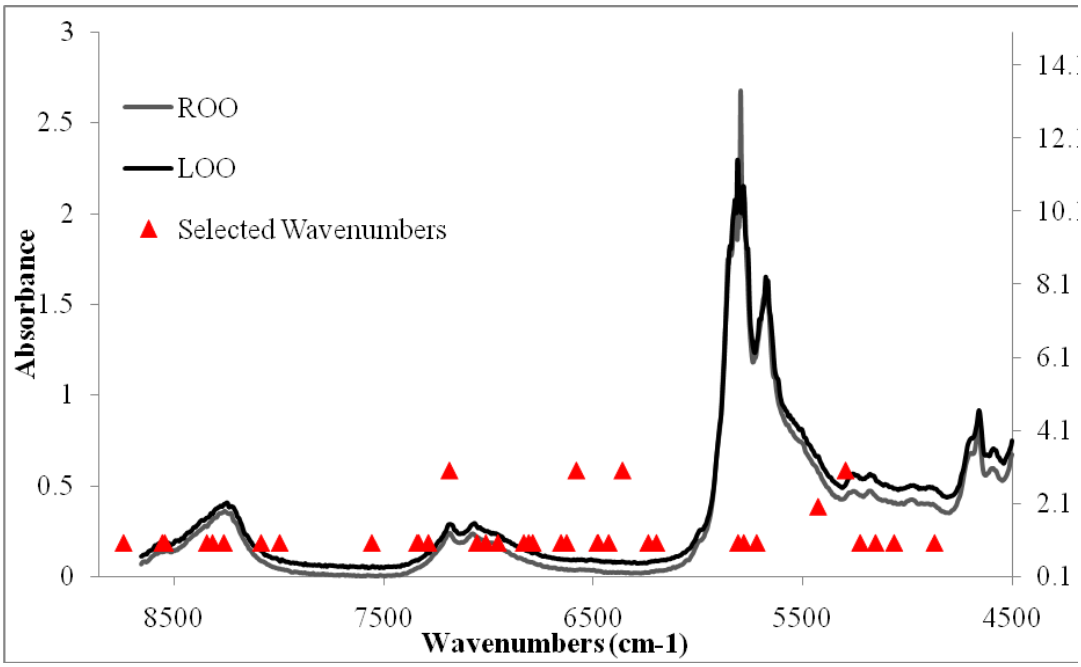
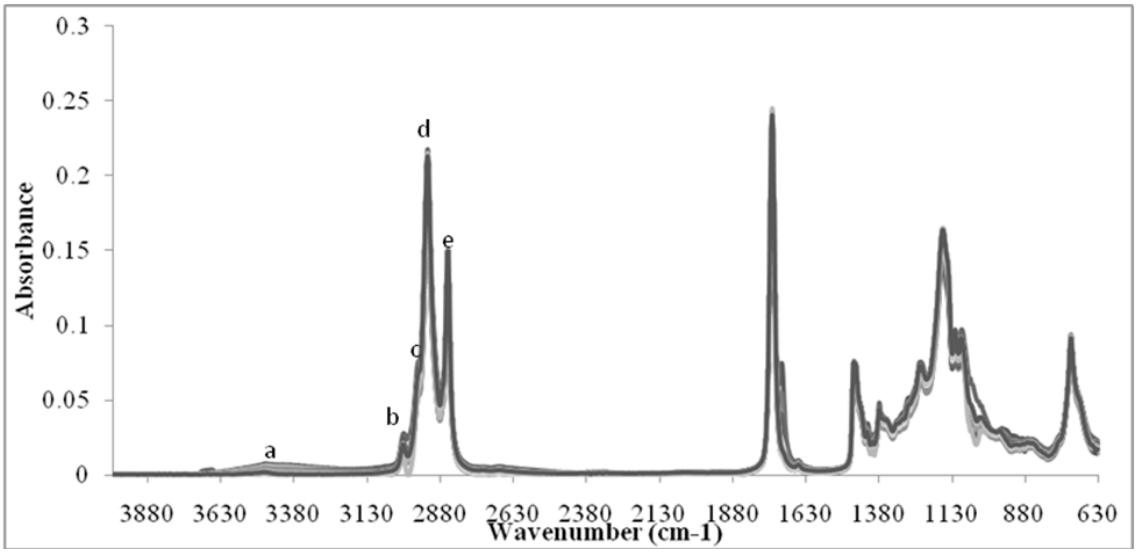


Figure 7.15. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis

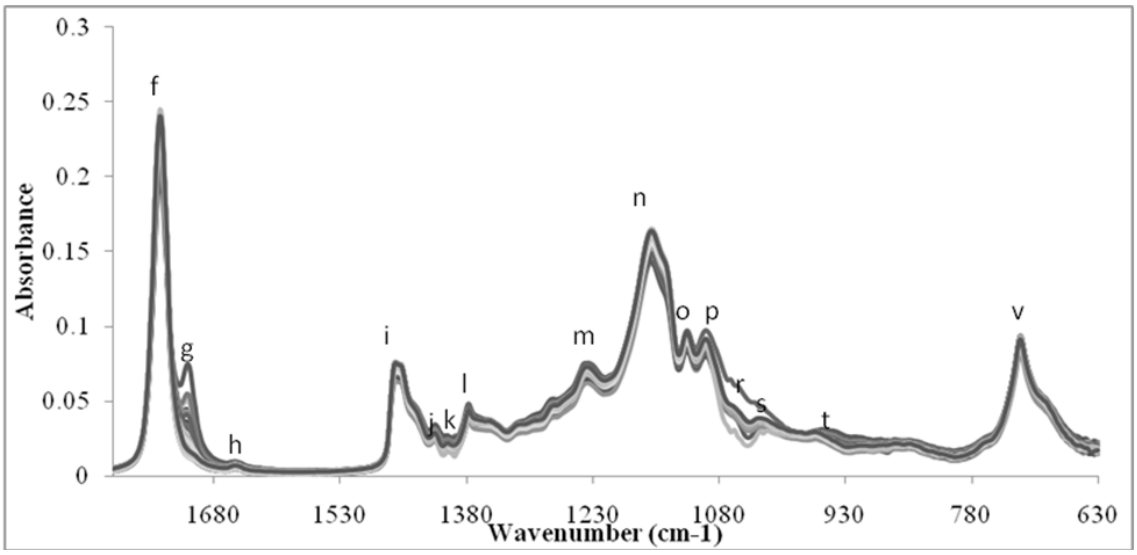
7.3. FTIR Results

7.3.1. FTIR Measurements of Olive Oil Samples

All the edible olive oils (extra virgin olive oils, EVOO and refined olive oils, ROO) and lampante olive oils (LOO) were measured using FTIR spectrometer equipped with attenuated total reflectance accessory attached diamond-ZnSe crystal. Due to the working range of the ATR crystal, spectra were taken in the range 4000–630 cm^{-1}. As it is seen from the spectra of olive oil samples (Figure 7.16), there are several infrared peaks including overlapping, since the chemical composition of olive oils includes basically fatty triglycerides esters with different structure, length, saturation of chains and minor components. Each infrared peak was coded by a letter in order to explain the corresponding functional groups.



(a)



(b)

Figure 7.16. FTIR spectra of olive oil samples measured in the range of a) 630 – 4000 cm^{-1}, b) 630 – 1800 cm^{-1} using ATR accessory attached diamond ZnSe crystal.

Table 7.4. Evaluation of FTIR spectrum
(Source: Guillén and Cabo 1997, Vlachos, et al. 2006)

Coded Infrared Peak	Wavenumber (cm^{-1})	Corresponding Functional Group
a	3470	overtone of the glyceride ester carbonyl absorption
b	3009	CH stretching of =CH (cis and trans double band)
c	2960	symmetric and asymmetric vibration of aliphatic CH_3 groups (seen as shoulder)
d	2925	asymmetric stretching of aliphatic CH_2 groups
e	2854	symmetric stretching of aliphatic CH_2 groups
f	1745	stretching of ester carbonyl functional group of triglycerides (C=O)
g	1710	acid group of free fatty acids (seen as shoulder)
h	1655	C=C stretching vibration of olefins
i	1460	bending vibration of CH_2 and CH_3 aliphatic groups
j	1418	rocking vibrations of CH bonds of cis-disubstituted olefins
k	1396	bending in plane of CH bonds of cis-disubstituted olefins
l	1379	symmetrical bending vibration of CH_2 bonds of cis-disubstituted olefins
m	1241	stretching vibration of C-O ester
n	1161	stretching vibration of C-O ester
o	1120	stretching vibration of C-O ester
p	1100	stretching vibration of C-O ester
r	1053	stretching vibration of C-O ester
s	1033	stretching vibration of C-O ester, asymmetric vibrations of C-C(=O)-O
t	950	bending vibration of out-of-plane of trans disubstituted olefinic groups.
v	723	methylene rocking vibration and out-of-plane bending vibration of *cis*-disubstituted olefins

In order to investigate the power of FTIR spectroscopy in the classification of olive oils, FTIR spectral results were examined in two different algorithms. Firstly, the singular value decomposition based principal component analysis (SVD-PCA) was performed using FTIR spectral data matrix. The same procedure was repeated using distance based genetic algorithm principal component analysis (GAPCAD). SVD-PCA generally is used to get prior information about the samples, whereas GAPCAD was designed as supervised classification methods. The main idea of combining the genetic algorithms and principal component analysis is only to use the wavenumbers or wavelengths which are included necessary information for the classification of samples. As it known, in supervised classification training set is used to define the rule of classification then this rule is tested by an independent sample set.

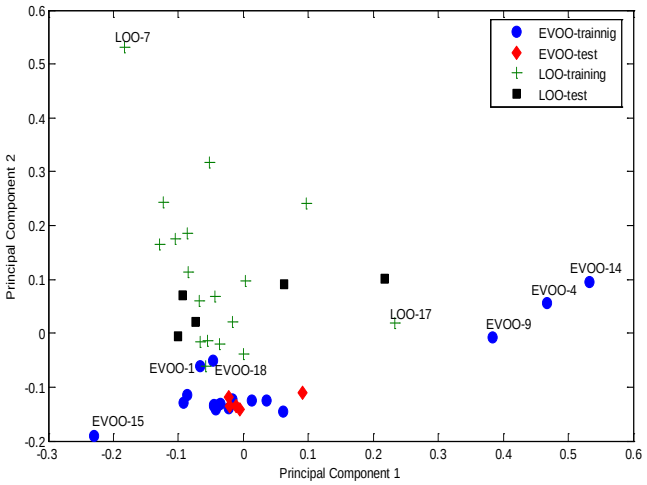
7.3.2. Classification Results of SVD-PCA and GAPCAD

7.3.2.1. Classification Results of Extra Virgin Olive Oils and Lampante Olive Oils

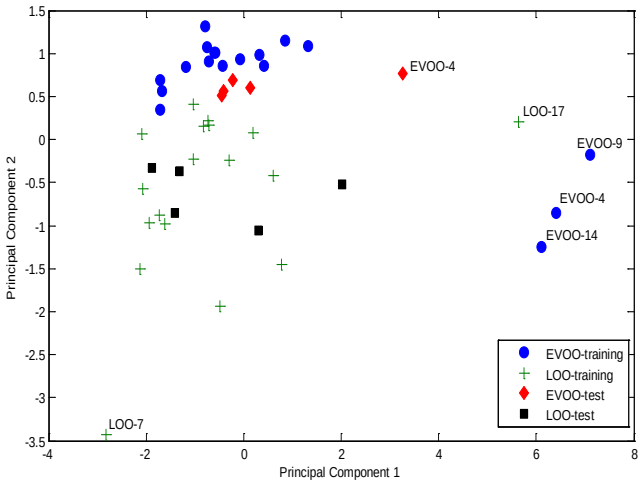
The FTIR spectral data contain totally 46 of olive oil samples. In the examination of SVD-PCA, all of them used to get the prior information about the classes of olive oil samples. On the other side, these olive oil samples divided into two samples set named as training set and test set. Training set includes totally 36 olive oil samples in which contain 18 of extra virgin olive oils (EVOO) and the remaining is lampante olive oil (LOO) samples. Independent sample set has totally 10 of olive oil samples that are 5 extra virgin olive oil samples and 5 lampante olive oils. Before running the algorithms, different pre-processing methods were applied to the spectral data matrix. The used pre-processing techniques were mean-centering and autoscaling.

Ten significant principal components were obtained with an 87.01% of explained variance at the end of the SVD-PCA calculation. The scores of first two principal components (with a 55.37% of explained variance) were plotted against each other in order to see the distribution of olive oil samples. The expected results from olive oil spectral data was two different classes were contained the extra virgin olive oils and lampante olive oils, respectively. However, as it is seen from the Figure 7.17.a., there is a mainly two different classes with some outliers. EVOO 4, 9, 14 samples are totally different from the other olive oil samples. On the other side, EVOO 1, and 18

samples have same chemical or organoleptic properties of lampante olive oil samples. Also EVOO 2, and 3 show similarities just like lampante olive oil samples. Lampante olive oil samples are also scattered in larger area than the extra virgin olive oil has done. The main reason of the larger scattering of LOO is to show acidities in a larger scale, since the LOO samples have the acidity values more than 3.3% (w/w) in 100 grams of olive oil. Also the LOO 7, and 17 show different properties than the other lampante olive oils.



(a)



(b)

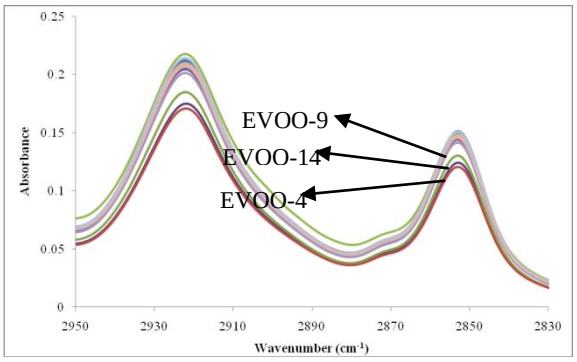
Figure 7.17. Score plot of principal components calculated from FTIR spectral data matrix of olive oil samples using a) SVD-PCA and b) GAPCAD for whole spectral data (ν = 4000–630 cm^{-1}).

GAPCAD algorithm was initiated with 100 genes and 100 iteration numbers. Six significant principal components were found out with a 96.19% explained variance. The score values of first two principal components (with a 68.01% of explained variance) were used to obtain the scattering of olive oil samples in the space. Figure 7.17.b. shows the classes of olive oil samples in the space. Again the EVOO samples numbered as 4, 9, and 14 show different properties than the other olive oil samples. Also LOO 17 has similar characteristics with EVOO samples 4, 9, and 14.

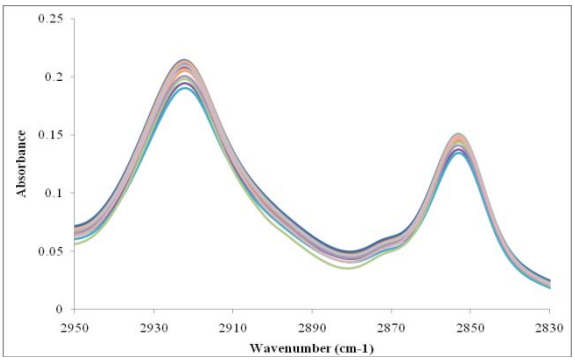
In order to investigate the reasons of different places in the space of these olive oil samples, FTIR spectra of olive oil samples were plotted against the wavenumber in different regions. As it is seen from the Figure 7.18.a, Figure 7.18.c, Figure 7.18.d, and Figure 7.18.f.; EVOO 4, 9, and 14 shows different intensities in the region of 2950–2830 cm^{-1}, and 1780–1700 cm^{-1}, respectively. In the first region, the symmetric and asymmetric vibrations of aliphatic CH_2 groups are exhibited and the vibrations of aliphatic CH_3 groups are also shows FTIR peaks as a shoulder. These aliphatic CH_2 and CH_3 groups assign the fatty acid chains. In the center of spectrum which is the region of 1780–1700 cm^{-1}, stretching vibration of C=O group of the triglyceride ester linkage is present. The region started at 3100 cm^{-1} and ended at 3600 cm^{-1} contains the vibrations of water (H–OH) and hydroperoxides (ROOH) and alcohols (ROH). Vlachos et al. studied on oxidation process of oils in the presence of heat and UV light. They proved that the spectral bands between the 3600–2700 cm^{-1} and around 1750 cm^{-1} have some changes. Since heat and UV light caused the production of saturated aldehydes functional groups and the secondary oxidation products. It can be concluded as the reasons of differentiation in the olive oil samples.

The loading plot of principal components were also plotted against the wavenumbers in order to observe which middle infrared region have the most contribution in the classification of extra virgin and lampante olive oil samples. In the score plot of principal components found out at the end of SVD-PCA process, the olive oil samples are mostly scattered mainly along principal component 2 (PC2). The infrared regions between the 3050–2800 cm^{-1}, the band around 1750 cm^{-1} show the most intense contributions on the classification of olive oil samples (Figure 7.19.a.). On the other side, the EVOO and LOO samples which samples have the different characteristics are mainly along on the positive scores of principal component 1 (PC1). According the loading plot PC1, the vibration of C=O has the largest and intense contribution (Figure 7.19.b.). PCA generally uses whole spectrum in order to define the

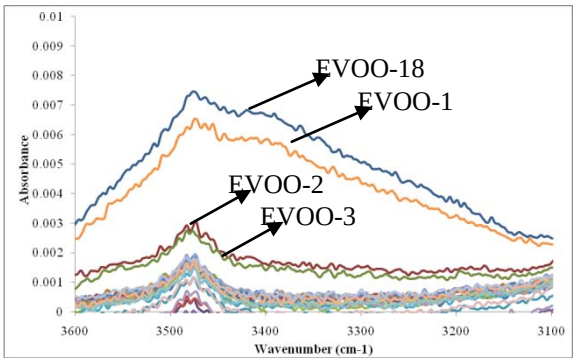
classes of samples, on the other hand GAPCAD try to find out the wavenumbers include the necessary information. The wavenumbers 3989, 3590, 2628, 2431, 2193, 2120, 1473, and 1130 cm^{-1} have the contribution in the classification of olive oil samples using GAPCAD. In these wavenumbers generally the lampante olive oil samples show differences.



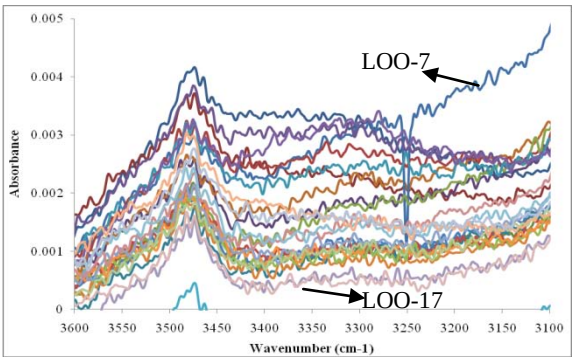
(a) FTIR spectra of EVOO in the range of 2950–2830 cm^{-1}.



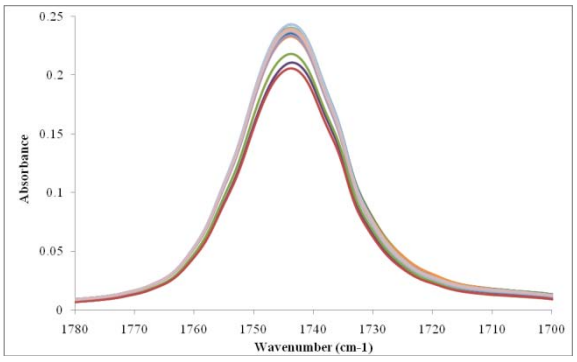
(d) FTIR spectra of LOO in the range of 2950–2830 cm^{-1}.



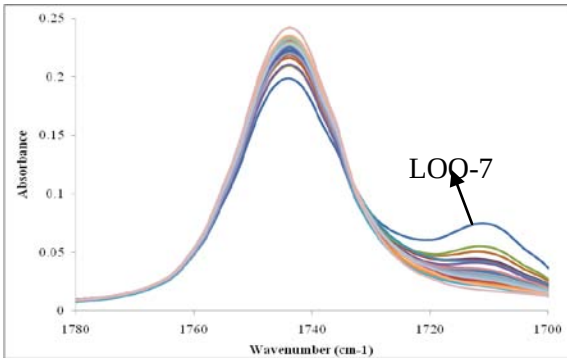
(b) FTIR spectra of EVOO in the range of 3600–3100 cm^{-1}



(e) FTIR spectra of LOO in the range of 3600–3100 cm^{-1}

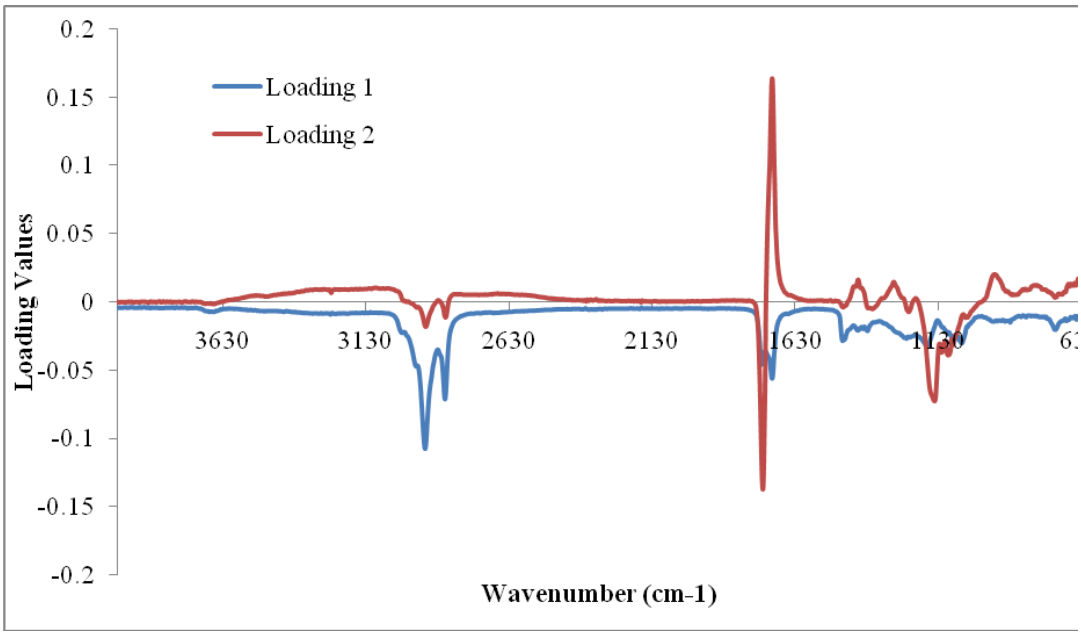


(c) FTIR spectra of EVOO in the range of 1780–1700 cm^{-1}

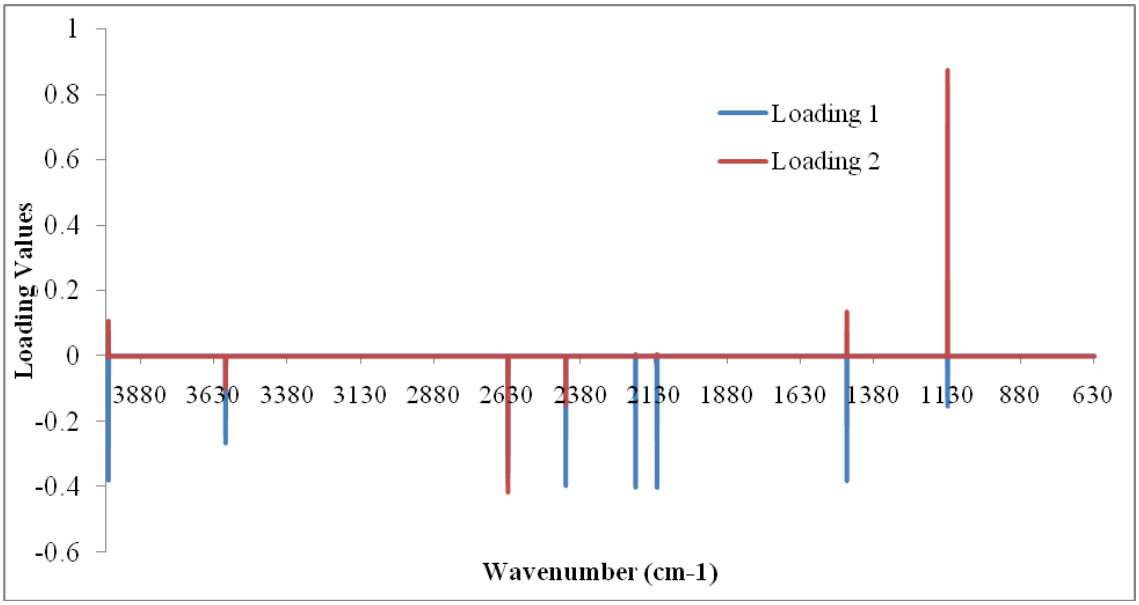


(f) FTIR spectra of LOO in the range of 1780–1700 cm^{-1}

Figure 7.18. FTIR spectra of olive oil samples in different regions.



(a)



(b)

Figure 7.19. Loading plots of FTIR spectral data of the olive oil samples (mean-centered data) obtained from the calculation of a) SVD-PCA, b) GAPCAD.

Due to the nature of the genetic algorithms, GAPCAD is based on a random design. Therefore, at each run different results will be obtained. To investigate which region or wavenumbers have the larger information about the classification of olive oil

samples, GAPCAD was run 100 times and the frequency of selected wavelengths was plotted against the wavenumbers.

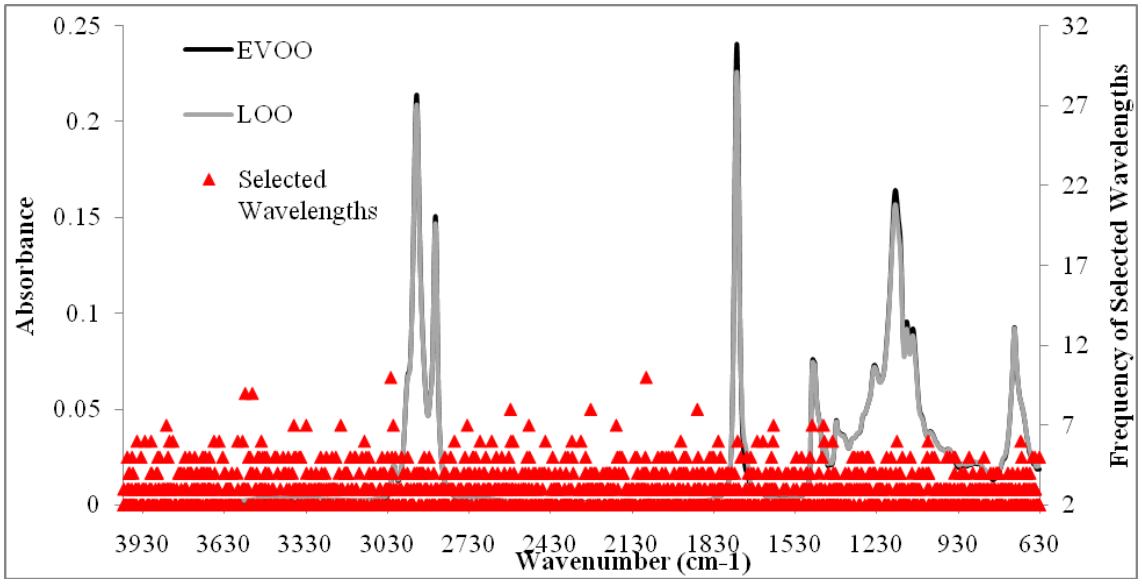


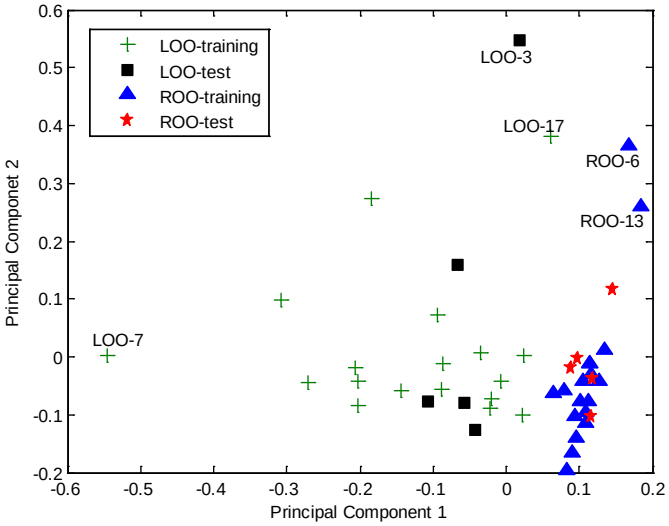
Figure 7.20. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of FTIR spectral data.

According to the frequency of selected wavenumbers, the vibration of ROOH and ROH groups are produced after the oxidation of olive oils take place in the classification olive oil samples. The spectral bands of LOO samples around 2130 cm^{-1} show differences unlike EVOO samples. The minor components exist in olive oils shows vibrations in the fingerprint region. They are also important in the classification olive oil samples.

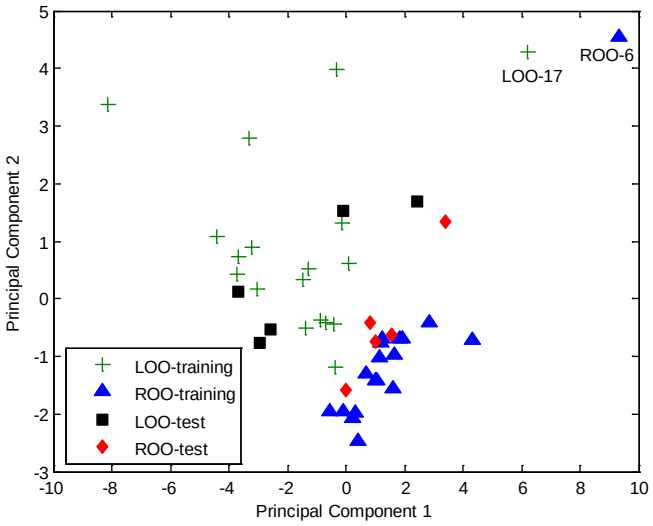
7.3.2.2. Classification Results of Refined Virgin Olive Oils and Lampante Olive Oils

To examine the success of GAPCAD, the classification of refined olive oil and lampante olive oil samples were investigated. Firstly the FTIR spectral data matrix of olive oil samples with a size of 46 x 3371(sample number x number of wavenumbers) was examined using SVD-PCA algorithm. The sample set containing 23 of refined olive oil and 23 of lampante olive oil samples were used to observe the classification rule. Mean-centering again was used as a pre-processing technique. Nine significant principal components (with an 84.23% of explained variance) were found out after the calculation

of SVD-PCA. The score values of first two principal components with a 53.60% of explained variance were used to visualize classes of refined olive oil and lampante olive oil samples.



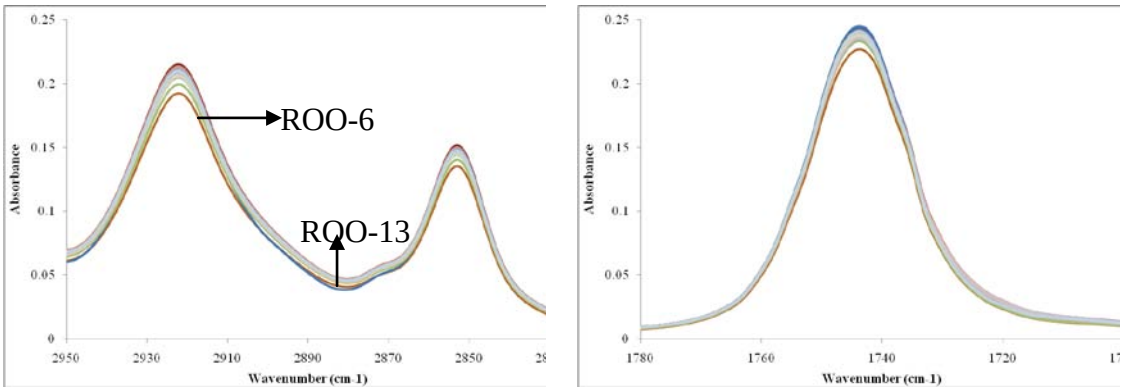
(a)



(b)

Figure 7.21. Score plot of principal components calculated from FTIR spectral data matrix of olive oil samples using a) SVD-PCA and b) GAPCAD for whole spectral data (ν = 4000–630 cm^{-1}).

This result proves that PC1 contains the most explained information about the classification of olive oil samples. LOO 3, 17 and ROO 6, 13 shows the different properties than the other olive oil samples. The FTIR spectra of ROO samples were investigated more detail and as it is seen from Figure 7.22.a. and Figure 7.22.b., ROO samples numbered as 6, 13 show different intensities than the other refined olive oil samples. The reasons of differences in lampante olive oil samples were explained above.



(a) FTIR spectra of ROO in the range of 2950–2830 cm^{-1}.

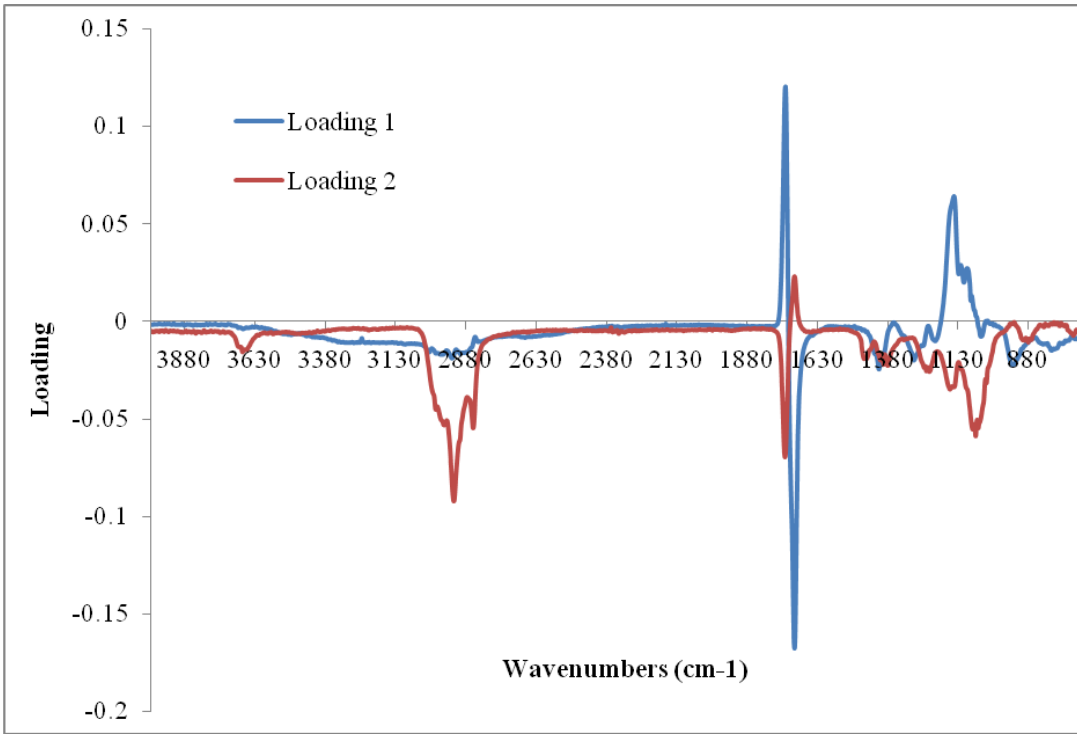
(b) FTIR spectra of ROO in the range of 1780–1700 cm^{-1}.

Figure 7.22. FTIR spectra of refined olive oil samples in different regions.

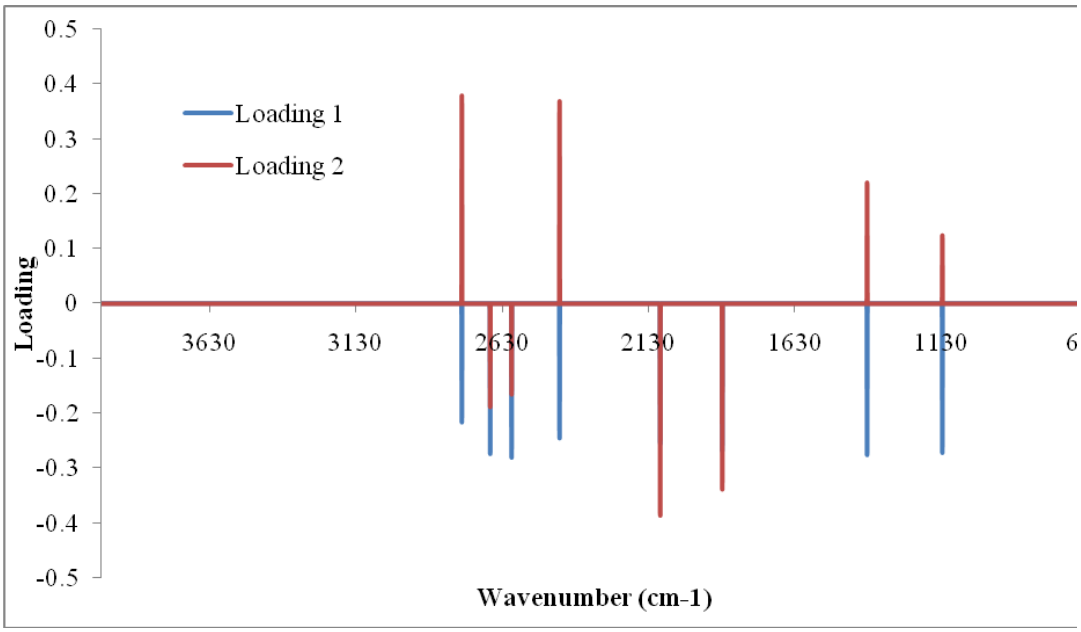
In the result of SVD-PCA, PC1 was the principal components that define the distribution of olive oil samples. As it is seen from the Figure 7.23.a., the region 3080 – 2780 cm^{-1} and the infrared peak of C=O group exists as a functional group in the fatty acid chain have the most contribution on the classification olive oil samples. The vibrations of minor components of olive oils have also effect in the classification olive oils, since these regions have the most intense peak area in the FTIR spectrum.

GAPCAD algorithm was initiated with 100 genes and 100 iteration numbers. Six significant principal components (93.45% of explained variance) were calculated at the end of the process. The first two principal components with a 64.95% of explained variance were plotted against each other in order to visualize the distribution of olive oil samples in the space. In the scatter plot of the scores values of PC1 vs. PC2 show nearly

equal contribution on the classification of refined olive oil and lampante olive oil samples.



(a)



(b)

Figure 7.23. Loading plots of FTIR spectral data of the olive oil samples (mean-centered data) obtained from the calculation of a) SVD-PCA, b) GAPCAD.

It can be easily concluded that the positive scores of PC1 and the negative scores of PC2 assign the distribution of refined olive oil samples, whereas the opposite of scores of principal components appoint the lampante olive oil samples. In the classification of ROO and LOO samples using GAPCAD, GA was selected nine wavenumbers with their corresponding absorbance values. These wavenumbers were 656, 1127, 1384, 1879, 2091, 2435, 2599, 2672, 2769 cm^{-1} and the score values of principal components were calculated according to these absorbance values. As it seen from the Figure 7.23.b., the wavenumbers which are 1879, 2091, 2435, 2599, 2672, 2769 cm^{-1} have the most contribution in the classification.

In order to have a better idea about the distribution of wavenumbers or region which have the most contribution on the classification, GAPCAD was run 100 times and the frequency of selected wavelengths was scattered against the wavenumbers. Figure 7.24 is a scatter plot of frequency of selected wavelengths. According to this plot, the region of aliphatic CH_n groups, the vibrations of oxidation products are existed in the 3600–3100 cm^{-1} are the mostly used ones in the classification. On the other hand the baseline in the region of 2600–2100 cm^{-1} also effects the classification.

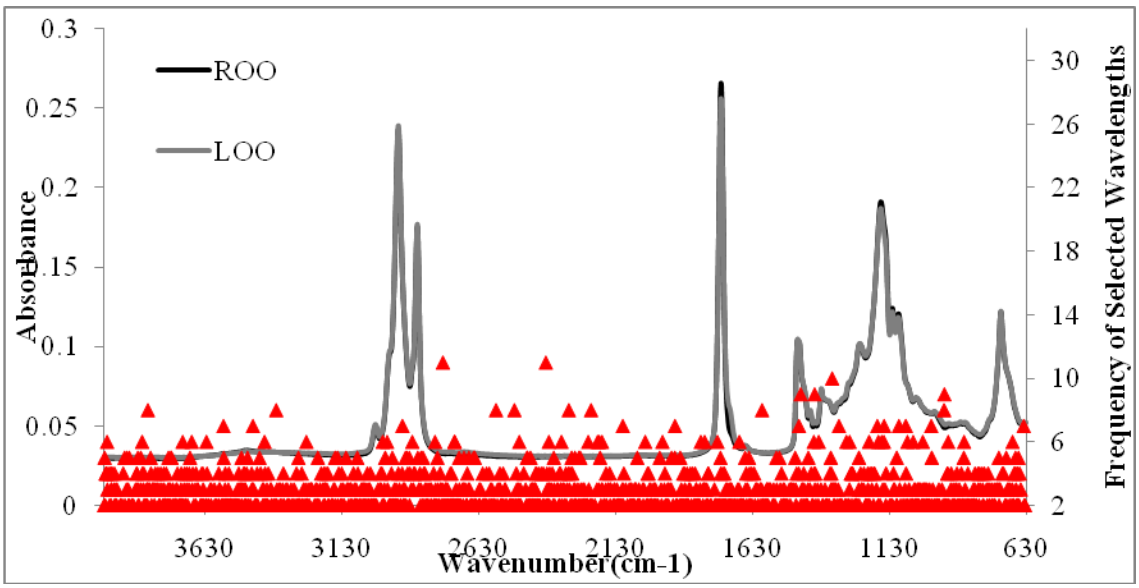


Figure 7.24. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of FTIR spectral data.

7.3.3. Classification Results of SIMCA and GADA

7.3.3.1. Classification Results of Extra Virgin Olive Oils and Lampante Olive Oils

The FTIR spectral data matrix was designed as including both extra virgin olive oil and lampante olive with their corresponding FTIR spectra. The observed spectral data matrix has 46 x 3371 (samples x wavenumbers) dimensions. 10 of them assign the test set of the olive oil samples, whereas the remaining are constructed the training set. Both sample sets include same numbers of extra virgin and lampante olive oil samples. The autoscaling was applied to the both training set and test set before starting the examination of SIMCA and GADA analysis.

SIMCA analysis found out eight principal components (with a 97.70% and a 98.20 % of explained variance for class of extra virgin olive and class of lampante olive oil samples, respectively) for both classes. The Cooman's plot was constructed by plotting the distances of extra virgin olive against the distances of lampante olive oils. The boundaries of each class were calculated as 1.35 at 95% confidence level.

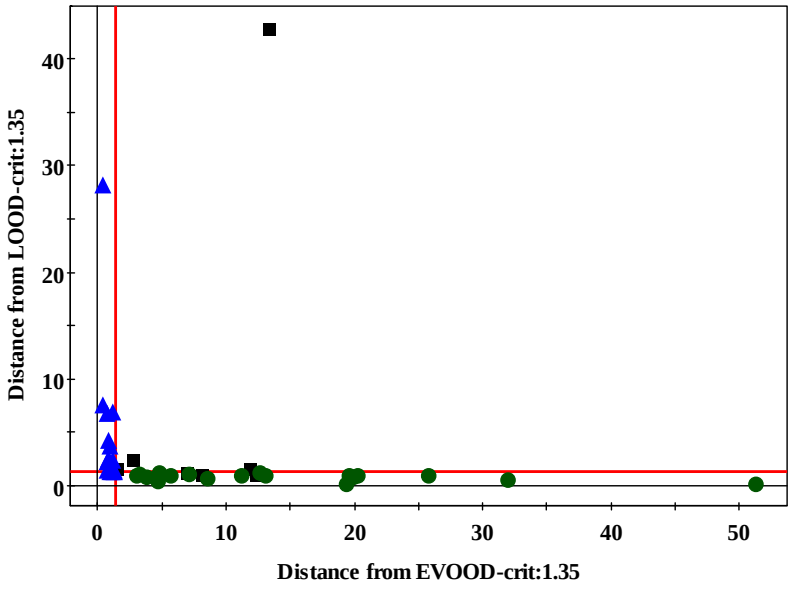


Figure 7.25. Cooman's plot of olive oil samples obtained from SIMCA analysis of FTIR spectra (triangle: EVOO-training, circle: LOO-training, box: test set).

According to the Cooman's plot (Figure 7.25), EVOO-3, 6, and 13 existing in the training set, and also EVOO-2, EVOO-4 existing in the test set are classified as belonging to both classes. O the other side, EVOO-1, EVOO-3, LOO-3, and LOO-4 existing in the test set are classified as not belonging to any classes.

GADA analysis was initiated with 6 genes and 10 iteration numbers. At the end of the analysis fifteen significant principal component analyses were found out with a 90.68% of explained variance. The Cooman's plot was constructed by using the distances between the extra virgin olive oil and lampante olive oil samples (). The critical limits were calculated as 5.11 at 95% confidence level. According to these boundaries, there are not any samples defined as belonging to the both classes. This is an improvement for the GADA analysis when the results of both techniques are compared Only EVOO-5, LOO-3, 4, and 5 existing in the test set are classified as not belonging to any classes.

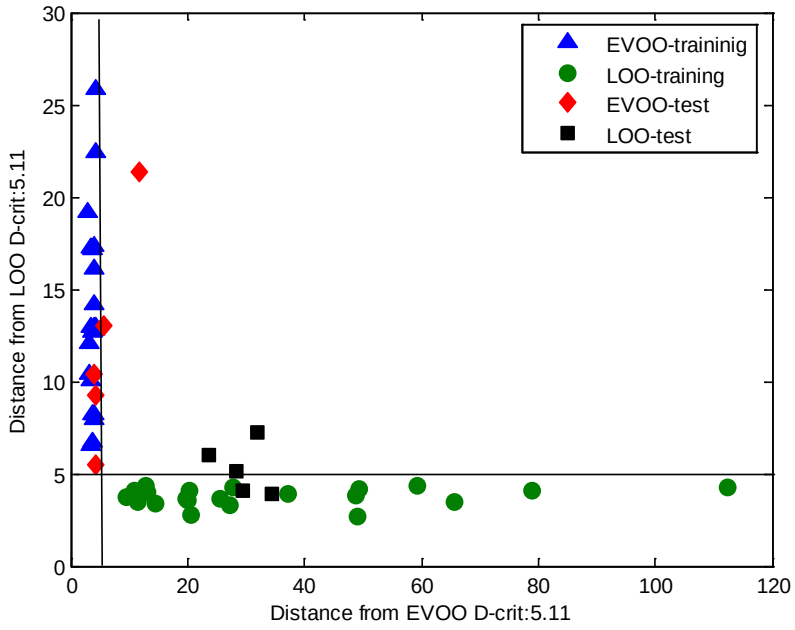


Figure 7.26. Cooman's plot of olive oil samples obtained from GADA analysis of FTIR spectra

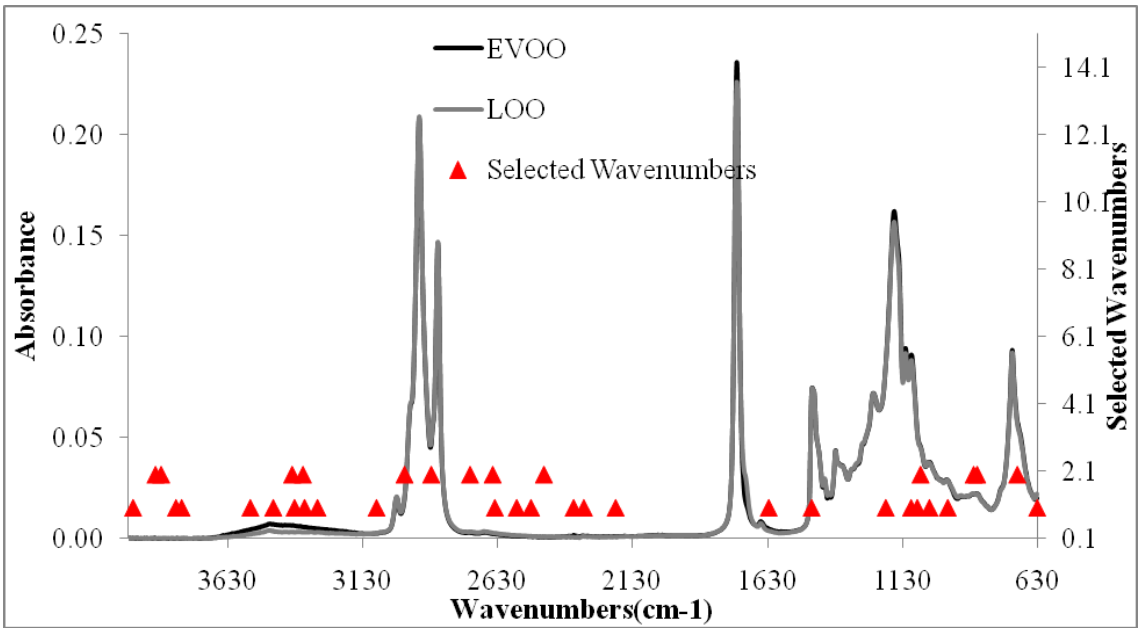


Figure 7.27. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis

The selected wavelengths or wavenumbers that were used in the classification are shown in Figure 7.27. The wavenumbers are existed in the fingerprint region of middle infrared were used in the classification. Since this region assigns the vibration of CO groups of esters and the olefins. The stretching vibrations of CH_2 and CH_3 groups are also used in the classification of olive oil samples.

7.3.3.2. Classification Results of Refined Olive Oils and Lampante Olive Oils

The classification of refined olive oil and lampante olive oil samples were examined using SIMCA and GADA analysis. The FTIR spectral data matrix was designed as including 36 samples of olive oil in the training set and 10 samples in the test set. The designed spectral data matrices have 36 x 3371 for training set and 10 x 3371 for test set. Both matrices were autoscaled before the classification procedure was started.

SIMCA analysis was performed after predefinition of each classes and test set. At the end of the analysis eight principal components (a 98.80% and a 98.40% of explained variances for the classes of refined olive and lampante olive oil samples, respectively) were found out for both classes. The critical limits were calculated as 1.35 for both vertical and horizontal lines at 95% confidence level. According to the

Cooman's plot (Figure 7.28) the samples of LOO-4 and LOO-14 are classified as belonging to the both classes whereas the samples ROO-3, ROO-4 existing in the test set and the test samples of lampante olive oils coded as LOO-1 and LOO-5 are classified as not belonging to the any classes.

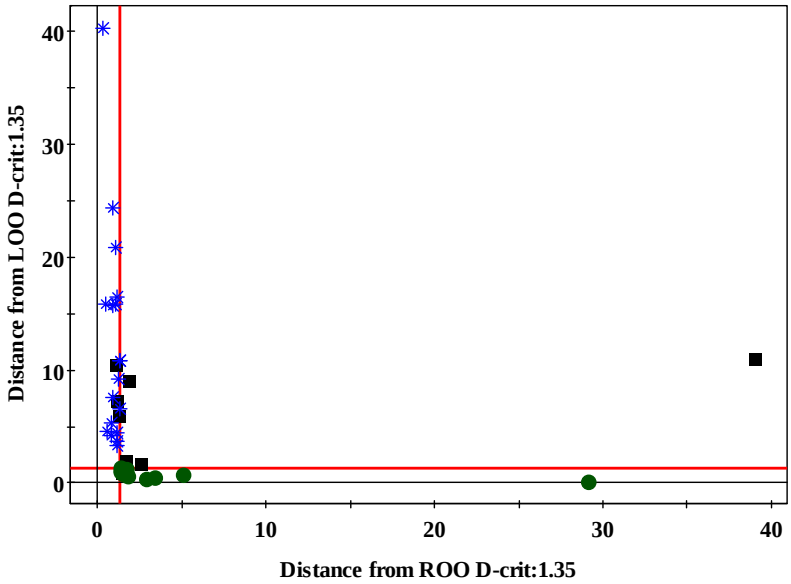


Figure 7.28. Cooman's plot of olive oil samples obtained from SIMCA analysis of FTIR spectra (star: ROO-training, circle: LOO-training, box: test set).

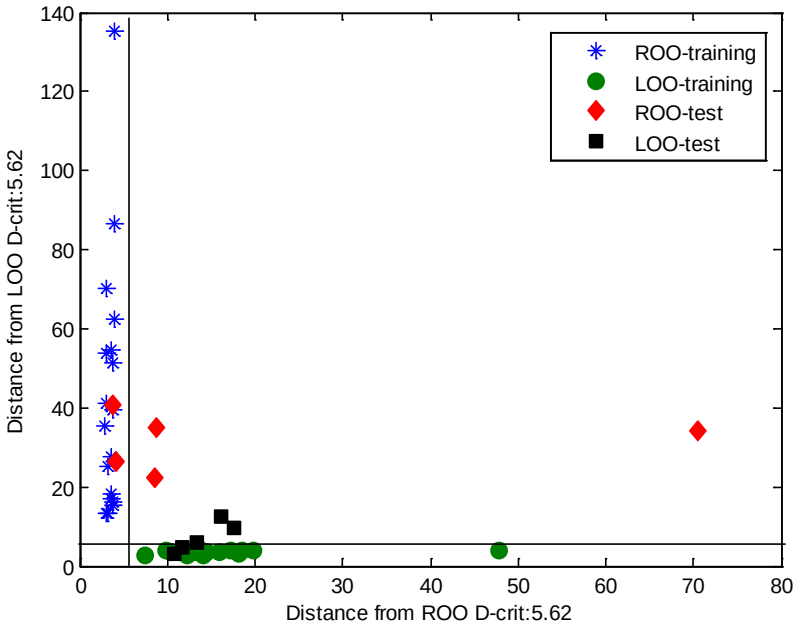


Figure 7.29. Cooman's plot of olive oil samples obtained from GADA analysis of FTIR spectra

GADA analysis was initiated with 6 genes and 10 iteration numbers. Totally thirteen principal components (90.80% of explained variance) were found out at the end of the analysis. The critical limits were calculated as 5.62 at 95% confidence level and these limits were used to separate the olive oil classes. The Cooman's plot was drawn to observe the distribution of olive oil samples in the space (Figure 7.29). The LOO-1 and LOO-5, ROO-3, 4, and 5 existing in the test set were found as not belonging to any classes predefined as refined olive oil and lampante olive. The same result was also obtained from the SIMCA analysis. This proves that FTIR spectra of these olive oil samples are different than the other olive oil samples. These samples can be expected as outliers.

The wavenumbers that were selected by genetic algorithms in the GADA analysis are shown in Figure 7.30. Totally sixty two of 3371 wavenumbers were used in the classification of olive oil samples. As it is seen from the plot of almost all selected the wavenumbers that have differentiation in the absorbance value have contribution on the identification of similarities or dissimilarities of the olive oils.

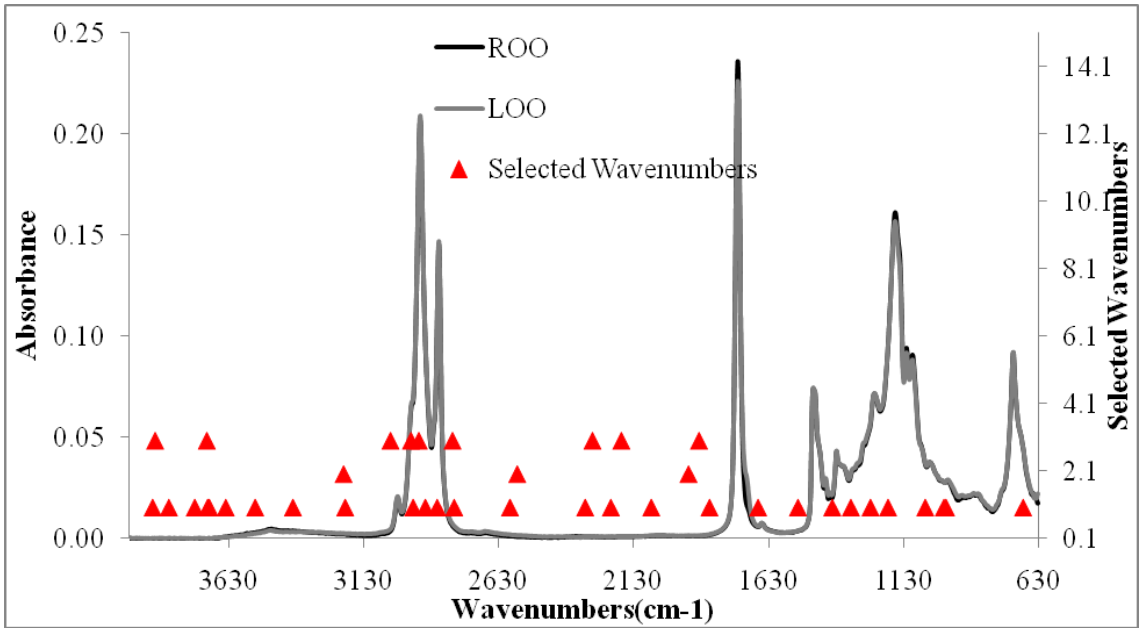
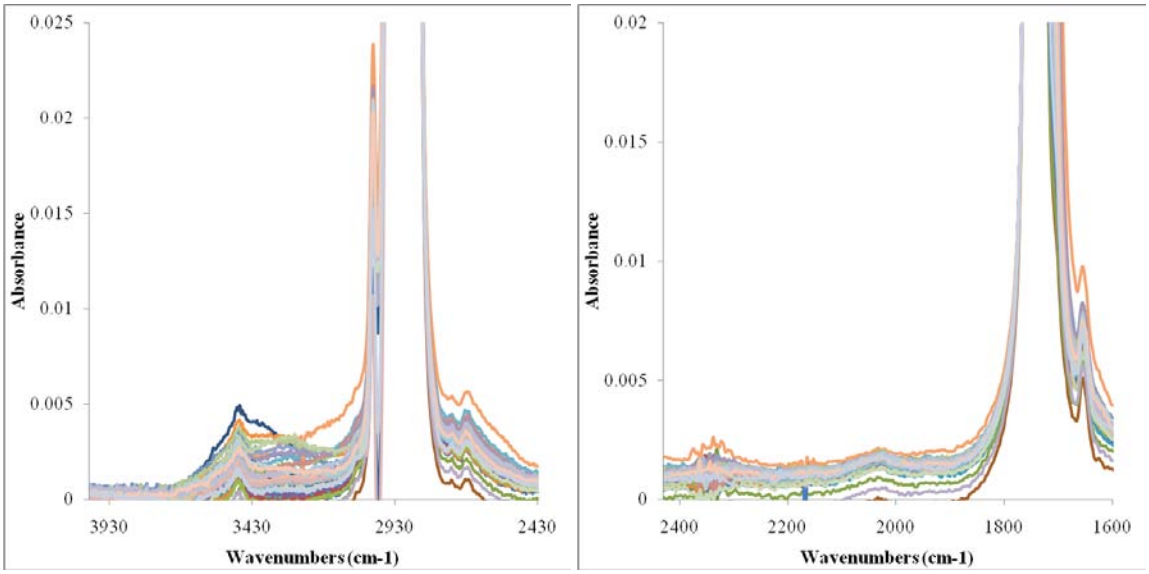


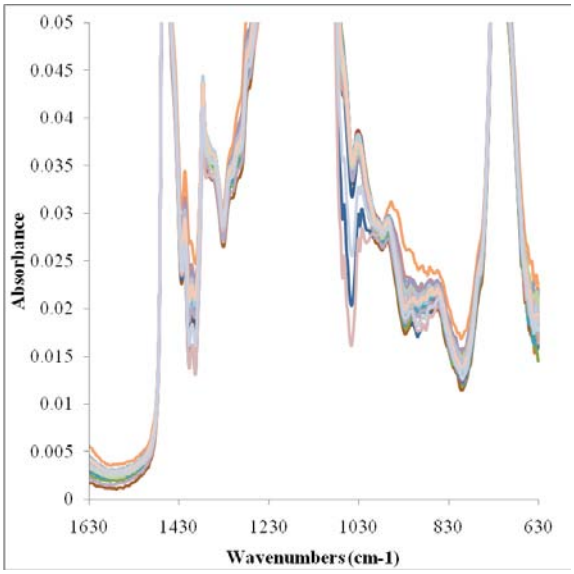
Figure 7.30. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis

To see the differentiation in the level absorbance values, the selected regions were plotted in a smaller scale (Figure 7.31). As it is seen from the Figure 7.31, the

intensity of peak height is based on the quality of olive oil. Genetic algorithm is selectable when spectral data matrix is used as an input.



(a) (b)



(c)

Figure 7.31. FTIR spectra of refined and lampante olive oil samples in different regions a) 4000 – 2430 cm^{-1}, b) 2430 – 1600 cm^{-1}, c) 1630 – 630 cm^{-1}.

7.4. Fluorescence Results

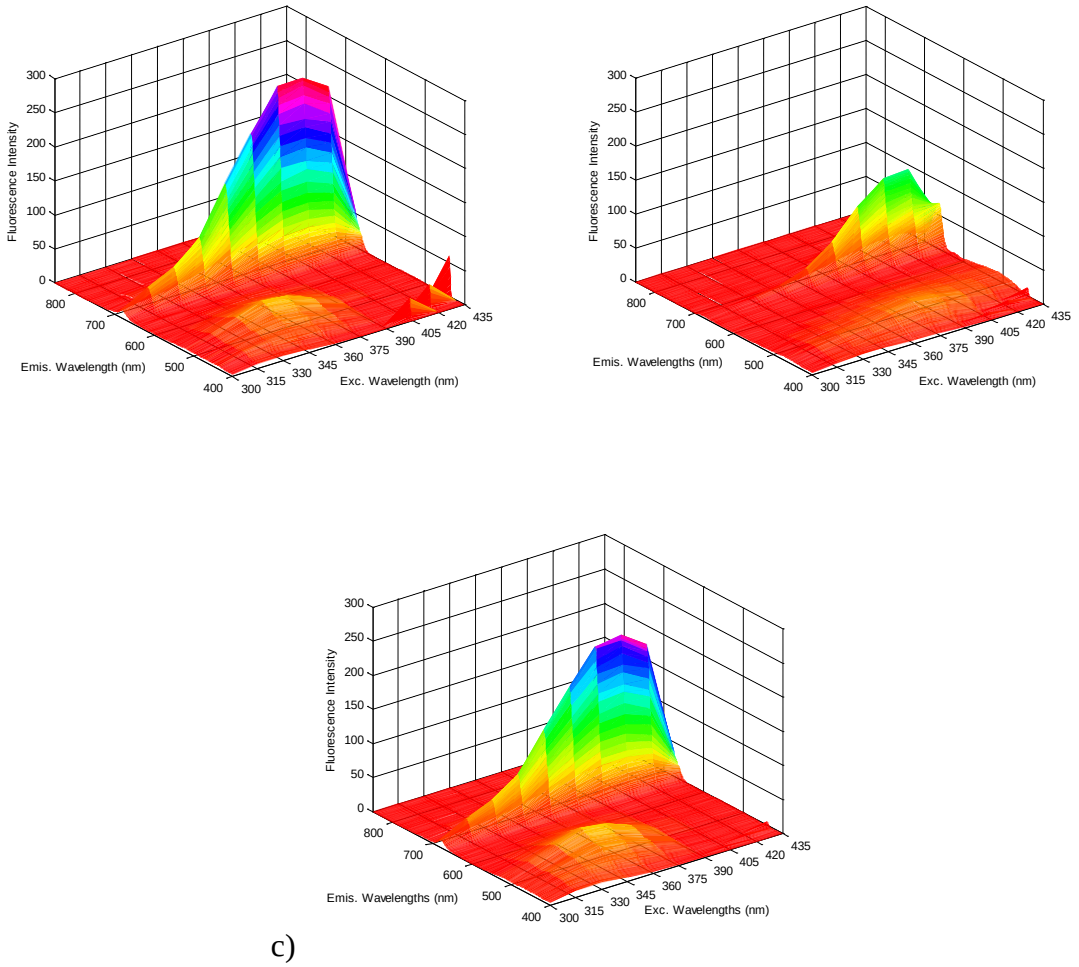
7.4.1. Excitation – Emission Fluorescence Results

7.4.1.1. Excitation – Emission Fluorescence Measurements of Olive Oil Samples

Kyriakids and Skarkalis studied on the fluorescence spectra measurement of olive oils and other vegetable oils. The emission spectra of olive oils were taken in the range of 400–700 nm at 365 nm excitation wavelength. In the fluorescence spectra of olive oils, four different fluorescence peaks were observed with a low intensity compared to the vegetable oils. Two medium peaks at 445 and 475 nm, one strong peak at 525 nm, and one weak peak to medium peak at 681 nm. Fluorescence at peak 680 nm was defined as typical of native chlorophyll. To proof that statement, the addition of chlorophyll was examined and the samples of virgin olive oil resulted in increase of the 680 nm peak. Also the heating experiment was performed and the fluorescence peak at 680 nm was decreased accompanied by a disappearance of green color of olive oil. As a consequent, due to the changes in the amount of chlorophyll, the chlorophyll acts as fluorescent quencher. Also another interesting result was observed from the heated olive oil sample. It was the increase of the intensity of peak at the region of 400–500 nm. The fluorescence peak at 525 nm appeared as a large peak on the spectra of virgin olive oil samples. At the beginning the reasons of appearance of this peak was connected to the parinaric acid, chlorophyll, and vitamin E. After some experiments, the peak was concluded as vitamin E, since the fluorescence spectra of pure vitamin E appeared as virgin olive oil samples. Lastly the peaks observed in the region 420–450 nm were obtained with different intensity among the virgin olive oils. The correlation studies of K_{270}, K_{232}, and acidity with the intensity of fluorescence peaks were examined. As it is known, these values (of K_{270}, K_{232}) are related to the oxidation state and hydrolysis products (acidity) of the oils. The peaks appeared in that region are proportional to these values. Finally the researchers indicated that the UV absorption at 270 and 232 nm were responsible for the intensity of fluorescence peak at 445 nm of good quality virgin olive oils that contains the monosaturated fatty acids and high content of phenolic antioxidants and could be partly due to its tocopherol contents. On the other hand, if this fluorescence peak was exhibited as a large peak, it could be due to the large percentages

of polyunsaturated fatty acids and their much higher percentages of oxidation products. (Kyriakids and Skarkalis 2000)

a) b)



c)

Figure 7.32. Excitation–emission fluorescence spectrum of a) extra virgin olive oil b) lampante olive oil c) refined olive oil between λ_{exc}=300–435 nm and λ_{em}=400–850 nm

In order to determine the whether the peaks in the fluorescent spectra of olive oils were related to the chlorophyll, phenolic antioxidants, vitamin E or oxidation products, as indicated above, the olive oil samples were purchased from TARİŞ, the Union of TARIS Olive and Olive Oil Co-operatives and their fluorescence spectra were measured in the range of 400–850 nm at 300–435 nm excitation wavelength. The fluorescence spectra of both extra virgin olive oil and lampante olive oil samples shows Rayleigh scattering which are appeared at the measuring emission wavelengths below

excitation wavelengths (Figure 7.32). Despite this scattering; extra virgin olive oil and lampante olive oil samples generally displayed the same pattern due the higher peak between the λ_{em}=650–700 nm at λ_{exc}=300–390 nm.

These intense peaks are the result of chlorophylls that are existed in the olive oils. In the lampante olive oil this peak appears much less intense than the extra virgin olive oils, the intensity difference of the chlorophyll peaks between the olive oil samples may be the results of differences at the values of K_{270}, K_{232} or the origin of the olive oil samples. There is also another less intense peak in the range of λ_{exc}=300–390 nm and λ_{em}=400-550 nm which is attributed to oxidation products (Figure 7.33.). This peak in the fluorescence spectrum of lampante olive oil shows a peak around 460 nm emission wavelength at 360–390 nm excitation wavelength whereas in the extra virgin olive oil shows much less intense peak. The peak at 460 nm refers the oxidation products or hydrolysis products that exist in the lampante olive oil samples. Due to stability of extra virgin olive oils to the oxidation this peak is less than lampante olive oil's peak. Extra virgin olive oils are quite stable to oxidation process due to their low fatty acid unsaturation and the high antioxidant activity of phenolic compounds and α-tocopherol (Vitamin E). Therefore the peaks around 520 nm can be attributed to Vitamin E (Guimet, et al. 2004, Guimet, et al. 2005, Kyriakidis and Skarkalis 2000). Oxidation products are formed when olive oils contact with air or light. The oxygen existing in the air has radical reactions between the double bonds of unsaturated fatty acids and the light accelerates the reaction. In the end conjugated hydroperoxides are formed. Due to the instability of these hydroperoxides, they quickly decompose into the aldehydes and ketones (Guimet, et al. 2004) and it is particularly known that these reactions are affecting the nutritional and safety properties of olive oils (Cheikhousman, et al. 2005).

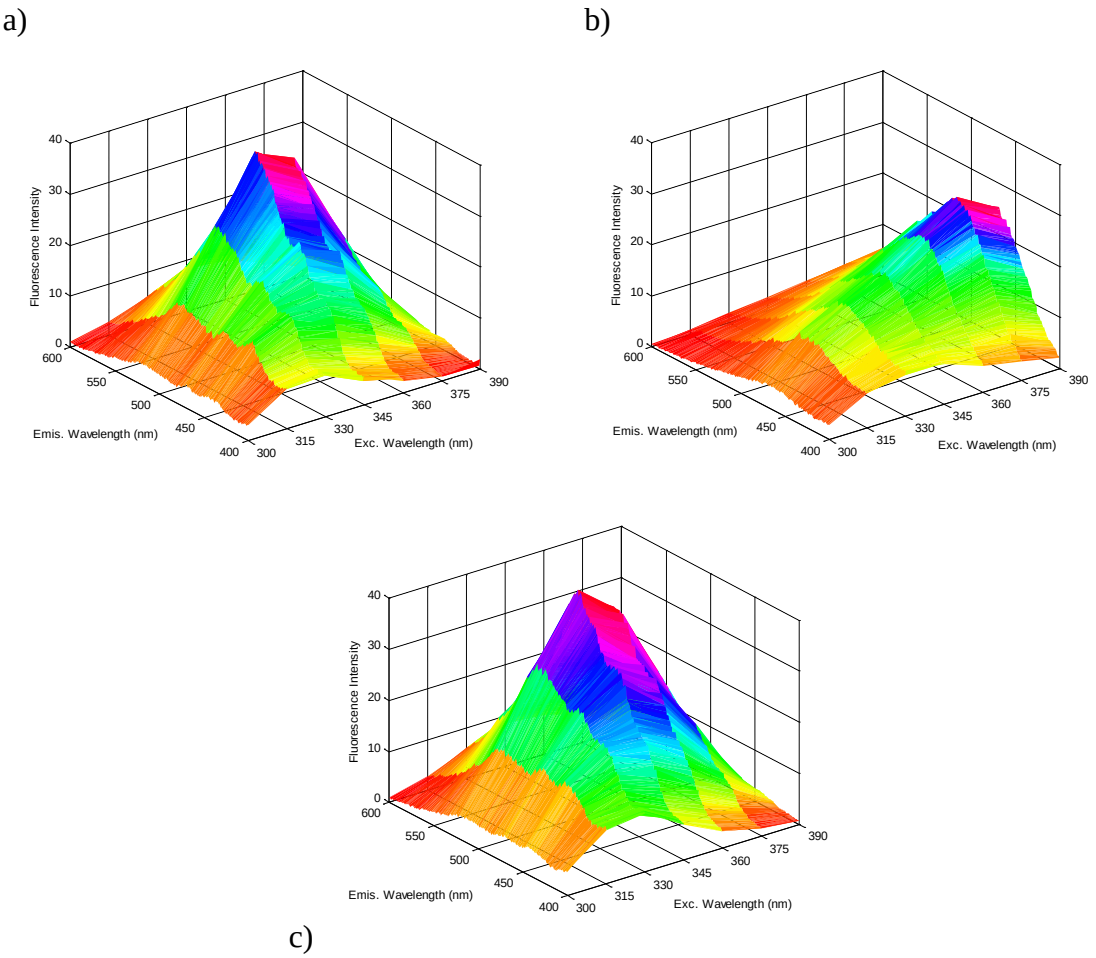


Figure 7.33. Excitation–emission fluorescence spectrum of a) extra virgin olive oil b) lampante olive oil c) refined olive oil between λ_{exc}=300–390 nm and λ_{em}=400–600 nm.

The fluorescence spectra of extra virgin olive oil and lampante olive oil can be seen in a two-dimensional way. In Figure 7.34, the existing peaks are clearly seen. The peaks around 400–550 nm are clearly different oil to oil. The acidity value of extra virgin olive oil is smaller than lampante olive oils; therefore the contents of monounsaturated and polyunsaturated fatty acids are caused the difference in fluorescence spectra. Also less intense peak around 470 nm for extra virgin olive oil is appeared in the fluorescence peak. Kyriakids and Skarkalis states in their study, this peak was due to oxidation of vitamin E. When it emitted the fluorescence light, vitamin E acetate was oxidized in that region.

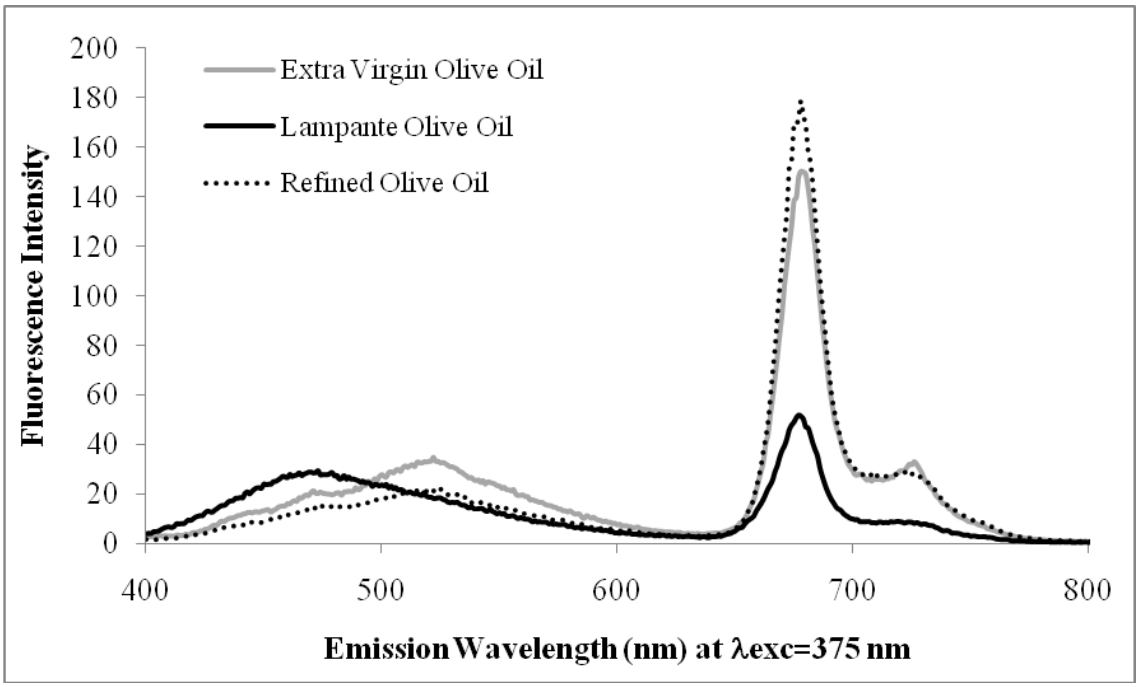


Figure 7.34. Excitation–emission fluorescence spectrum of extra virgin olive oil, refined olive oil, and lampante olive oil samples were measured between λ_{em}=400–600 nm at 375 nm excitation.

In excitation–emission fluorescence studies, mainly four different compounds (chlorophyll, vitamin E, phenolic antioxidants and oxidation products) act as main fluorescent components in olive oils. Synchronous fluorescence could be beneficial for the analysis of olive oil samples. It collects the emission only from the waveband where the absorption and emission bands overlap by the specified wavelength interval. This collection reduces the complexity of fluorescence spectra of fluorescent compounds. As a result, the selectivity of for individual components is considerably improved; much more fluorescent compounds will be detected.

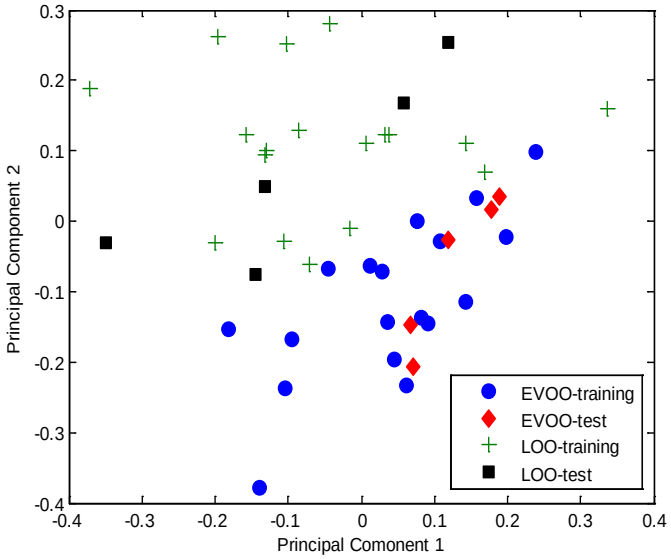
7.4.1.2. Classification Results of SVD-PCA and GAPCAD

7.4.1.2.1. Classification of Extra Virgin Olive Oils and Lampante Olive Oils

In order to classify the extra virgin olive oil (EVOO) and lampante olive oil (LOO) samples, both singular value decomposition based principal component analysis (SVD-PCA) and distance based genetic algorithm principal component analysis (GAPCAD) were performed. The success of GAPCAD algorithm was compared to the results of SVD-PCA. SVD–PCA was calculated from the emission spectra oils between the $\lambda_{emis.}$ = 400–800 nm measured at $\lambda_{exc.}$ = 300–435 nm with 15 nm increments. The observed three-dimension array data was concatenated in order to obtain two-way array spectral data matrix and EEF spectral data matrix was observed in 46 x 4510 (sample number x ($\lambda_{exc.}$ x $\lambda_{emis.}$)) dimensions. The olive oil samples of both training and test set were combined and used to calculate the principal components of the spectral data matrix. Totally seven significant principal components were calculated with 92.47% of explained variance. The scores of first two principal components (70.00% of explained variance) were used to see the classes of olive oil samples in the space. Figure 7.35.a. shows the scatter plot of scores of PC1 and PC2.

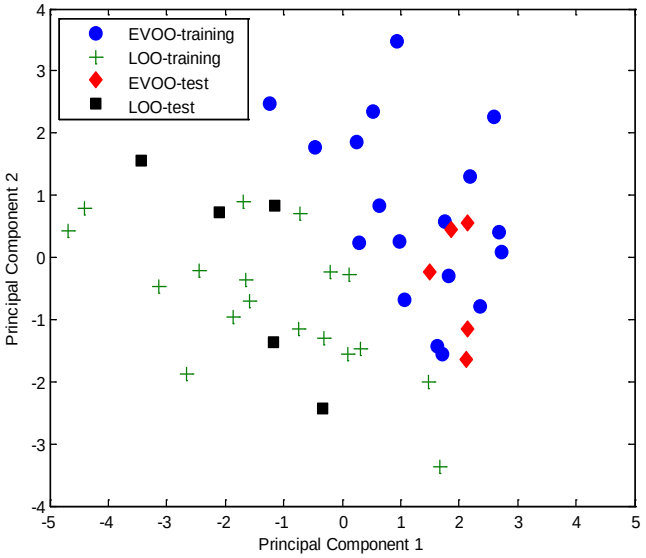
In the examination of GAPCAD algorithm two different sample set were used named as training and test set. The training set contains totally thirty six olive oil samples, half of them are extra virgin olive oil and the other remaining are lampante olive oil samples. The test set includes totally ten olive oil samples divided into two groups with five olive oil samples of each oil classes. The algorithm was initiated with 50 genes and 100 iteration numbers. For autoscaled data, the number of significant principal components was four with %97.91 of explained variance. The first two principal components (% 70.87 of explained variance) were used to differentiate the olive oil samples into two different classes (Figure 7.35.b.).

SVD-PCA



(a)Whole EEF spectral data

GAPCAD



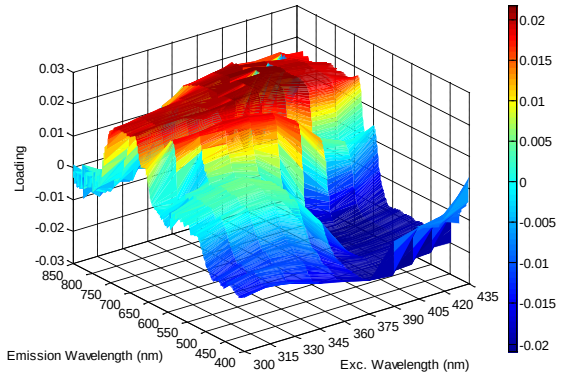
(b) Whole EEF spectral data

Figure 7.35. Score plot of principal components calculated from unfolded EEF data matrix of olive oil samples using SVD-PCA and GAPCAD for whole spectral data ($\lambda_{emis.}$ = 400–850 nm at $\lambda_{exc.}$ = 300–435 nm with 15 nm increments).

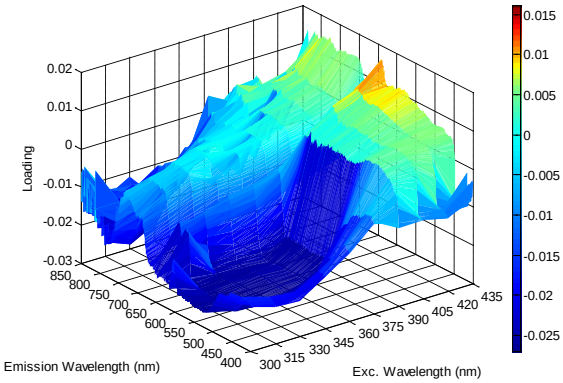
The distribution of olive oils obtained from the scores of principal components that were found out using SVD-PCA is along on both the PC1 and PC2. Generally, the positive scores of PC1 and negative scores of PC2 refer the EVOO samples whereas LOO samples are distributed on the combination of negative scores of PC1 and positive scores of PC2. The reasons of scattering on both principal components can be easily seen from the loading plots of PC1 and PC2 vs. refolded EEF matrix (Figure 7.36.a and Figure 7.36.b). The most contribution of PC1 comes from both the fluorescence peaks of chlorophylls ($\lambda_{emis.}$ = 600–700 nm at all excitation wavelengths) and tocopherol ($\lambda_{emis.}$ = 500–600 nm at $\lambda_{exc.}$ = 375–435 nm). The loading plot of PC2 shows only the contribution of tocopherol ($\lambda_{emis.}$ = 500–550 nm) at low excitation wavelengths. As it mentioned before, the main differences between the extra virgin olive and lampante olive are the amount of tocopherol and oxidation products. The fluorescence properties of tocopherol have weighted effects in both PC1 and PC2. Therefore the distributions of olive oil samples in the space are along on both scores.

On the other hand, the score plot obtained from GAPCAD calculation, the two different types of olive oil samples are separated mainly along principal component 1. Extra virgin olive oils mostly have positive scores, whereas the lampante olive oil samples mostly have negative scores on the score plot. Due to wavelength selection property of GAPCAD, it can be easily obtained which wavelength(s) were used in the classification of olive oil samples.

SVD-PCA

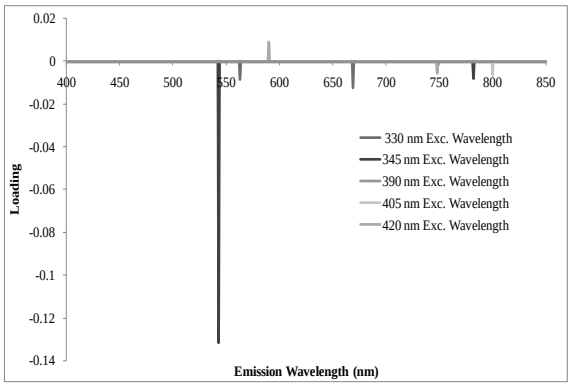


a) Loading plot of PC1 calculated from the whole EEF spectral data

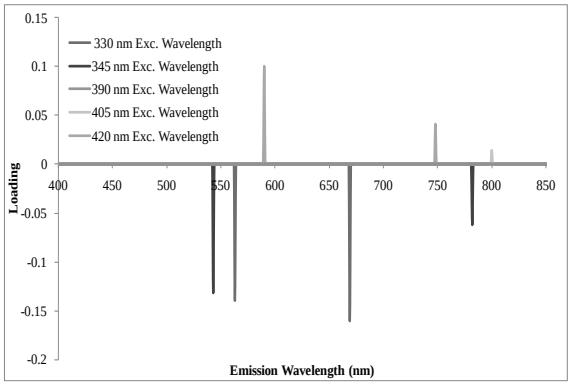


b) Loading plot of PC2 calculated from the whole EEF spectral data

GAPCAD



c) Loading plot of PC1 calculated from the whole EEF spectral data



d) Loading plot of PC2 calculated from the whole EEF spectral data

Figure 7.36. Loading plots of refolded EEF spectral data ($\lambda_{emis.}$ = 400–850 nm at $\lambda_{exc.}$ = 300–435 nm).of the olive oil samples (autoscaled data) a) Principal Component 1, b) Principal Component 2 obtained from SVD-PCA, c) Principal Component 1, b) Principal Component 2 obtained from GAPCAD.

In GAPCAD calculation, only seven emission wavelengths at different excitation wavelengths were used to classify the extra virgin olive oil and lampante olive oil samples. These wavelengths are 564, 669 nm ($\lambda_{exc.}$ =330 nm), 542, 782 nm ($\lambda_{exc.}$ =345 nm), 800 nm ($\lambda_{exc.}$ =405 nm), 590, 748 nm ($\lambda_{exc.}$ =420 nm), respectively. These wavelengths generally refer the fluorescence peaks of chlorophylls and tocopherol. Mainly the fluorescence peak of tocopherol at 345 nm excitation

wavelength has the most weighted loading values of PC1. This is the main reason why PC1 is more effective in the classification. As it mentioned before the EVOO are mostly yielded tocopherol. The results also show that again the chlorophylls fluorescence bands are valuable in the classification of olive oil samples (Figure 7.36.c and Figure 7.36.d), due to the large and intense fluorescence peak of chlorophylls.

In order to prove the effect of chlorophyll bands on to the classification results, the GAPCAD was performed hundred times and the distribution of selected emission wavelengths at a various excitation wavelengths was investigated. Figure 7.37 shows the frequency of selected wavelengths after 100 runs of GAPCAD processing. Generally the chlorophylls fluorescence peaks are the most selected emission wavelengths at a various excitation wavelengths. This is not an amazing result, since the chlorophylls fluorescence bands have a large peak area on the whole spectra. For this reason, the EEF matrix was designed without chlorophyll fluorescence peaks, is just including the emission wavelength 400–600 nm with excitation wavelengths ($\lambda_{exc.}$ = 300–435 nm).

To perform the classification analysis, EEF was designed without chlorophyll peak region and concatenation was done to observe two way array of 36 x 2010 (samples x {λ_{exc} x λ_{em}}) for the training set. The same process was done to the test sample set and the array of 10 x 2010 (samples x {λ_{exc} x λ_{em}}) data matrix was observed. For the SVD-PCA calculation training and test sample set were combined and evaluated together. In both algorithms, autoscaling preprocessing technique was applied to the spectral data matrix.

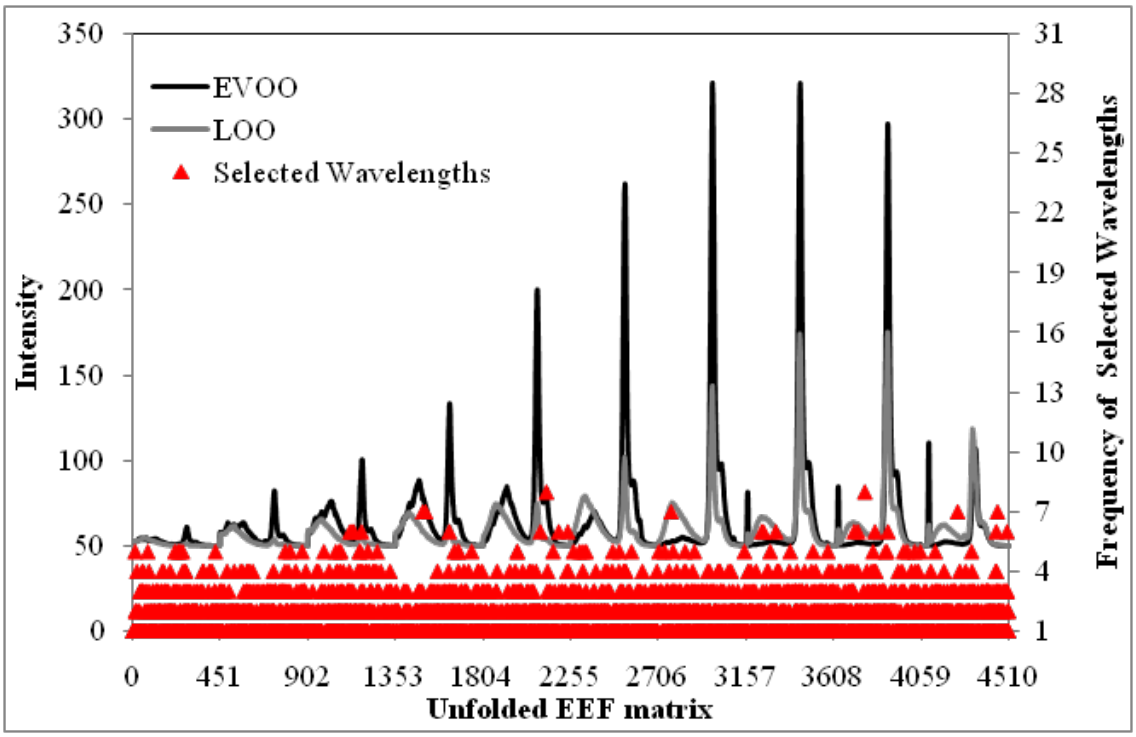
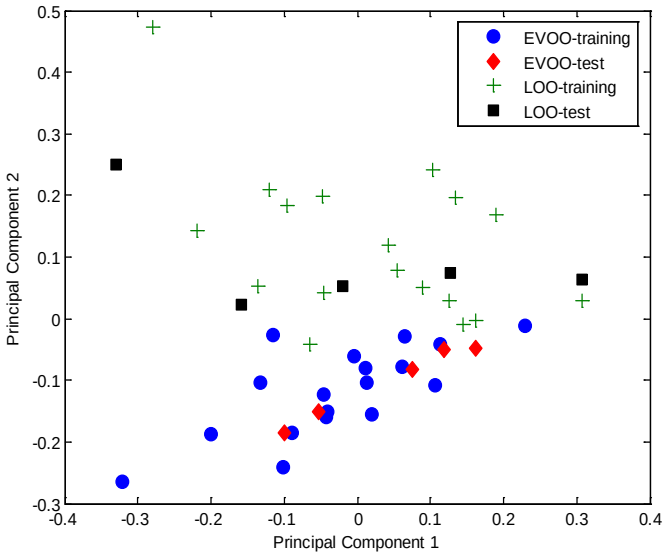


Figure 7.37. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of whole EEF spectral data.

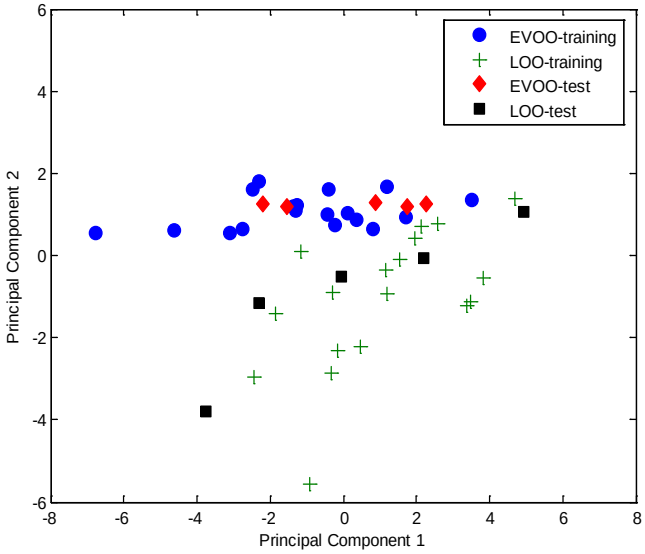
In the examination of SVD-PCA algorithm, ten significant principal components were calculated with an 81.07% of explained variance. The scores of first two principal components (with 46.06% of explained variance) were plotted and the distributions of olive oil samples were obtained (Figure 7.38.a). On the other side, distance based genetic algorithm principal component analysis was set to initiate with 50 gene and 50 iteration numbers and at the end of the analysis the first five principal component scores was found as 94.38% of explained variance value. The score plot of the first two principal components (75.49% of explained variance) were plotted to see the distribution of olive oil samples (Figure 7.38.b).

SVD-PCA



(a) EEF spectral data without chlorophyll fluorescence peaks.

GAPCAD

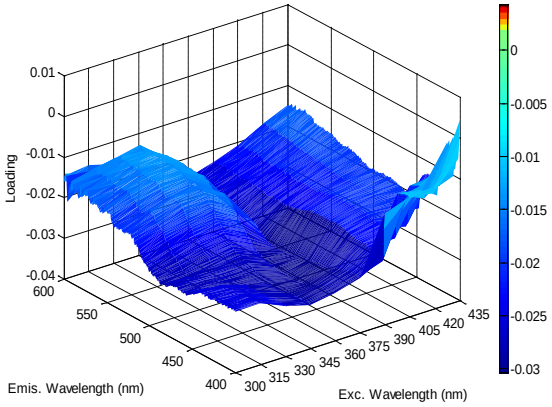


(b) EEF spectral data without chlorophyll fluorescence peaks.

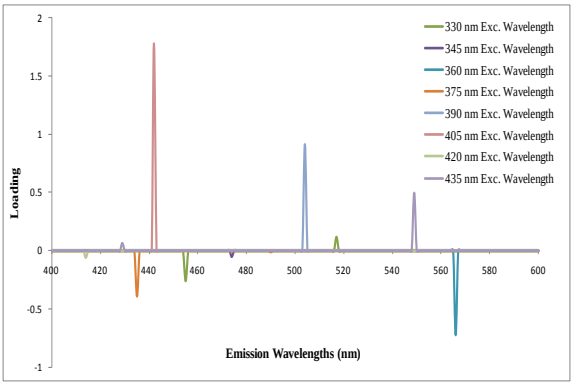
Figure 7.38. Score plot of principal components calculated from unfolded EEF data matrix of olive oil samples using SVD-PCA and GAPCAD for spectral data without chlorophyll peaks ($\lambda_{emis.}$ = 400–600 nm at $\lambda_{exc.}$ = 300–435 nm with 15 nm increments).

As it is seen from the figure, score plots obtained from both algorithms shows the same distributions. Only the negative and positive scores of principal components refer different type of olive oils. They are mainly along on the PC2. To visualize the effect of florescent compounds those are the constituents of olive oils, the loading values of each principal components were plotted against refolded EEF spectral data matrix. The loading plot of PC2 shows that the emission wavelengths in the range of 550 – 600 nm at low excitation wavelengths are the most contributed florescence region in the classification. This range refers the tocopherol compounds of olive oils, therefore the extra virgin olive oil samples are mainly distributed on negative scores of PC2. In GAPCAD calculation, only eleven emission wavelengths at different excitation wavelengths were used to classify the extra virgin olive oil and lampante olive oil samples. These wavelengths are 455, 518 nm ($\lambda_{exc.}$ =330 nm), 474 nm ($\lambda_{exc.}$ =345 nm), 566 nm ($\lambda_{exc.}$ =360 nm), 436 and 490 nm ($\lambda_{exc.}$ =375 nm), 442 nm ($\lambda_{exc.}$ =405 nm), 415, 430 nm ($\lambda_{exc.}$ =420 nm), 549 nm ($\lambda_{exc.}$ =435 nm), respectively. Among all these selected wavelengths, the emission wavelength of 518 nm at 330 nm excitation has the most contribution in the classification of olive oil samples. As it is seen from the score plot of olive oils obtained from GAPCAD calculation, the positive scores of PC2 refer the distribution of extra virgin olive oil samples.

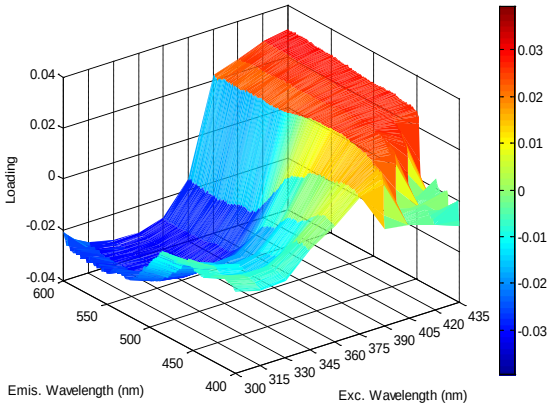
SVD-PCA GAPCAD



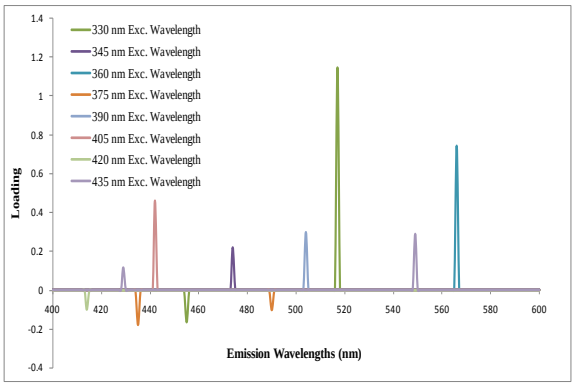
a) Loading plot of PC1 calculated from the EEF spectral data without chlorophyll fluorescence peak



c) Loading plot of PC1 calculated from the EEF spectral data without chlorophyll fluorescence peak



b)) Loading plot of PC2 calculated from the EEF spectral data without chlorophyll fluorescence peak



d) Loading plot of PC2 calculated from the EEF spectral data without chlorophyll fluorescence peak

Figure 7.39. Loading plots of refolded EEF spectral data ($\lambda_{emis.}$ = 400–600 nm at $\lambda_{exc.}$ = 300–435 nm).of the olive oil samples (autoscaled data) a) Principal Component 1, b) Principal Component 2 obtained from SVD-PCA c) Principal Component 1, b) Principal Component 2 obtained from GAPCAD.

GAPCAD was performed 100 times, in order to see the real effect of fluorescent compounds. At the end of the performing, the frequency of selected wavelengths was determined and plotted. Figure 10 proves that the fluorescence property of tocopherol is the dominant factor in the classification of extra virgin and lampante olive oil samples.

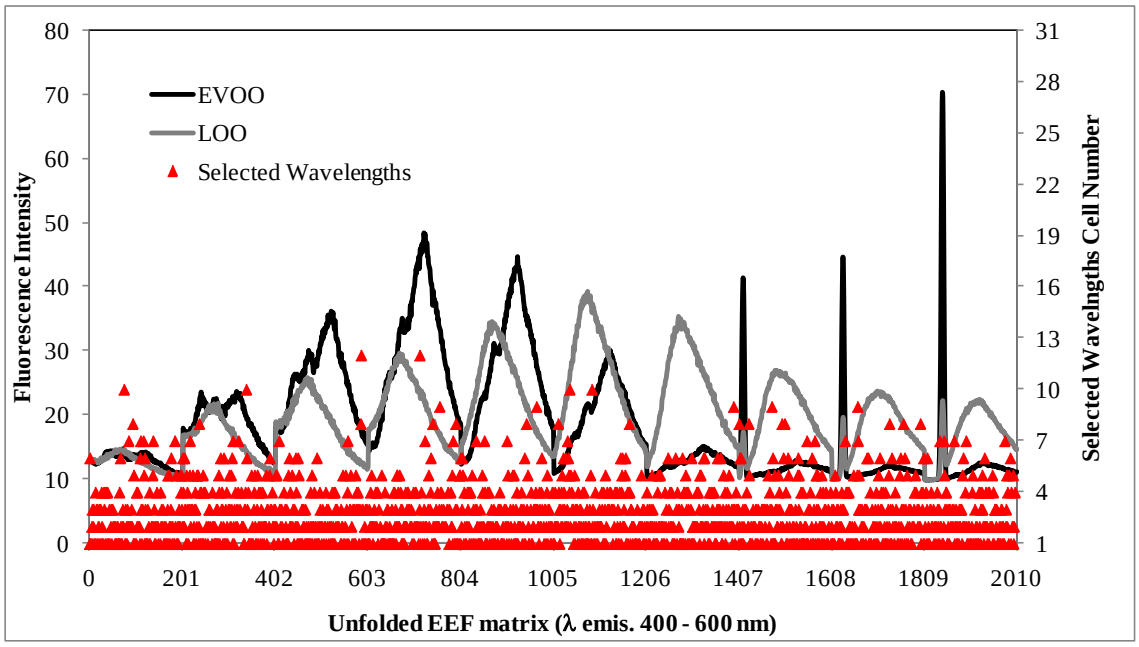


Figure 7.40. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of EEF spectral data without chlorophyll fluorescence peak.

7.4.1.2.2. Classification of Refined Olive Oils and Lampante Olive Oils

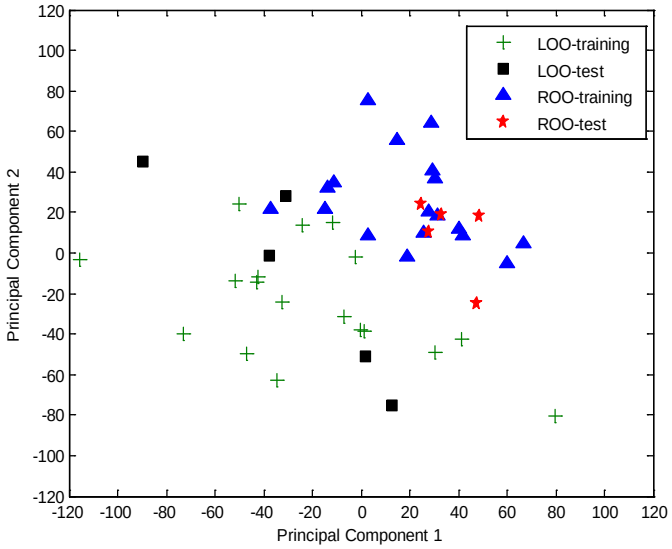
Singular value decomposition based principal component analysis (SVD–PCA) was calculated from the emission spectra oils between the $\lambda_{emis.}$ = 400–850 nm measured at $\lambda_{exc.}$=300–435 nm with 15 nm increments. The observed three-dimension array data was concatenated in order to obtain two-way array spectral data matrix and EEF spectral data matrix was observed in 46 x 4510 (sample number x ($\lambda_{exc.}$ x $\lambda_{emis.}$)) dimensions. Distance based genetic algorithm principal component analysis was also performed to the classification of refined olive oil and lampante olive oil samples. The training set contains totally thirty six olive oil samples, half of them are refined olive oil and the other remaining are lampante olive oil samples. The test set includes totally ten olive oil samples divided into two groups with five olive oil samples of each oil classes. Spectral data matrices of training set and test set were observed with same

concatenation processing. The resulted training and test set's spectral data matrices have 36 x 4510, 10 x 4510 (sample number x ($\lambda_{exc.}$ x $\lambda_{emis.}$)) dimensions, respectively. Due to the high intense fluorescence peak of chlorophylls, autoscaling preprocessing technique was performed before the decomposition EEF spectral data matrix. The aim of the preprocessing is to make use of the contributions of other species exist in the olive oils.

In SVD-PCA calculation, ten significant principal components with 94.17% of explained variance were found for autoscaled spectral data. The first two principal components were used to identify the distribution of olive oil samples (Figure 7.41.a). These two principal components have 67.18% of explained variance. However it must be noted that both groups of olive oil samples overlapped slightly in the classification process using SVD-PCA. Both the scores values of PC1 and PC2 have same contribution on the classification of olive oil samples. On the other side, GAPCAD algorithm was initiated with 50 genes and 100 iteration numbers. For autoscaled data, the number of significant principal components was four with %94.01 of explained variance. The first two principal components (% 76.88 of explained variance) were used to differentiate the two types of olive oil samples into two different classes (Figure 7.41.b). As it seen from the score plot, the two different types of olive oil samples are separated mainly along principal component 1. Refined olive oils mostly have positive scores, whereas the lampante olive oil samples mostly have negative scores on the score plot.

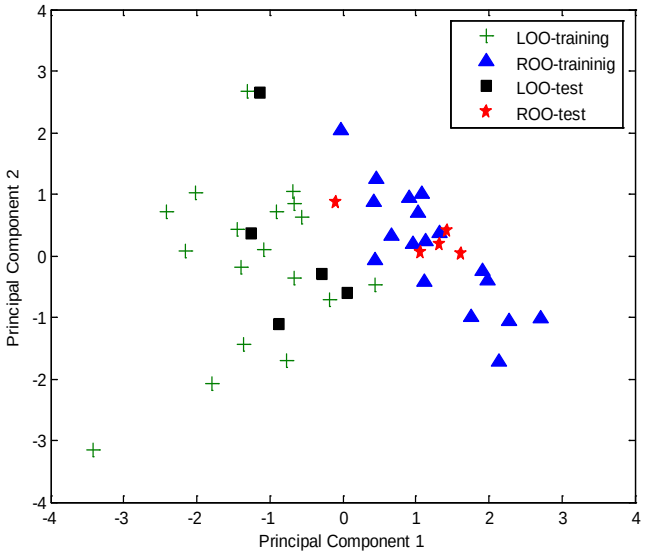
To observe the which wavelength region(s) have the most contribution on the classification of olive oil samples, the loading plots of PC1 and PC2 were drawn against to the refolded EEF spectra. In SVD-PCA calculation, the contribution of oxidation products and vitamin E are seen (at higher excitation wavelengths) from the loading plot of PC1 (Figure 7.42.a) and also the contribution of chlorophyll fluorescence peaks is similar as the other components. On the other hand, the loading plot of PC2 shows that the tocopherol and oxidation products (the peaks at lower excitation wavelengths) have the larger contributions, as it is seen from Figure 7.42.b.

SVD-PCA



(a) Whole EEF spectral data

GAPCAD

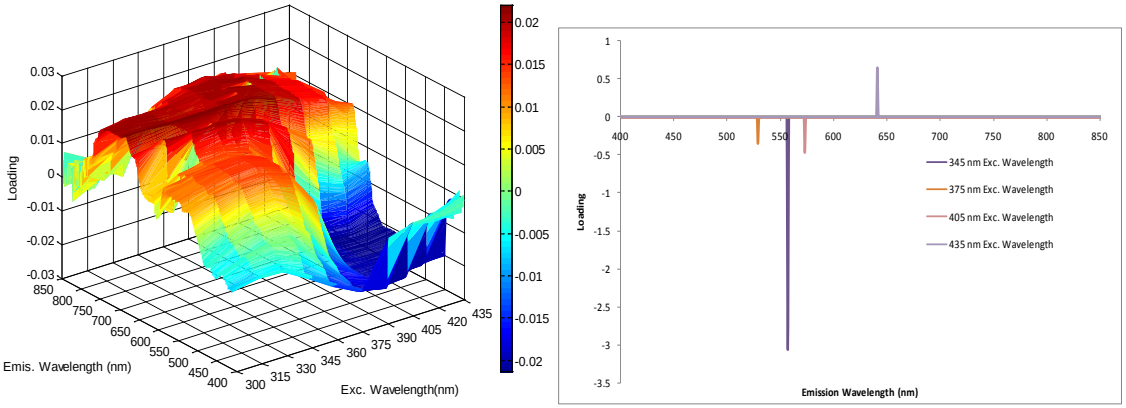


(b) Whole EEF spectral data

Figure 7.41. Score plot of principal components calculated from unfolded EEF data matrix of olive oil samples using SVD-PCA and GAPCAD for whole spectral data ($\lambda_{emis.}$ = 400–850 nm at $\lambda_{exc.}$ = 300–435 nm with 15 nm increments).

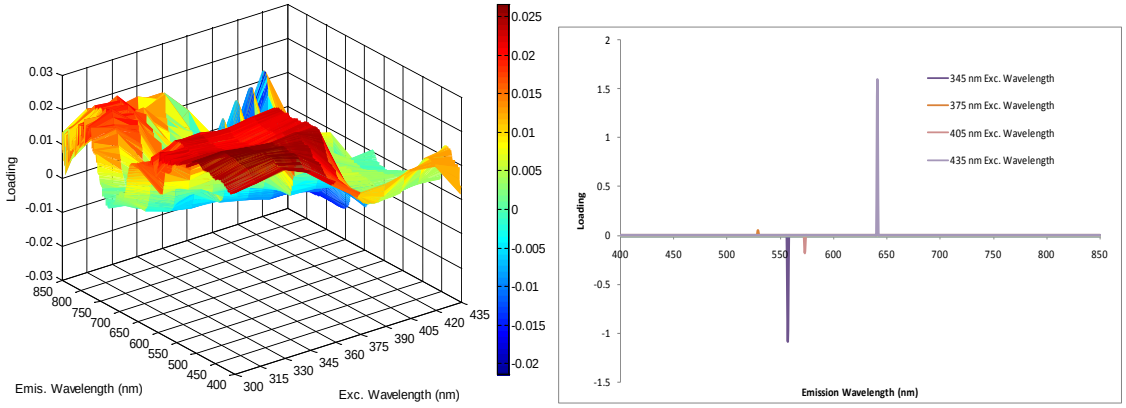
Due to the nature of genetic algorithms, the wavelengths which have desired knowledge about samples will be selected and the corresponding fluorescence properties will be used in the classification of olive oil samples. In the examination of GAPCAD, only four emission wavelengths at four different excitation wavelengths were used in the calculation of principal components and the scores of these principal components were used to classify the two types of olive oil samples. All these emission wavelengths are 557 nm ($\lambda_{exc.}$ = 345 nm), 529 nm ($\lambda_{exc.}$ = 375 nm), 573 nm ($\lambda_{exc.}$ = 405 nm), 641 nm ($\lambda_{exc.}$ = 435 nm). The loading plot of these selected wavelengths at different excitation wavelengths is shown in Figure 7.42. The selected wavelengths generally related to the vitamin E contents of olive oils which are the refined olive oil samples. Refined olive oil samples generally called as pure olive oil, since they are obtained from refining virgin olive oils which have higher acidity value and/or organoleptic (taste and aroma) defects and they are eliminated after refining of virgin olive oil samples. Therefore the contribution of vitamin E is higher than the other constituents, as it is seen from Figure 7.42. This contribution also explains the reasons of distribution of olive oil samples mainly on PC1 (Figure 7.41).

SVD-PCA GAPCAD



a) Loading plot of PC1 calculated from the whole EEF spectral data

c) Loading plot of PC1 calculated from the whole EEF spectral data



b) Loading plot of PC2 calculated from the whole EEF spectral data

d) Loading plot of PC2 calculated from the whole EEF spectral data

Figure 7.42. Loading plots of refolded EEF spectral data ($\lambda_{emis.}$ = 400–850 nm at $\lambda_{exc.}$ = 300–435 nm).of the olive oil samples (autoscaled data) a) Principal Component 1, b) Principal Component 2 obtained from SVD-PCA, c) Principal Component 1, b) Principal Component 2 obtained from GAPCAD.

In order to prove the effect of vitamin E bands in the classification of olive oil samples, the GAPCAD was performed hundred times and the distribution of selected emission wavelengths at a various excitation wavelengths was investigated. Figure 7.43 shows the frequency of selected wavelengths after 100 runs of GAPCAD processing. At lower excitation wavelengths, the fluorescence peaks of vitamin E are mostly used in

the classification of refined and lampante olive oil samples. Due to the predomination of vitamin E fluorescence bands, either SVD-PCA or GAPCAD algorithm were performed to the whole EEF spectral data.

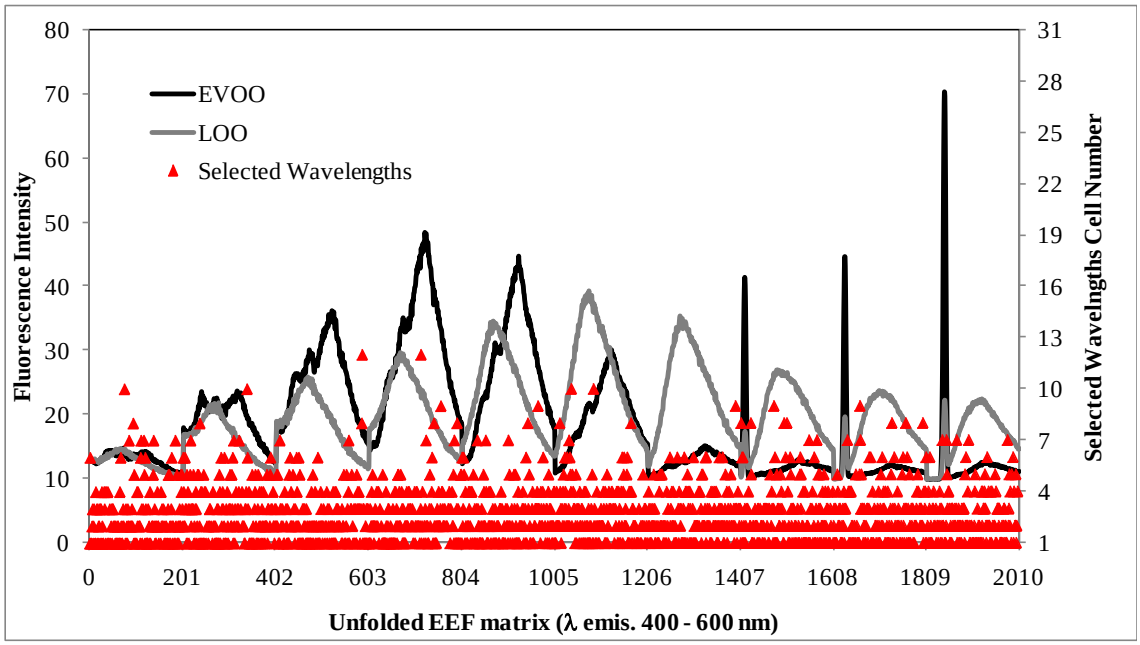
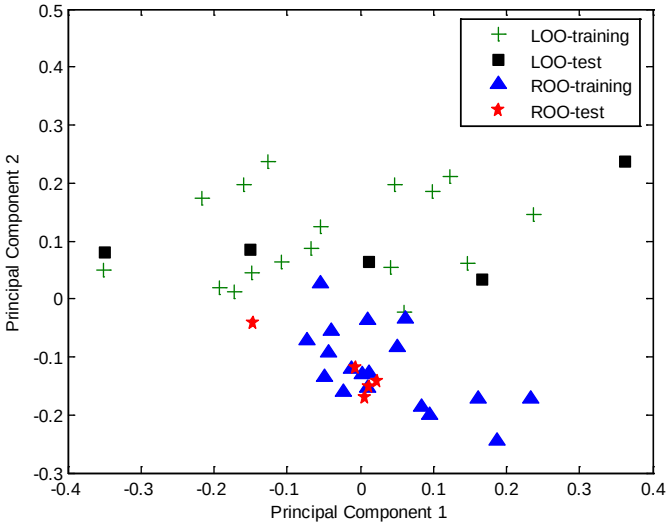


Figure 7.43. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of whole EEF spectral data.

Also in order to compare the differences between the classification results of spectra data matrix with and without chlorophyll fluorescence peaks, spectral data matrix was redesigned without the fluorescence peak of chlorophylls. The new designed spectral data matrix has 36 x 2010 and 10 x 2010 dimensions for training and test set, respectively. As it mentioned before training and test sample sets were combined and the total spectral data matrix (with a 46 x 2010 dimension) was used in the analysis of SVD-PCA.

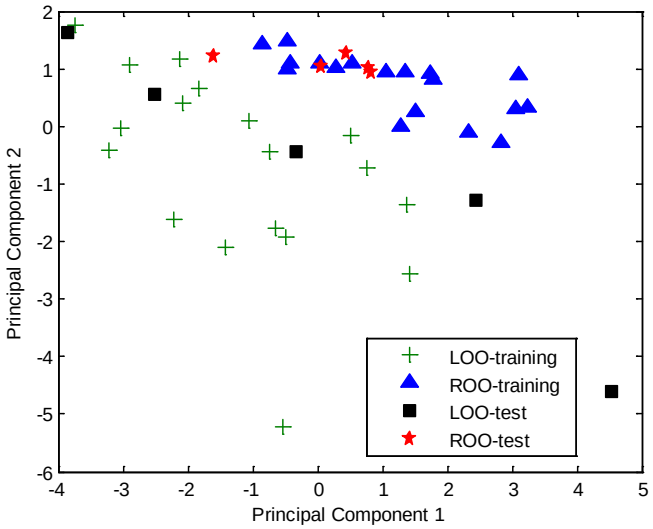
For the examination of autoscaled spectral data, ten significant principal components (with a 77.21 % of explained variance) were calculated and the scores of first two principal components (with a 46.76 % of explained variance) were used to visualize the classes of olive oils (Figure 7.44.a.).

SVD-PCA



(a) EEF spectral data without chlorophylls

GAPCAD



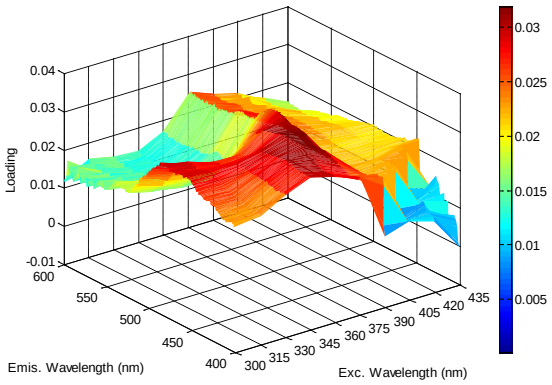
(b) EEF spectral data without chlorophylls

Figure 7.44. Score plot of principal components calculated from unfolded EEF data matrix of olive oil samples using a) SVD-PCA and b) GAPCAD for spectral data without chlorophyll peaks ($\lambda_{emis.}$ = 400–600 nm at $\lambda_{exc.}$ = 300–435 nm with 15 nm increments).

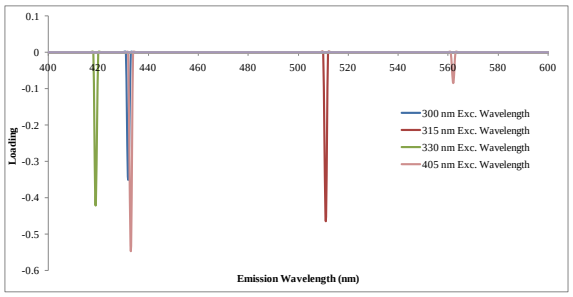
The results of SVD-PCA show that the lampante olive oils are mainly along on the positive scores of PC2, whereas the negative scores of PC2 assign the refined olive oils. Due to the broad range in the acidity value of LOO, LOO samples are more scattered than ROO samples. The reason of scattering on PC2 becomes from the contribution of emission wavelengths in the range of 500–600 nm at low excitation wavelengths (Figure 7.45.b). On the other hand there is no intense contribution from PC1.

On the other side, GAPCAD algorithm was initiated by predefinition of 50 genes and 100 iterations for training set. After the decision rule was obtained for the autoscaled EEF spectral data, it applied to the test set and the distributions of olive oils samples were obtained by plotting the first two principal components. Four significant principal components were found out with a 99.00% of explained variance. The first two principal components have a 92.78% of explained variance. As it is seen from Figure 7.44.b., the classes of olive oil samples are scattered on both PC1 and PC2.

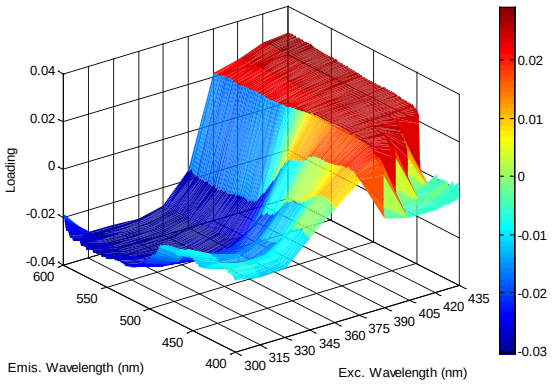
SVD-PCA GAPCAD



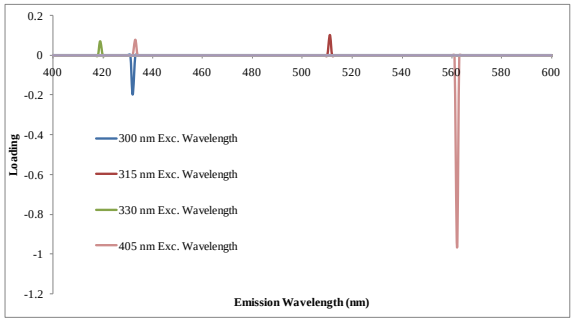
a) Loading plot of PC1 calculated from the EEF spectral data without fluorescence peak of chlorophyll.



c) Loading plot of PC1 calculated from the EEF spectral data without fluorescence peak of chlorophyll.



b) Loading plot of PC2 calculated from the EEF spectral data without fluorescence peak of chlorophyll.



d) Loading plot of PC2 calculated from the EEF spectral data without fluorescence peak of chlorophyll.

Figure 7.45. Loading plots of refolded EEF spectral data ($\lambda_{emis.}$ = 400–600 nm at $\lambda_{exc.}$ = 300–435 nm).of the olive oil samples (autoscaled data) a) Principal Component 1, b) Principal Component 2 obtained from SVD-PCA, c) Principal Component 1, b) Principal Component 2 obtained from GAPCAD.

In order to understand the effect principal components, the loading values of each principal component were plotted to the refolded EEF spectral data for both the results of SVD-PCA and GAPCAD. As it mentioned before, the results of SVD-PCA shows that the most contribution becomes from the fluorescence peaks of tocopherols ($\lambda_{emis.}$ = 500 – 600 nm at $\lambda_{exc.}$ = 300 – 360 nm). From the results that obtained from

GAPCAD algorithm, only five emission wavelengths at 300, 315, 330, and 405 nm excitations were used to classify the olive oil samples. The selected emission wavelengths which have the largest loading values of PC1 are 419 nm (at 330 nm excitation), 433 nm (at 405 nm excitation), and 437 nm (at 300 nm excitation) refer the oxidation products of lampante olive oils and the 511 nm emission at 315 nm excitation assign the existed of tocopherol in the olive oil samples (Figure 7.45.c). The fluorescence peak of tocopherol at 562 nm emission ($\lambda_{exc.}$ = 405 nm) have the largest contribution of PC2 in the classification of olive oil samples (Figure 7.45.d). Therefore the olive oil samples are scattered on both PC1 and PC2 (Figure 7.44.b).

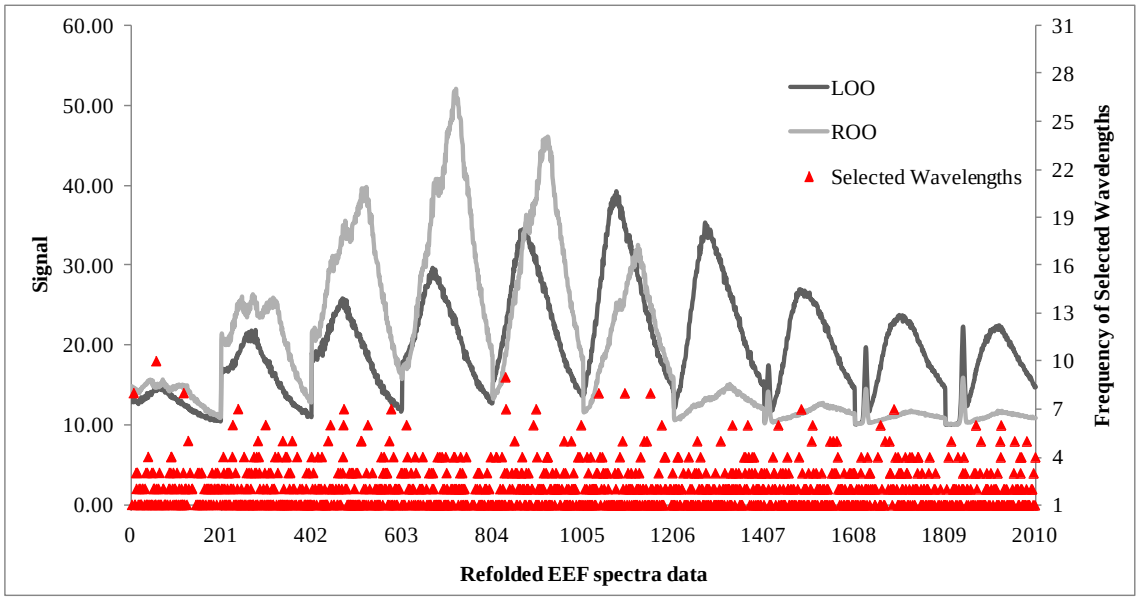


Figure 7.46. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of EEF spectral data without chlorophyll fluorescence peaks

In order to see the frequency of selected wavelengths, GAPCAD was performed 100 times. Figure 7.46 shows that, the emission in the range of 420–450 nm and the fluorescence peaks of tocopherols (emission in the range of 500–550 nm) are the mostly used wavelengths.

7.5. Classification Results of SIMCA and GADA

7.5.1.1.1. Classification of Extra Virgin Olive Oil and Lampante Olive Oil Samples.

Classification of two different olive oil types named as EVOO and LOO were studied by using genetic algorithm based discriminant analysis (GADA) and SIMCA. In order to achieve the classification performance of GADA, the results of SIMCA was compared to the results of GADA. Totally forty six olive oil samples including 23 of EVOO and 23 LOO olive oil samples were used to construct the data set of olive oils. These data matrix were divided into two different set named as training and test set. Training set contains totally thirty six olive oil samples with their corresponding excitation–emission fluorescence (EEF) spectra. Each spectrum contains totally 4510 emission wavelengths at 10 different excitation wavelengths. Therefore the spectral data matrix was concatenated in order to obtain two-dimension array data matrix with 36 x 4510 dimensions. Same procedure was repeated for the independent test set. Test set includes ten olive oil samples with 4510 emission wavelengths. Before starting the classification analysis, in both analyses autoscaling was used as preprocessing technique. Two categories were predefined as class 1 including 18 samples of extra virgin olive oils, class 2 for 18 samples of lampante olive oil samples and 5 of extra virgin olive oil and 5 of lampante olive oil samples building test set.

Classification of olive oil samples by SIMCA analysis of EEF spectra (whole spectra) is shown in (Figure 7.47). Vertical and horizontal lines in Cooman's plot indicate the boundaries for classifying samples at a 5% significance level. The olive oil samples existed in the training set do not place in the outlier region which is indicated as neither extra virgin olive oil nor lampante olive oil samples. There is only one LOO sample in the outlier region. Two samples of EVOO and one of LOO sample can be assigned as belonging to both classes. Eight significant principal components were calculated for both EVOO and LOO classes with a 71.6% and 73.7% of total variances, respectively.

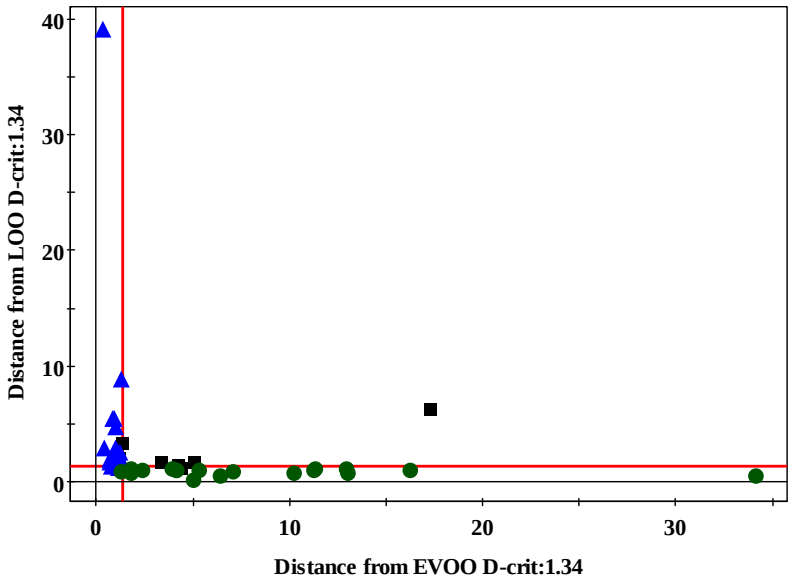


Figure 7.47. Cooman's plot of olive oil samples obtained from SIMCA analysis of EEF spectra (triangle: EVOO-training, circle: LOO-training, box: test set)

GADA was used to determine which variables better modulate and discriminate between the classes or the categories established depending on variety of olives. Same data matrix was used to build the classes of olive oil samples. The algorithm was initiated with 10 genes and 10 iteration number at 95% confidence level. 10 significant principal components were calculated with a 91.02% of explained variance. The Cooman's plot of olive oil samples were shown in Figure 7.48. From the results of GADA analysis, three of LOO samples existing in test set were classified as a third class which is not belonging to both classes. And also one of the LOO sample was found as an outlier at 95% confidence level. On the other side the classification results of EVOO samples was better than the results of LOO samples. The main reason for that kind of classification is the diversity in the amount of free fatty acids values of lampante olive oils.

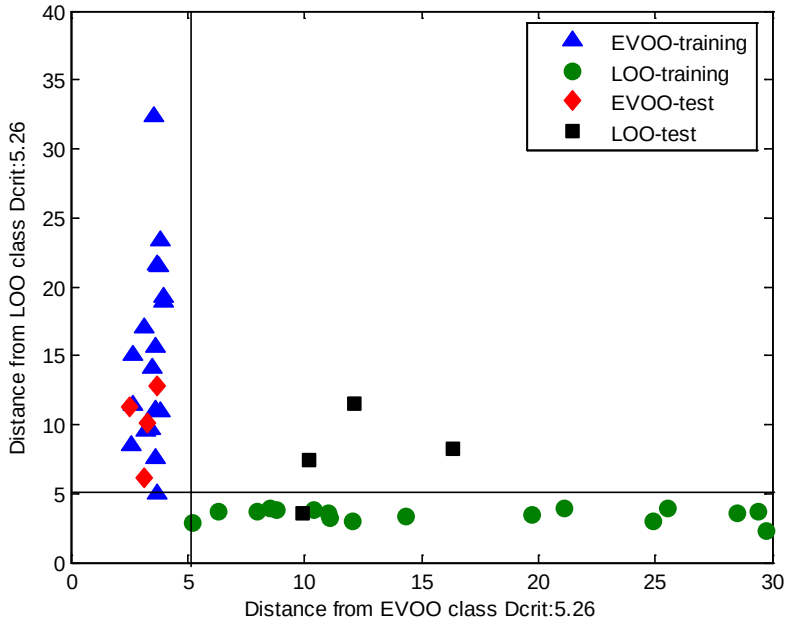


Figure 7.48. Cooman's plot of olive oil samples obtained from GADA analysis of EEF spectral matrix.

Due to the working principle of GADA that is based on the natural evolution, the emission wavelengths at different excitation wavelengths were selected according to the intensity values of components existing in the olive oil samples. These selected wavelengths are shown in Figure 7.49. The most important point of these wavelengths is that contain the necessary information for the classification of extra virgin olive oil samples and lampante olive oil samples. At the end of the analysis, these wavelengths are generally related to the fluorescence peaks of chlorophyll compound existing in the olive oil samples and at lower excitation wavelengths the peak appeared around 445 nm emission wavelength were used in the classification of samples. This peak generally refers the monosaturated fatty acids and high content of phenolic antioxidants which characterize the good quality virgin olive oils. Also the excitation wavelengths at 300 and 360 nm contain the most selected emission wavelengths.

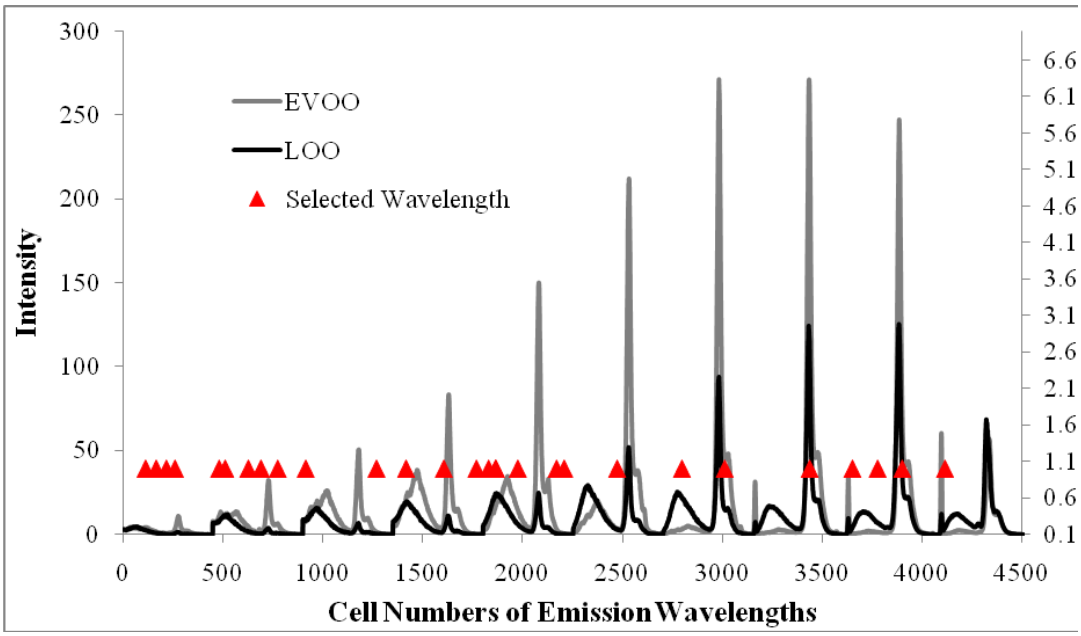


Figure 7.49. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis.

7.5.1.1.2. Classification of Refined Olive Oils and Lampante Olive Oil Samples

GADA were also examined in the classification of refined olive oil and lampante olive oil samples using EEF spectral data matrix. The observed three-dimension array data was concatenated in order to obtain two-way array spectral data matrix and EEF spectral data matrix was observed in 46 x 4510 (sample number x ($\lambda_{exc.}$ x $\lambda_{emis.}$)) dimensions. In that time, two classes were predefined as class 1, 18 of refined olive oil samples and class 2, 18 of lampante olive oil samples. An independent test set was built using different 5 of refined olive samples and 5 of lampante olive oil samples. Before starting the analysis of SIMCA and GADA, spectral data matrix was autoscaled.

8 principal components (with a 73.7% and 70.7% of total variance for class 1 and class 2, respectively) were calculated for both class 1 and class 2 in the analysis of SIMCA At 95% confidence level critical limits were found as 1.34 and these limits were drawn as vertically and horizontally in the Cooman's plot (Figure 7.50). As it is seen from the plot ROO-22 was appeared as an outlier or assigned as not belonging to the both classes.

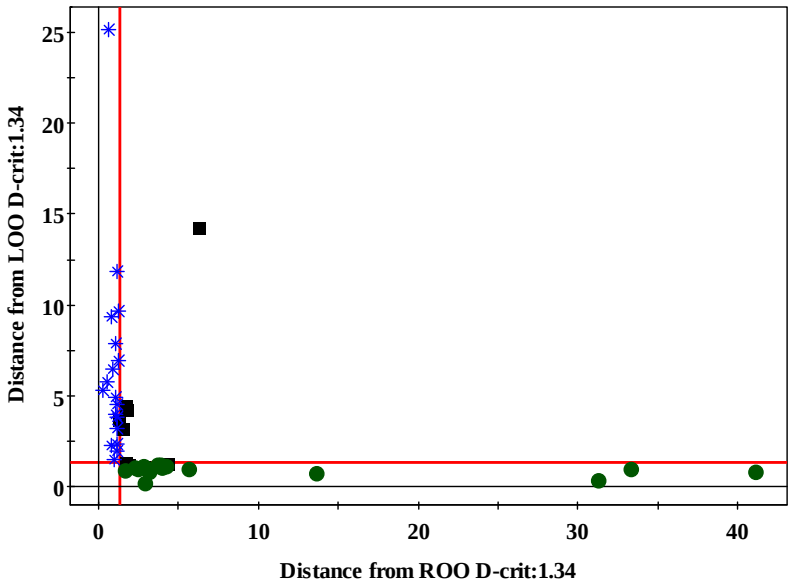


Figure 7.50. Cooman's plot of olive oil samples obtained from SIMCA analysis of EEF spectra (star: ROO-training, circle: LOO-training, box: test set)

GADA analysis was initiated with 6 genes and 10 iteration numbers at 95% confidence level. 15 principal components were used to identify the olive oil classes in Cooman's plot. The vertical and horizontal critical limits were found as 12.49 for both classes at 95% confidence level with (m-PC-1) degrees of freedom for each class. As it is seen from the Figure 7.51, the training sets of each olive oil type were classified in a level of 94.4% (17/18) for ROO and 88.89% (16/18) for LOO samples. On the other hand the test set of olive oil samples were classified in a level of 70% (7/10). According to the results assigned in Cooman's plot one of ROO and two of LOO have similar properties, and also one LOO sample existing in the test set have similar properties.

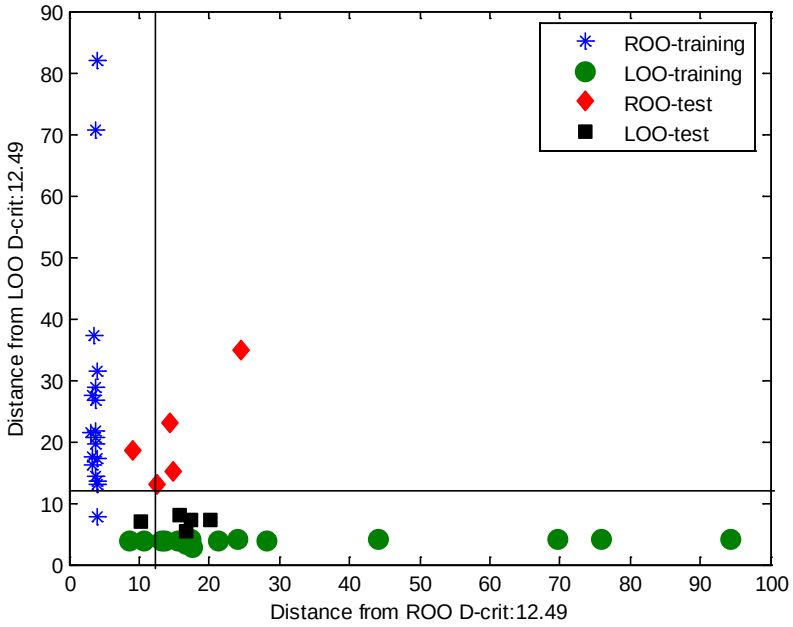


Figure 7.51. Cooman's plot of olive oil samples obtained from GADA analysis of EEF spectra.

As it is mentioned before, genetic algorithm is imposed to the discriminant analysis to select the wavelengths which contains the necessary information that will be useful in the classification of samples. At the end of the analysis, forty six wavelengths at different excitation wavelengths were selected. Spectral data matrix was contained totally 4510 emission wavelengths, and GADA was used only 46 of them in the classification of refined olive oil and lampante olive oil samples. These wavelengths generally assign the fluorescence peaks of chlorophylls and tocopherols existing in the olive oil samples. Especially the 300, 315, 330, and 435 nm excitation wavelengths were used in the classification olive oil samples. The emission wavelengths of tocopherols at 300, 315 and 330 nm excitation have the large contribution on the classification whereas the emission wavelengths of chlorophylls have contribution only at 435 nm excitation.

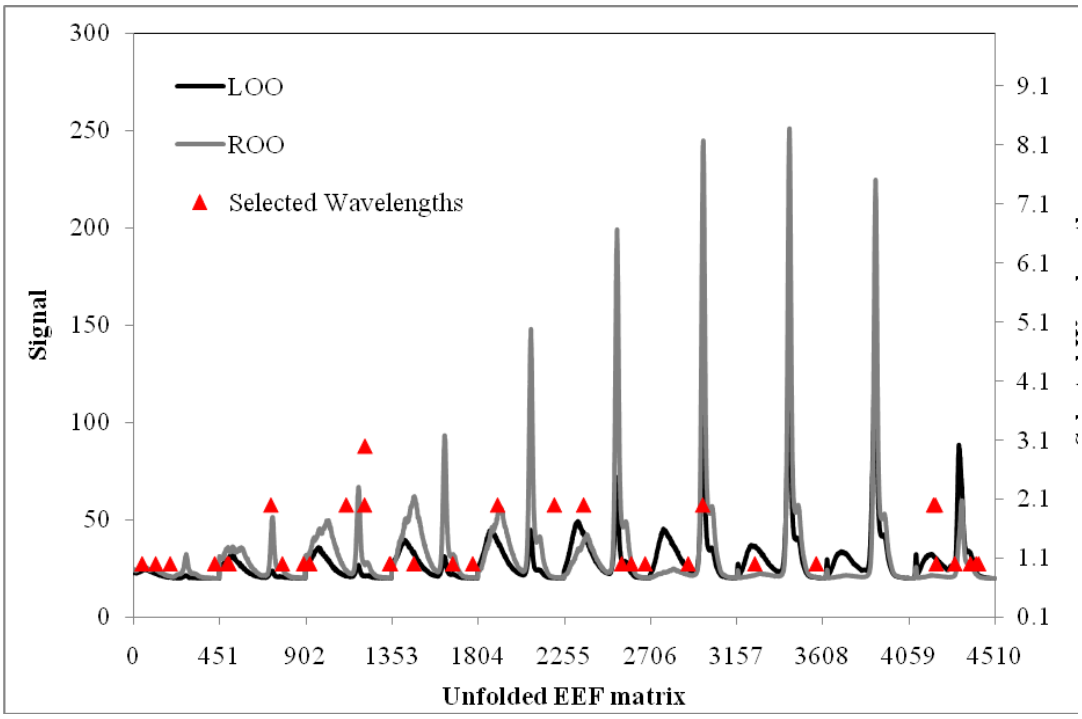


Figure 7.52. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis.

7.5.2. Total Synchronous Fluorescence Results

7.5.2.1. Total Synchronous Fluorescence Measurements of Olive Oil Samples

In the literature, the synchronous fluorescence studies indicates two main fluorescent compounds named as tocopherol and chlorophyll in the olive oil samples (Sikorska et al., 2004, Sikorska et al., 2005). The synchronous fluorescence spectra of pure α-tocopherol and bacteriopheophytin *c* (indicates chlorophyll) were studied in the range of 250–700 nm with 10, 20, 30, 60 and 80 nm offset values were studied. Bacteriopheophytin *c* was used instead of chlorophyll compound. Bacteriopheophytin *c* differs in structure of chlorophylls *a* and *b*, in that the pyrrole ring IV is not reduced, and the position 17 is esterified by an acrylic residue instead of a propionic group, terminal carboxylic group being generally not esterified (Schoefs, 2002). However the fluorescence spectra of the pigments of chlorophylls group are very similar. The resulted synchronous spectra at 60 and 80 nm offset values of chlorophyll exhibits two

fluorescence peaks at 665, 610 and 665, 603 nm, respectively. The synchronous fluorescence spectrum at 10 nm offset value of α-tocopherol in *n*-hexane shows a narrow band with high intensity at 301 nm. However at higher offset values, the maximum of band is shifted to the 317–319 nm with large and weak intensity depending on oil. (Sikorska, et al. 2005).

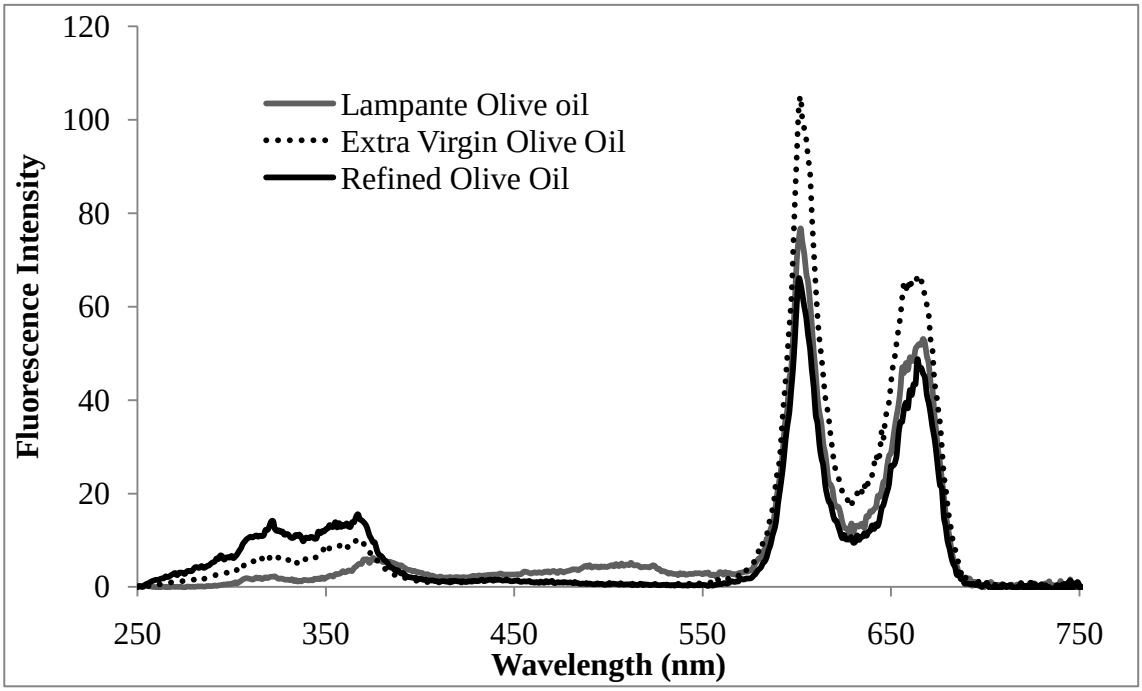


Figure 7.53. Synchronous fluorescence spectra of extra virgin olive, refined olive, lampante olive oil by synchronous scanning the excitation and emission monochromator maintained an offset value of 80 nm in the spectral range 250–750 nm.

In the literature another synchronous fluorescence study indicates that, the synchronous fluorescence spectral differences were based on only the acidity of lampante and virgin olive oils (Poulli, et al. 2005). Poulli, et al. states that acidity of olive oil is based on the hydrolytic rancidity; therefore the expected free fatty acids are indicated as linoleic, palmitic and oleic acid. The pure forms of these acids were observed and their synchronous fluorescence spectra with an 80 nm offset value were taken to prove the bands of oleic, palmitic, and linoleic acid. The fluorescence bands of

oleic, butyric and linoleic acids were found as 405 nm, 273 nm, and 325 nm, respectively.

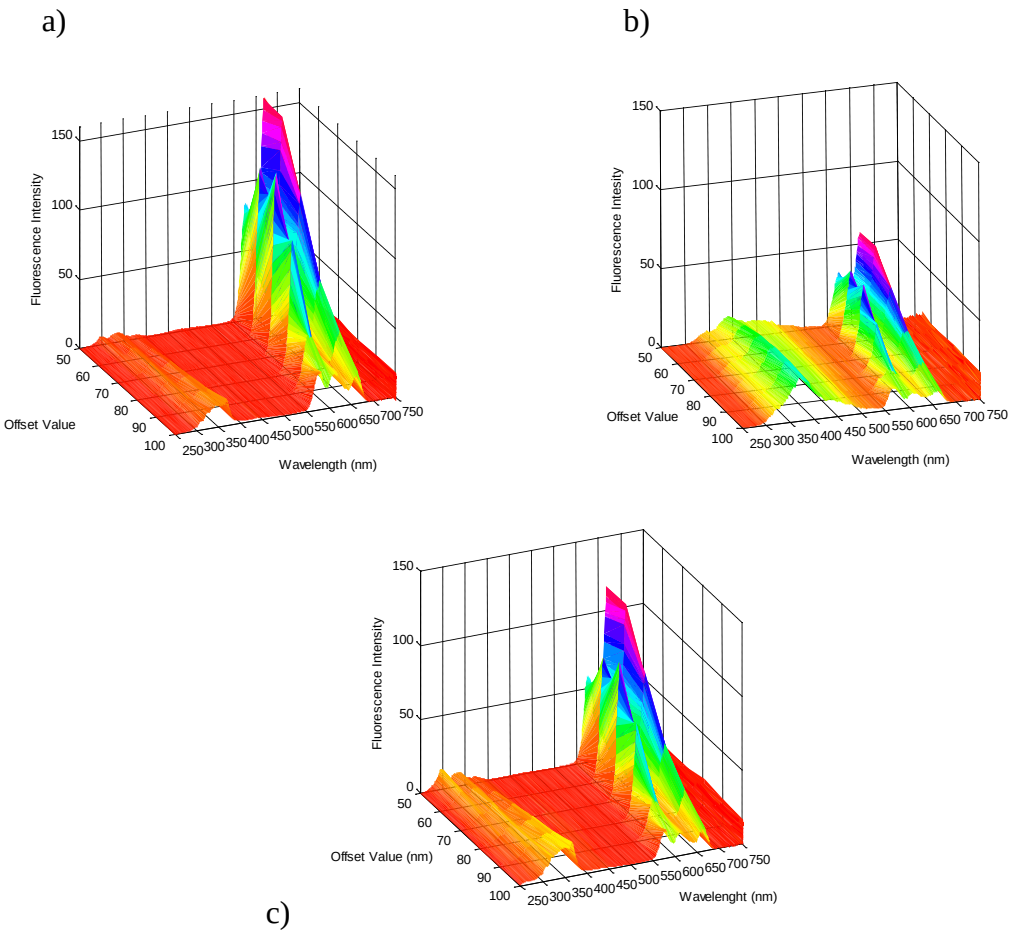


Figure 7.54. Total synchronous fluorescence spectrum of a) extra virgin olive oil, b) lampante olive oil, and c) refined olive oil samples were measured between λ_{em}=250–800 nm at 50 – 100 nm offset values with 10 nm increments.

The total synchronous fluorescence spectra of extra virgin olive oil, refined olive oil and lampante olive oil are shown in Figure 7.54. These spectra show the synchronous fluorescence spectra of oils in the range of 250–750 nm with an 80 nm offset value. The two intense fluorescence bands at 600 and 660 nm are referred the chlorophylls. The weak and large fluorescence bands of olive oils at the region of 280–370 nm are exhibited the tocopherol bands. Due to the differences in acidity, origin of olive oils, the intense of fluorescence bands are changed oil to oil. This feature may be helped on the classification of oils according to their quality. In the experimental

studies, total synchronous fluorescence was examine in the range of 250–800 nm with 10 nm wavelength increments of the offset values in the range of 50–100 nm. In the classification studies the whole three-dimensional fluorescence spectra were taken as a data matrix.

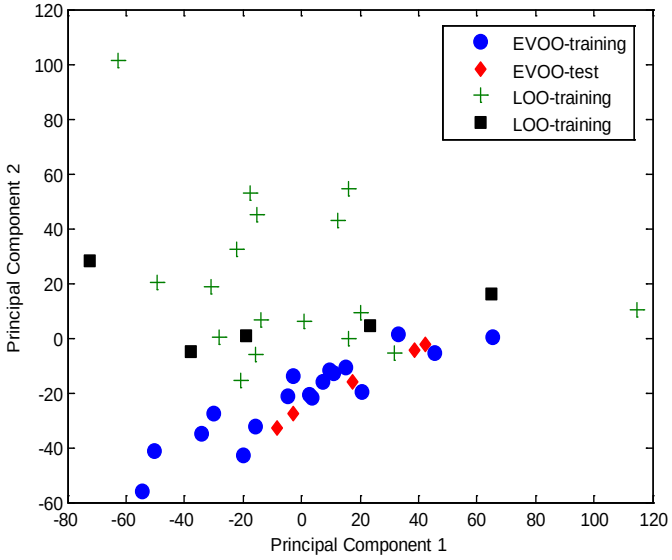
7.5.2.2. Classification Results of SVD-PCA and GAPCAD

7.5.2.2.1. Classification of Extra Virgin Olive Oils and Lampante Olive Oils

To observe the classification of extra virgin olive oil and lampante oil, total synchronous fluorescence spectral data matrix of olive oil samples was performed using singular value decomposition based principal component analysis (SVD-PCA) and distance based genetic algorithm principal component analysis (GAPCAD). Firstly, SVD-PCA was performed to evaluate which classes of olive oils exist in a data set without using any prior information about class memberships. Then GAPCAD was examined and the results were compared.

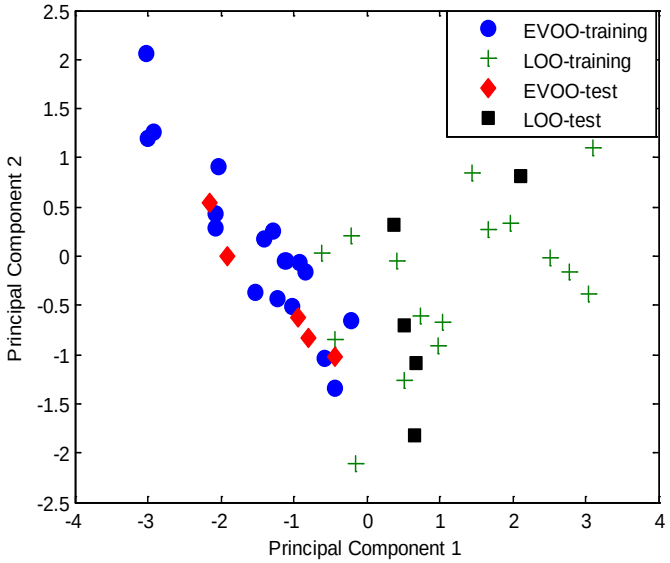
In the examination of SVD-PCA algorithm, the unfolded TSyF spectral data matrix with a 46 x 3306 dimension was constructed by the combination of training and test sample sets. The spectral data matrix was autoscaled before the examination of PCA. Due to the large and intense fluorescence peak of chlorophylls, autoscaled was used as a preprocessing technique. Ten significant principal components explained 85.11% of explained variance. The scores of the first two principal components with a 66.32% of explained variance was plotted to see the distribution of olive oil samples (Figure 7.55.a.). The samples generally lay on principal component 2 and there are no distinguishable two different classes on the score plot.

SVD-PCA



(a)

GAPCAD



(b)

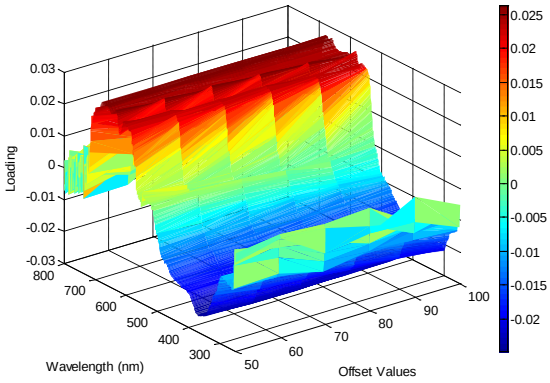
Figure 7.55. Score plot of principal components calculated from unfolded TSyF data matrix of olive oil samples using a) SVD-PCA and b) GAPCAD for whole spectral data (λ = 250–800 nm at $\Delta\lambda$= 50–100 nm with 10 nm increments).

In the examination of GAPCAD, totally thirty six olive oil samples were set into the training set, whereas ten olive oil samples were in the test set. The half of the training samples set includes extra virgin olive oil samples and the other half is

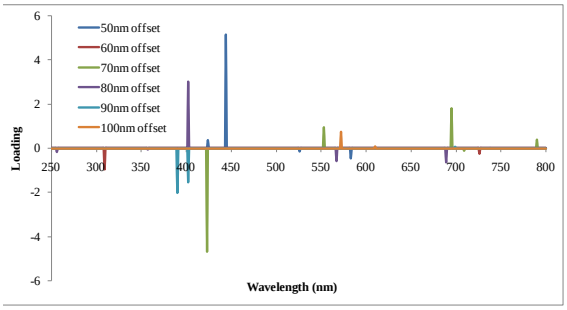
lampante olive oil samples. Test sample set was designed same with training sample set. The whole total synchronous fluorescence spectral data matrix contains 3306 wavelengths with a six different offset values ($\Delta\lambda$=50–100 nm with a 10 nm increments). The algorithm was initiated with 50 genes and 100 iteration numbers. For autoscaled data, the number of significant principal components was three with %93.17 of explained variance. The score plot of first two principal components was used to define the classes of olive oil samples (Figure 7.55.b.)The first two principal components have %68.75 of explained variance. As it is seen from the score plot of principal components, there are two different classes with a few overlapped lampante olive oil samples. Lampante olive oils are generally the mixtures of the virgin olive oils which are not fit to consumption. These olive oils can be used for the refining step or only industrial applications. Therefore the overlapping of lampante olive oil samples can be explained as some olive oil samples have extreme characteristics just like extra virgin olive oil samples or the mixing ratio of virgin olive oil to olive oil can be higher. The distribution of olive oil samples generally are laid on principal component 1. The extra virgin olive oil samples have negative scores whereas lampante olive oil samples have positive scores.

In order to compare the classification results, the loading plots of olive oil samples were plotted to obtain better visualization. The first two principal components which were obtained from SVD-PCA calculation were used to evaluate the classes of olive oils; therefore the loading values of first two principal components were plotted against the refolded TSyF spectral data. The loading plot of PC1 shows both the contribution of chlorophylls (λ=570–670 nm) and tocopherol (λ=350–400 nm) at all offset values. On the other side, the oleic acid (λ=450–500 nm) and tocopherol (λ=350–400 nm) show the most contribution in the classification of olive oil samples. As it mentioned before, the main differences between the extra virgin olive and lampante olive are the amount of tocopherol and the amount of the oleic acid. The fluorescence properties of tocopherol have weighted effects in both PC1 and PC2. Therefore the distributions of olive oil samples in the space are mainly along on PC2.

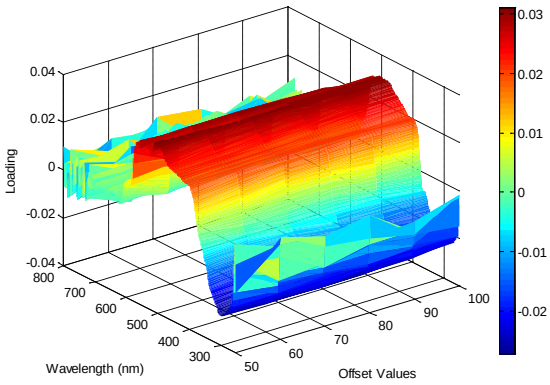
SVD-PCA GAPCAD



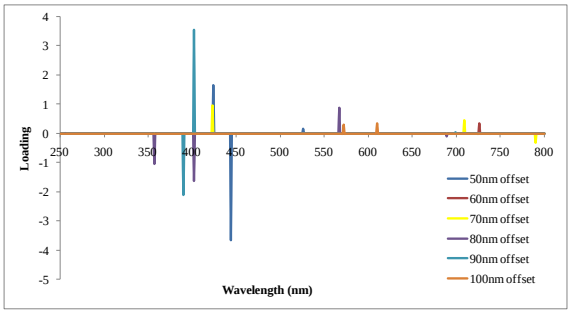
a) Loading plot of PC1 calculated from the whole TSyF spectral data



c) Loading plot of PC1 calculated from the whole TSyF spectral data



b) Loading plot of PC2 calculated from the whole TSyF spectral data



d) Loading plot of PC2 calculated from the whole TSyF spectral data

Figure 7.56. Loading plots of refolded TSyF spectral data (λ = 250–800 nm at Δλ = 50–100 nm).of the olive oil samples (autoscaled data) a) Principal Component 1, b) Principal Component 2 obtained from SVD-PCA, c) Principal Component 1, d) Principal Component 2 obtained from GAPCAD.

The classes of EVOO and LOO samples were generated using totally twenty five wavelengths with their corresponding fluorescence intensity values at different offset values by using GAPCAD calculation. As it is seen from the Figure 7.56.c and Figure 7.56.d, the fluorescence peaks of chlorophylls, oleic acid, and tocopherols were

used in the classification of olive oil samples. The score plot of olive oil samples proves that olive oil samples are mainly along on the PC1.

The weight of loadings that were existed in PC1 comes from the fluorescence peaks of oleic acids. The same result can be also easily seen from the loading plot of PC2. After 100 runs of GAPCAD algorithm, the frequency of selected wavelengths will show the trend of the wavelengths selection in the TSyF whole spectra. According to the Figure 7.57, at all offset values, tocopherols and oleic acid at all offset values (except 50 nm offset value) have the most contribution in the classification of olive oils. On the other hand, the effects of chlorophylls are also seen at lower offset values.

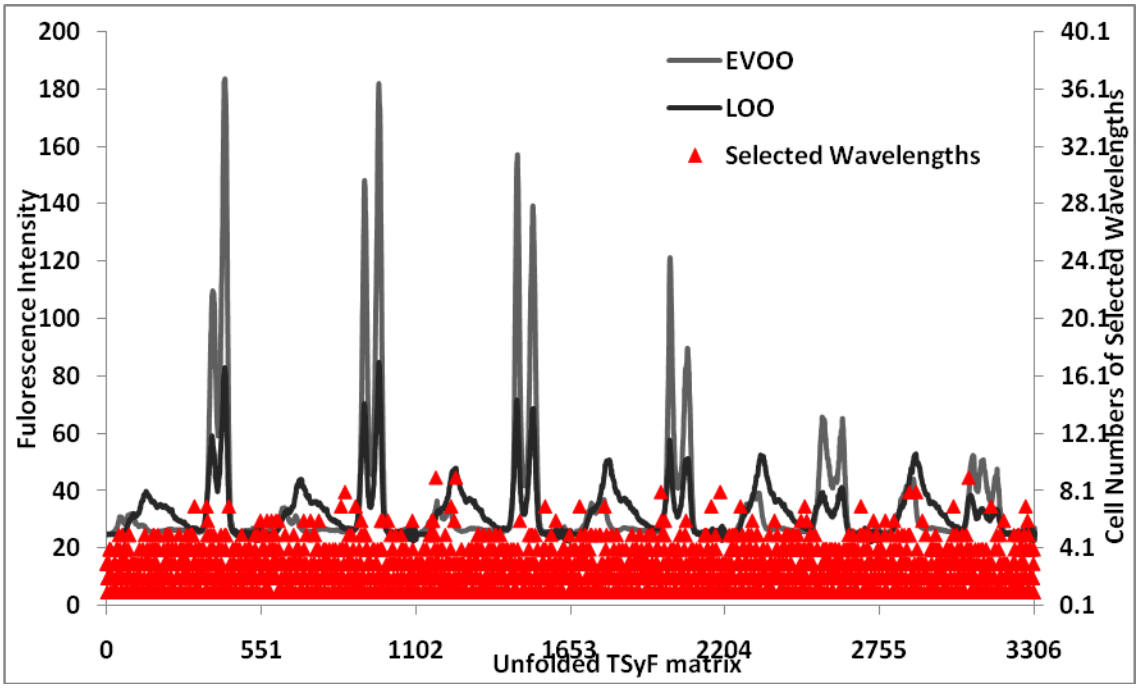


Figure 7.57. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of whole TSyF spectral data.

7.5.2.2.2. Classification of Refined Olive Oils and Lampante Olive Oils

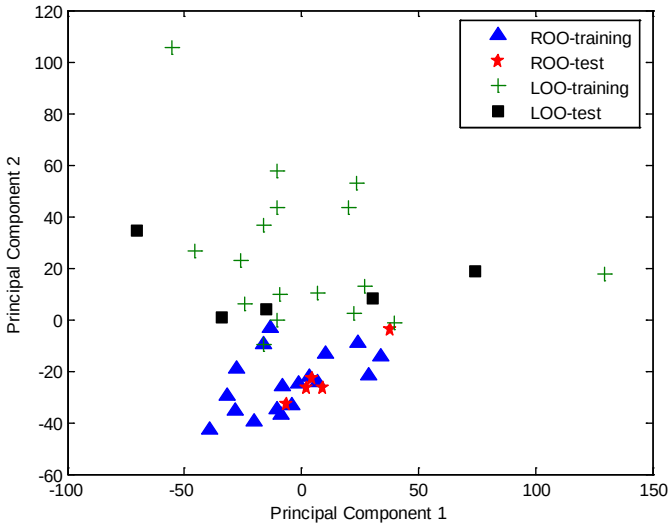
Principal component analysis (PCA) based on singular value decomposition (SVD) was performed as an unsupervised classification technique in order to get prior information on the classification of the lampante olive oil and refined olive oil samples. On the other side, GAPCAD was performed as a supervised technique using two

different samples sets. The training set is used to construct the rules of the classification and then it is tested by an independent sample set named as test set. In the PCA studies, both training and test sample set was combined in order to see the classification results. The concatenated TSyF spectral data matrix with a 46 x 3306 dimension was firstly autoscaled and then the scores of principal components were found. Totally ten significant principal components were found with 85.03% of explained variance. The scores of first two principal components which have totally 64.44% of explained variance were used to identify the classes of olive oil samples.

Figure 7.58.a. shows the scatter plot of PC1 and PC2 which has two different classes in the space. As it is seen from the scatter plot, there are a few overlapped olive oil samples which show similar properties. The score plot principal components shows the distribution of olive oil samples on principal component 2 (PC2) on where the lampante olive oils lay on positive scores and refined olive oils are gathered on negative scores of PC2.

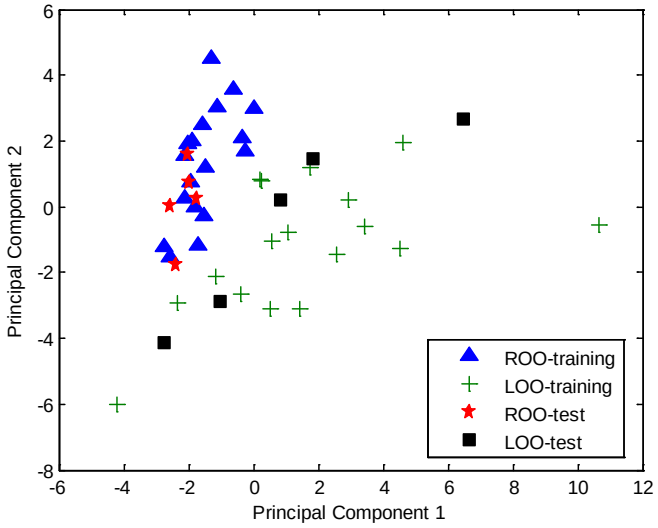
GAPCAD method was also performed to distinguish the refined olive oil and lampante olive oil samples using total synchronous fluorescence (TSyF) spectroscopy. TSyF spectral data matrix contains the wavelengths in the range of 250–800 nm at different offset values which starts 50 nm and ends 100 nm of offset values with a 10 nm increments. Totally 6x551 (Δλ x wavelengths) wavelengths with their corresponding fluorescence intensities were observed in a three-way stacked array. This three-way array was concatenated and at the end 3306 fluorescence points were observed for each sample. Training set contains totally 36 samples in which include 18 of refined olive oil samples and 18 lampante olive oil samples. 10 independent refined and lampante olive oil samples construct the test set. The GAPCAD method was initiated with 50 genes and 100 iterations. Totally seven significant principal components were found with 82.47% of explained variance. The scores of first two principal components which have totally 39.52% of explained variance were used to identify the classes of olive oil samples. Figure 7.58b shows the scatter plot of PC1 and PC2 which has two different classes in the space. Both principal components have similar amount of explained variance, therefore there is no distinguishable principal component in the space.

SVD-PCA



(a)

GAPCAD

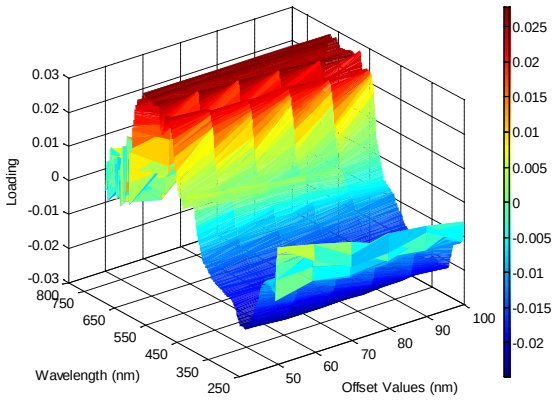


(b)

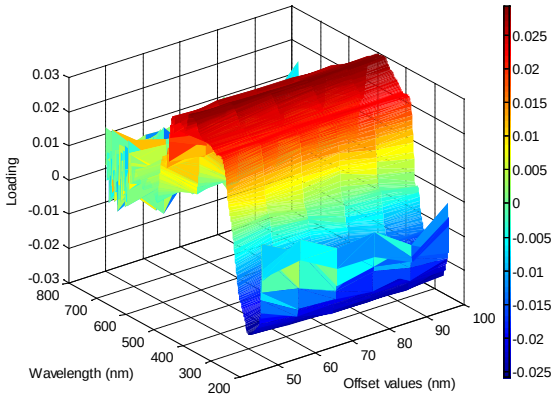
Figure 7.58. Score plot of principal components calculated from unfolded TSyF data matrix of olive oil samples using a) SVD-PCA and b) GAPCAD for whole spectral data (λ= 250–800 nm at $\Delta\lambda$= 50–100 nm with 10 nm increments).

The loading values of first two principal components calculated using SVD-PCA was plotted against refolded TSyF spectra (Figure 7.59.a. and Figure 7.59.b.). The loading values of PC2 prove that the contribution of oxidation products is larger than the PC1. Both principal components have similar distribution on the fluorescence peaks of chlorophylls. The interesting result obtained from the GAPCAD is the selected wavelengths. As it is seen from the Figure 7.59.c and Figure 7.59.d., oxidation products and the oleic acid are selective in the identification of classes of olive oil samples. Refined olive oils have the lowest acidity value among the types of olive oils. Due to the high stability of refined olive oil samples, the content of α-tocopherol is higher than the lampante olive oil samples. This difference can be also seen from the TSyF spectra in the region of 270–450 nm. Generally these fluorescence peaks existing in the region of 270 –450 nm were selected to use in the identification of the olive oil classes. To see the frequency of selected wavelengths, GAPCAD was performed 100 times (Figure 7.60).

SVD-PCA

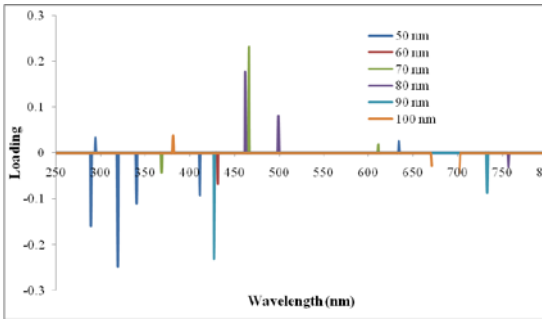


a) Loading plot of PC1 calculated from the whole TSyF spectral data

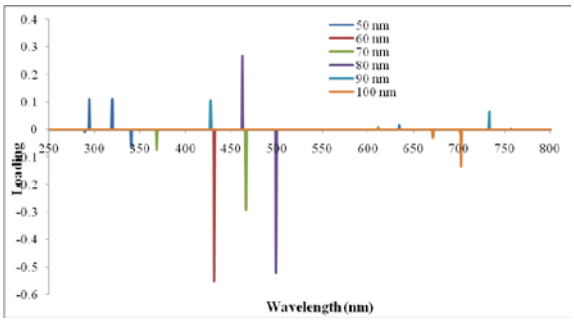


b) Loading plot of PC2 calculated from the whole TSyF spectral data

GAPCAD



c) Loading plot of PC1 calculated from the whole TSyF spectral data



d) Loading plot of PC2 calculated from the whole TSyF spectral data

Figure 7.59. Loading plots of refolded TSyF spectral data (λ = 250–800 nm at Δλ = 50–100 nm).of the olive oil samples (autoscaled data) a) Principal Component 1, b) Principal Component 2 obtained from SVD-PCA, c) Principal Component 1, d) Principal Component 2 obtained from GAPCAD

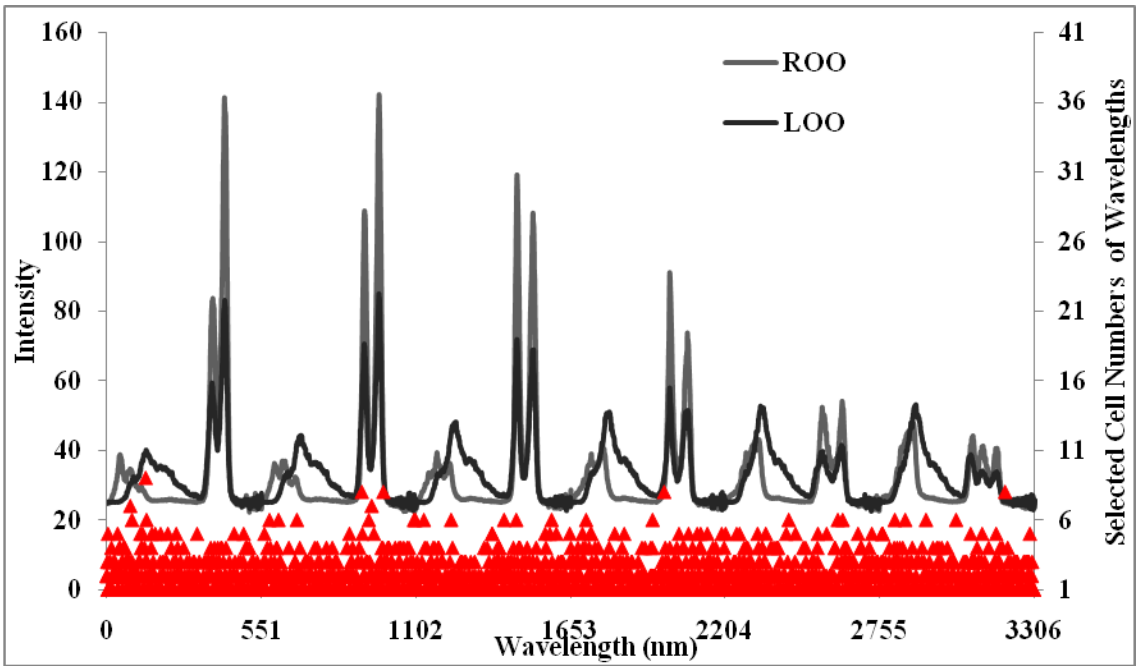


Figure 7.60. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of whole TSyF spectral data.

7.5.2.3. Classification Results of SIMCA and GADA

7.5.2.3.1. Classification Results of Extra Virgin Olive Oils and Lampante Olive Oils

To observe the classification of extra virgin olive oil and lampante oil, total synchronous fluorescence spectral data matrix of olive oil samples was performed using SIMCA and genetic algorithm discriminant analysis (GADA). Firstly, SIMCA was performed and then GAPCAD was examined. After all, the results were compared to get the performance of GADA analysis. In the examination of SIMCA algorithm, the unfolded TSyF spectral data matrix with a 46 x 3306 dimension was constructed by the combination of training and test sample sets. 18 of EVOO samples and 18 of LOO samples were predefined as class 1 and class2, respectively. The remaining samples were built the test set in the testing step of classification analysis.

At the end of the SIMCA analysis, 8 significant principal components with a 88.04% and 90.4% of explained variance were found out. These principal components were used to calculate the distances between the olive oil classes. The critical distances

of each class were obtained as 1.35 at 95% confidence level. Cooman's plot was drawn to visualize the olive oil classes (Figure 7.61). According to the plot, four of EVOO (EVOO-7, 8, 13, 17) samples and one of the LOO (LOO-14) sample show similar properties, whereas LOO-20 ad LOO-22 were found as not belonging to both class of EVOO and class LOO. The remaining samples of training set were correctly classified.

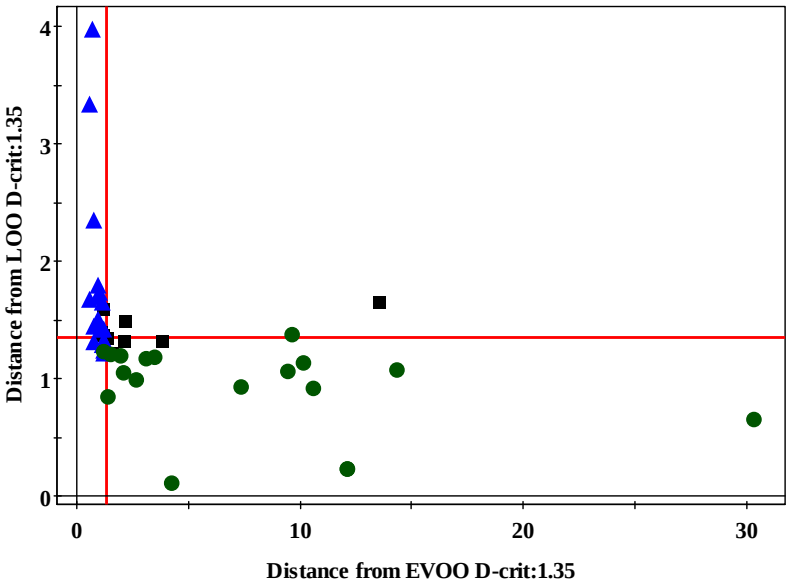


Figure 7.61. Cooman's plot of olive oil samples obtained from SIMCA analysis of TSyF spectra (triangle: EVOO-training, circle: LOO-training, box: test set)

The same training and test sets of unfolded TSyF spectral data matrix of olive oil samples were examined with GADA analysis. The algorithm as initiated with 10 iterations and 10 genes at 95% confidence level. Totally twelve significant principal components were found out with 91.07% of explained variance. Critical limits were calculated as 3.84 for both vertical and horizontal lines.

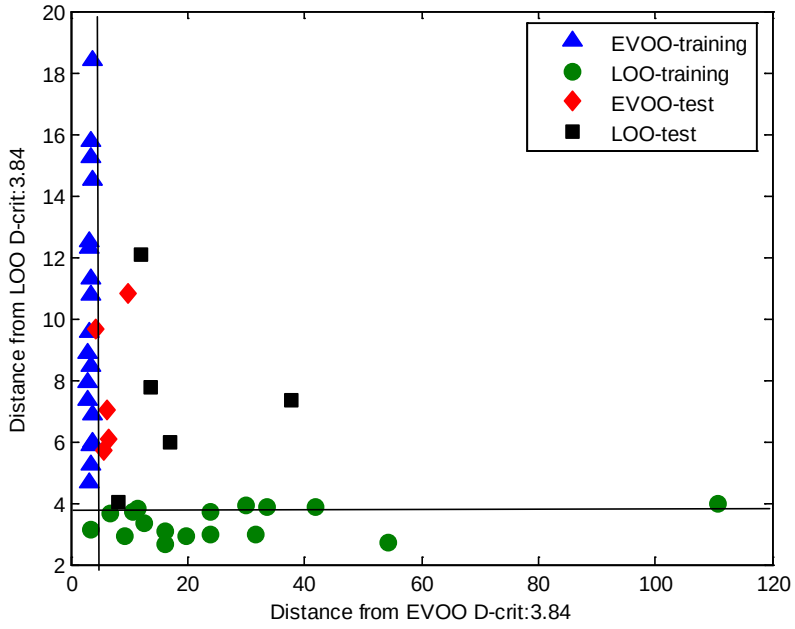


Figure 7.62. Cooman's plot of olive oil samples obtained from GADA analysis of TSyF spectra

According to the Cooman's plot (Figure 7.62) obtained from the analysis of GADA, LOO-14 shows similar properties with extra virgin olive oil samples. There are no extra virgin olive oil samples show similar properties like as LOO-14. It can be concluded that GADA analysis was classified olive oil samples better than SIMCA analysis. However in the test step, the olive oil samples existing in the test set were found as not belonging to the extra virgin olive oil or lampante olive oil samples. Genetic algorithm existing in GADA analysis was chosen the wavelengths which contain the better information for the classification of extra virgin olive oil and lampante olive oil samples. To get the reasons of this classification the selected wavelengths were plotted against the unfolded TSyF spectral data matrix. Totally thirty one wavelengths at different wavelength interval were selected from 3306 wavelengths and they were used to classify the samples. As it is seen from the Figure 7.63, the selected wavelengths generally assign the oxidation products existing in the olive oil samples. As it mentioned before the oxidation products show fluorescence property in the region of 450–500 nm in synchronous fluorescence spectroscopy. In this region the acidity value of lampante olive oil is higher than the acidity of extra virgin olive oil, and also the antioxidant contents of extra virgin olive oil is larger than the lampante olive oil

samples. Therefore lampante olive oil samples show higher fluorescence intensity in that region. GADA generally used these wavelengths in the classification of olive oil samples.

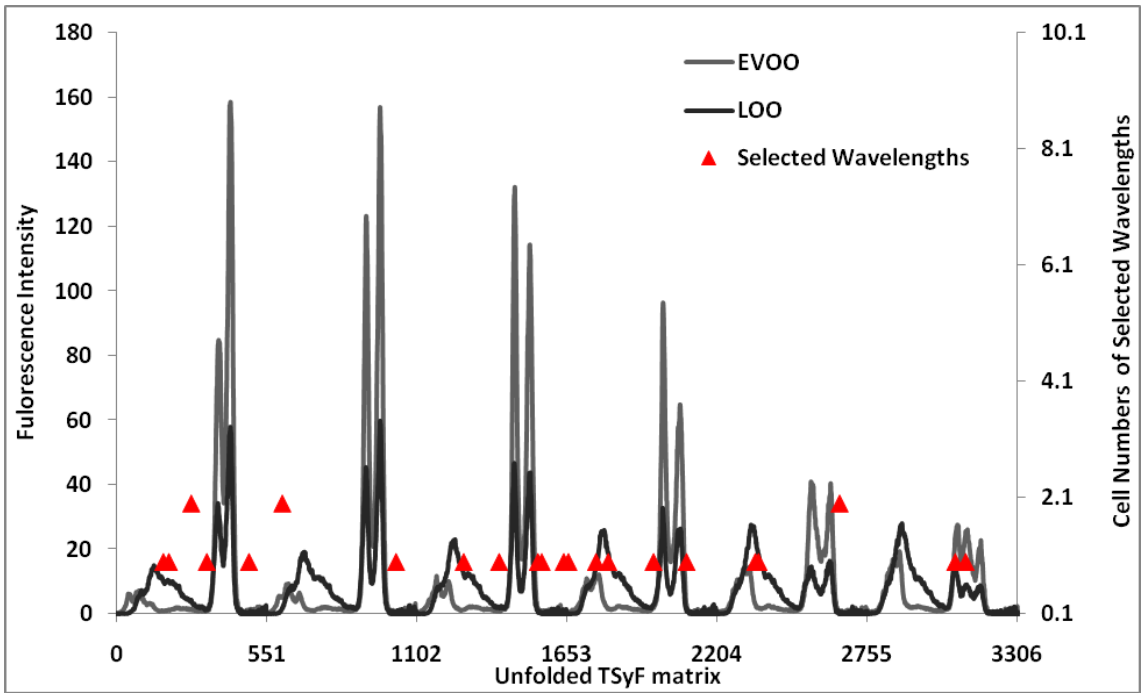


Figure 7.63. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis.

7.5.2.3.2. Classification Results of Refined Olive Oils and Lampante Olive Oils

In the SIMCA studies, both training and test sample sets were predefined in order to see the classification results. The concatenated TSyF spectral data matrix with a 46 x 3306 dimension was firstly autoscaled and then the critical limits of each sample classes were calculated at 95% confidence level. Totally eight significant principal components were found with 88.40% and 90.40% of explained variances for each class. Cooman's plot of ROO (class 1) and LOO (class 2) was drawn to observe the distribution of olive oil samples in the space (Figure 7.64).

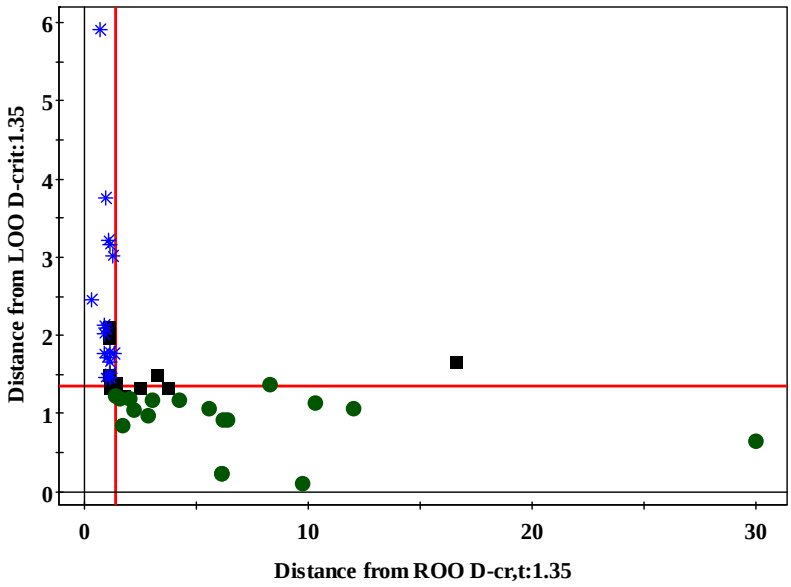


Figure 7.64. Cooman's plot of olive oil samples obtained from SIMCA analysis of TSyF spectra

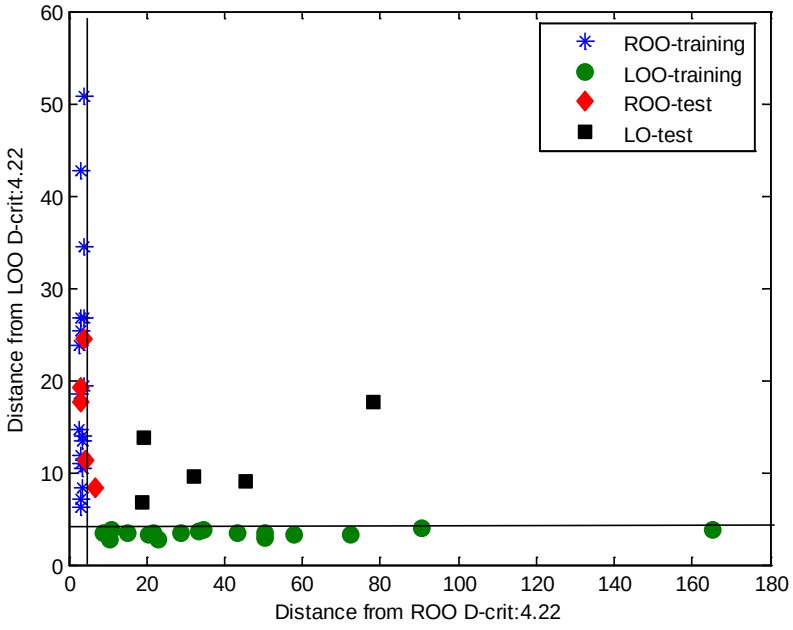


Figure 7.65. Cooman's plot of olive oil samples obtained from GADA analysis of TSyF spectra

The sample coded as LOO-8 is show different properties than the other olive oil samples existing in the training set. On the other side the distribution of lampante olive oil samples have larger variation than the refined olive oil samples. It is the expected results since the acidity value of these olive oil shows diversity. In the test step, the olive oil samples coded as LOO-2 and 4 also do not belong to the either refined olive oil or lampante olive oil.

The same preprocessing technique was also applied to spectral data matrix before starting the GADA analysis. The algorithm was initiated with 6 genes and 10 iteration numbers. Totally 13 significant principal components were found out with a 92.14% of explained variance. The critical limits of each class found as 4.22 at 95% confidence level. The vertical and horizontal lines used at these distances in the constructing of Cooman's plot (Figure 7.65). According to this plot, the classification result of training set is better than the result of SIMCA analysis, whereas in the test step, the samples of LOO and the ROO-4 do not belong to the both classes.

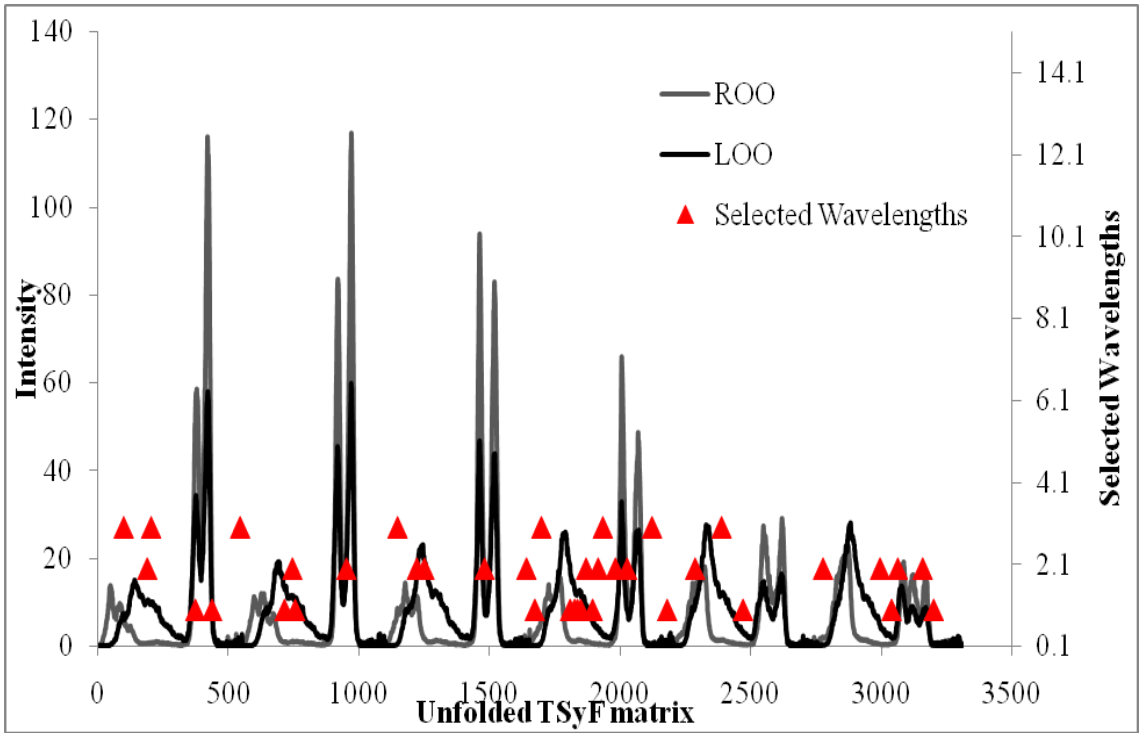


Figure 7.66. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis

Totally seventy wavelengths have different wavelength interval were chosen by genetic algorithm among 3306 wavelengths for this classification. To visualize the selected wavelengths, the plot of selected wavelengths vs. unfolded TSyF data matrix was plotted (Figure 7.66).

As it is seen from the plot of selected wavelengths, 80 nm of the wavelengths interval contains the most selected wavelengths. And also, due to fluorescence intensity of the oxidation products of lampante olive oil in the range of 450–500 nm, this region has the most contribution on the classification of olive oil samples. As a result the GA was selected the wavelengths based on the variety of fluorescence intensity of olive oil samples.

7.6. Classification of Vegetable Oil

7.6.1. NIR Measurements of Vegetable Oils

Three types of vegetable oils were obtained from local markets. These are olive oil, sunflower oil, and corn oil. All types of oils include different trade mark. Totally 34 vegetable oils were analyzed using NIR spectrometer. Typical NIR spectra of vegetable oils are shown in Figure 7.67. As it is seen from the NIR plot of vegetable oils, there are various overlapped peaks. These bands are the results of the overtones and the combinations of fundamental vibrations that occur in the middle infrared region. The absorbance peak existing at 1720 nm is the characteristics of the CH vibration of various chemical groups. This variety depends on the amount of the triglyceride that is analyzed. The second absorption peak in the area of 2143 nm is the characteristic peak of CH vibration of *cis*-unsaturation and it is more intense in polyunsaturated than in monounsaturated fatty acid spectra. Saturated and *trans* fatty acids show weak bands and maxima in the area of 2128–2131 nm. The absorbance peak at 1800nm refers the saturated fatty acids groups.

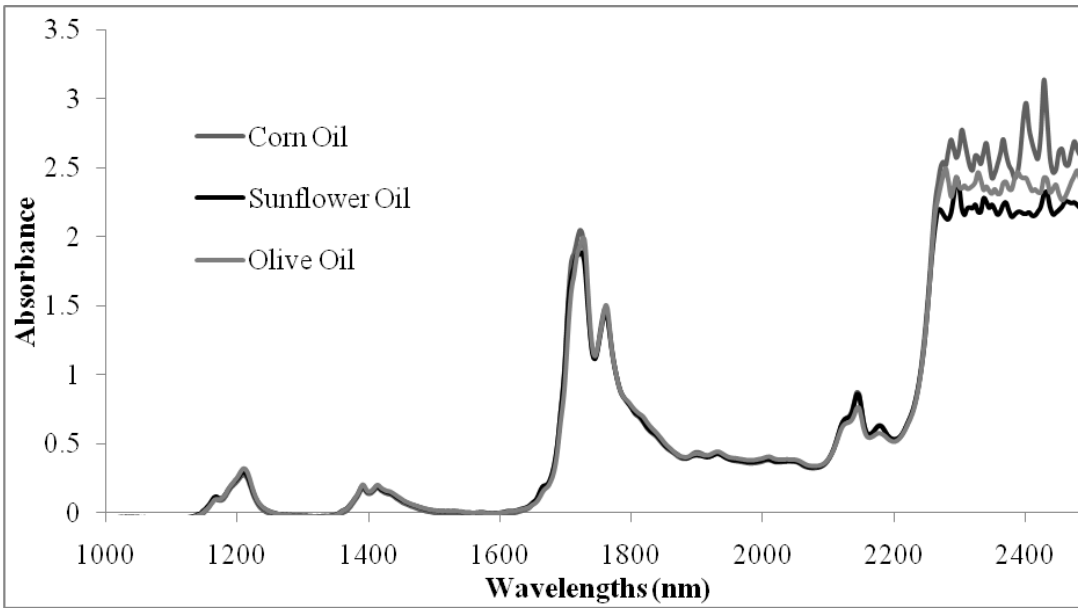


Figure 7.67. NIR spectra of corn oil, sunflower oil, and olive oil.

Oils that have a high amount of polyunsaturated fatty acids have a maximum absorption band at lower wavelengths in the vicinity of 1720 nm. If they are rich in monounsaturated fatty acids, there are two absorbance peaks at 1720 and 2140 nm. Sunflower oil has a maximum intensity nearly 1720 nm; corn oil shows an absorbance peak nearly 1722 nm, and high-oleic sunflower, olive oils have an absorbance peak at 1724 nm (Harwood and Aparicio, 2000). The spectral regions of 1100–1300 and 2050–2230 nm also assign the spectral feature characteristics of vegetable species. Table 7.5 shows the wavelengths that assign a high coefficient of correlation between the intensities of absorbance and the chemical indexes.

Table 7.5. Relevant near-infrared wavelengths (nm) of several lipids and bands that are correlated with some chemical indexes (R>0.90) (Source: Harwood and Aparicio 2000)

	Spectral Regions		
	Second Overtone	**First Overtone**	**Combination**
Tricaprin (C10:0)		1726, 1800	2128
Triolein (cis C18:1)		1725	2143
Trilinolein (cis C18:2)		1665, 1717	2143
Trilinoelaidin (trans C18:2)		1725, 1800	2131
Trilinolenin (cisC18:3)		1665, 1712	2143
MUFA		1724, 1766	2358
PUFA	1162, 1212*	1724, 1766	2136,2176
		1730*	2224,2310
			2348*, 2434*
IV	1164	1664, 1714	2144, 2178
		1740*,1784*	2340*, 2444*

*: Negative correlation coefficient.
MUFA: monounsaturated fatty acids
PUFA: polyunsaturated fatty acids
IV: iodine value

7.6.2. Classification Results of Vegetable Oils Using SIMCA and GADA

NIR spectra of three different vegetable oils were used to design the spectral data matrix. These three types of vegetable oils are olive oil, corn oil, and sunflower oils. The expected classes should be the classes or clusters of these oils. The designed spectral data matrix has a 54 x 780 (samples x wavenumbers) dimensions. The classes of olive oil and corn oil samples were predefined as class 1 and class 2 for the training set, whereas the set of sunflower oil was used a test set. Before starting the classification studies using SIMCA and GADA, the data matrix was autoscaled.

SIMCA analysis was found out 8 principal components (98.60% and 96.80% explained variances for class1 and 2, respectively) for both classes. The critical limits were calculated as 1.38 at 95% confidence level. The Cooman's plot was plotted and the vertical and horizontal limits were used to identify the boundaries of vegetable oil classes (Figure 7.68). These limits prove that the olive oils and corn oils are constructed as separate classes, whereas sunflower oil samples are classified as not belonging to any classes.

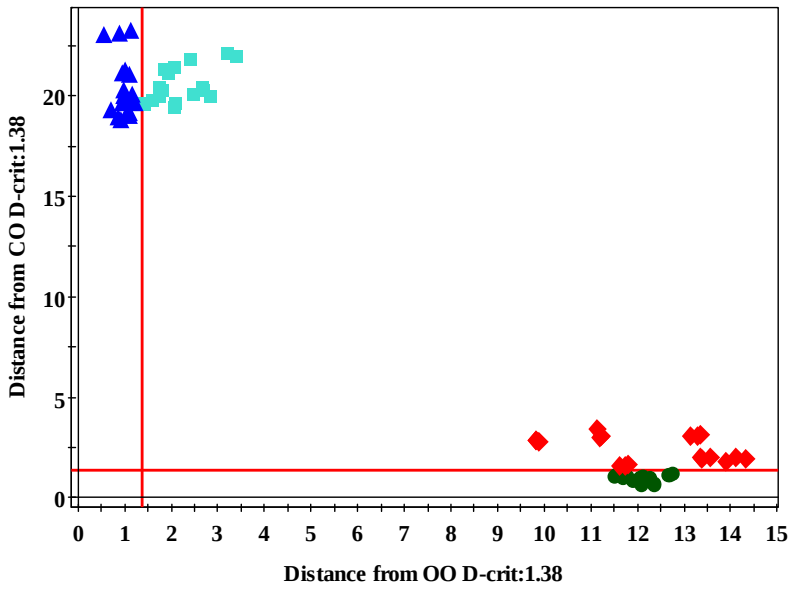


Figure 7.68. Cooman's plot of vegetable oil samples obtained from SIMCA analysis of NIR spectra (triangle: olive oil, circle: corn oil, diamond: sunflower oil, box: olive oil-test)

GADA analysis was initiated with 6 genes and 10 iterations. Totally eleven principal components was found out with a 90.28% of explained variance at the end of the algorithm. The critical limits were calculated as 7.27 at the 95% confidence level. These horizontal and vertical limits were used in the Cooman's plot (Figure 7.69). This plot shows the distribution of vegetable oils in the space. The same results were obtained as they in SIMCA. The main difference between the GADA and SIMCA is the wavelength selection. Since GA is used as wavelength selection toll, whereas SIMCA uses the all the wavelengths that are exist in the spectral data matrix. The selected wavenumbers are plotted against wavenumber region that was used in the measurement of vegetable oils.

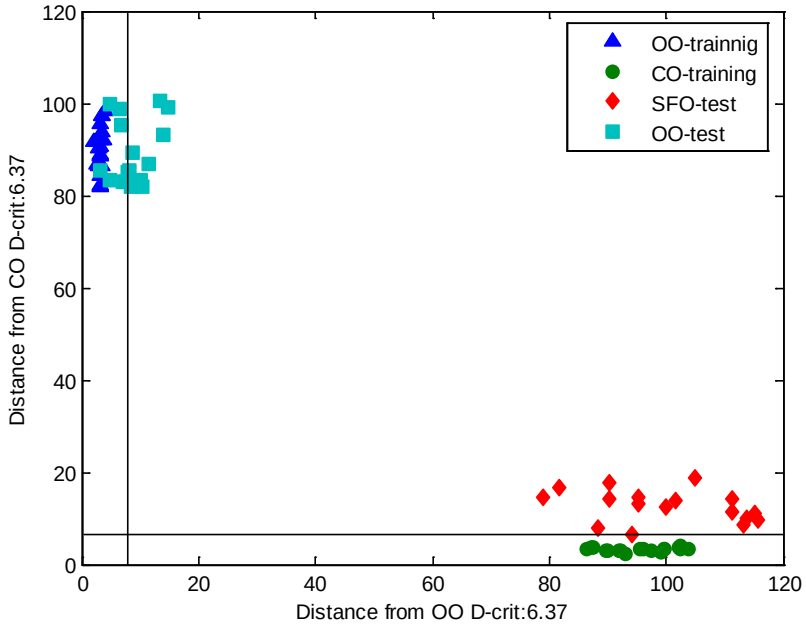


Figure 7.69. Cooman's plot of vegetable oil samples obtained from GADA analysis of NIR spectra

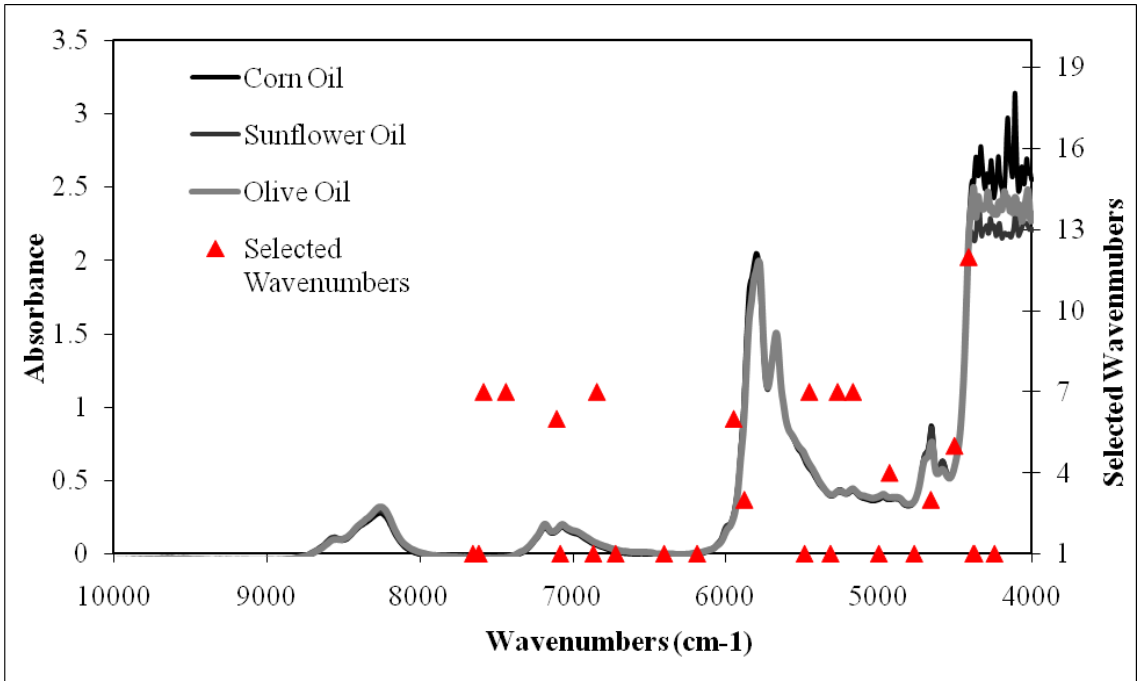


Figure 7.70. The plot of selected wavelengths used in the classification of olive oil samples using GADA analysis.

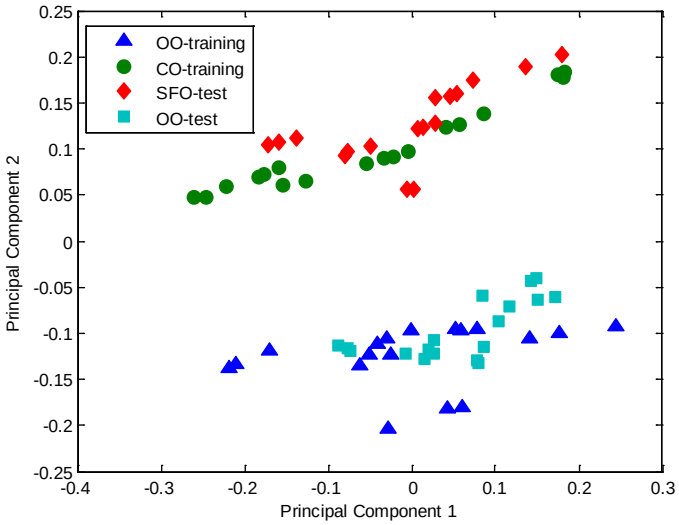
According to the plot of selected wavenumbers, the absorbance peaks of polyunsaturated and monounsaturated fatty acids are generally used in the classification of three types of vegetable oils. Also the intensity of CH groups that existing in the triglycerides have a contribution on the classification of vegetable oils.

7.6.3. Classification Results of Vegetable Oils Using SVD-PCA and GAPCAD

The classification studies of vegetable oils were examined using SVD-PCA and GAPCAD. The spectral data matrix was designed as including the all three types of vegetable oils. The training set was included the olive oil and corn oil samples whereas test set contained only the sunflower oil and olive oil that were different from the samples of the training set. The spectral data matrix of training set and test set with 36 x 780 and 36 x 780 dimensions, respectively. Again, autoscaling was used as a preprocessing technique. The results of both SVD-PCA and GAPCAD were compared to achieve the success of the genetic algorithm based classification method.

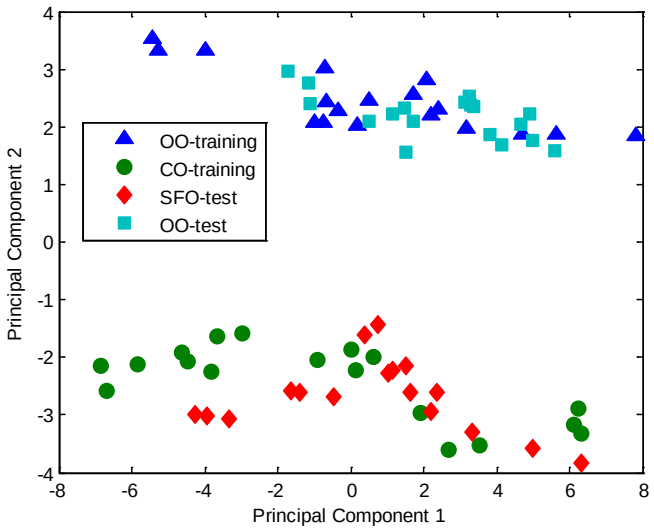
In SVD-PCA calculation, six significant principal components were found out with 57.90% of explained variance. Principal component 1 (PC1) and principal component 2 (PC2) which have totally 36.54% of explained variance were plotted against to observe the distribution of vegetable oil samples. As it is seen from the Figure 7.71.a, there are two different classes. The vegetable oil samples were classified as pulp and seed oil. CO and SFO samples generally lay on the positive scores of PC2, whereas OO samples are distributed on the negative scores of PC2. Same spectral data matrices were examined in GAPCAD calculations. GAPCAD algorithm was initiated with 10 genes and 10 iterations. At the end of the calculation 8 significant principal components (95.50% of total explained variance) were found out. The first two principal components with a 65.80% of explained variance were used to represent the score plot of vegetable oil samples. As it is seen from the Figure 7.71.b, the olive oil samples lay on positive scores of PC2 and the sunflower oil and corn oil samples lay on negative scores of PC2. GAPCAD was also classified the vegetable oil samples as pulp and seed oil.

SVD-PCA



(a)

GAPCAD



(b)

Figure 7.71. Score plot of principal components calculated from NIR spectral data matrix of vegetable oil samples using a) SVD-PCA, b) GAPCAD.

In order to obtain which region(s) or wavenumbers of near infrared region have a contribution on the classification of vegetable oil samples, loading matrices that were obtained from both classification techniques were investigated. The loading values of

first two principal components were plotted against the wavenumber. The more intense value of loadings means the more contribution on the classification of samples. Figure 7.72 show the loading plot of NIR spectral data obtained from SVD-PCA and GAPCAD. Only the first two principal components were taken into the consideration, since the scores plot of vegetable oils were obtained using both of them.

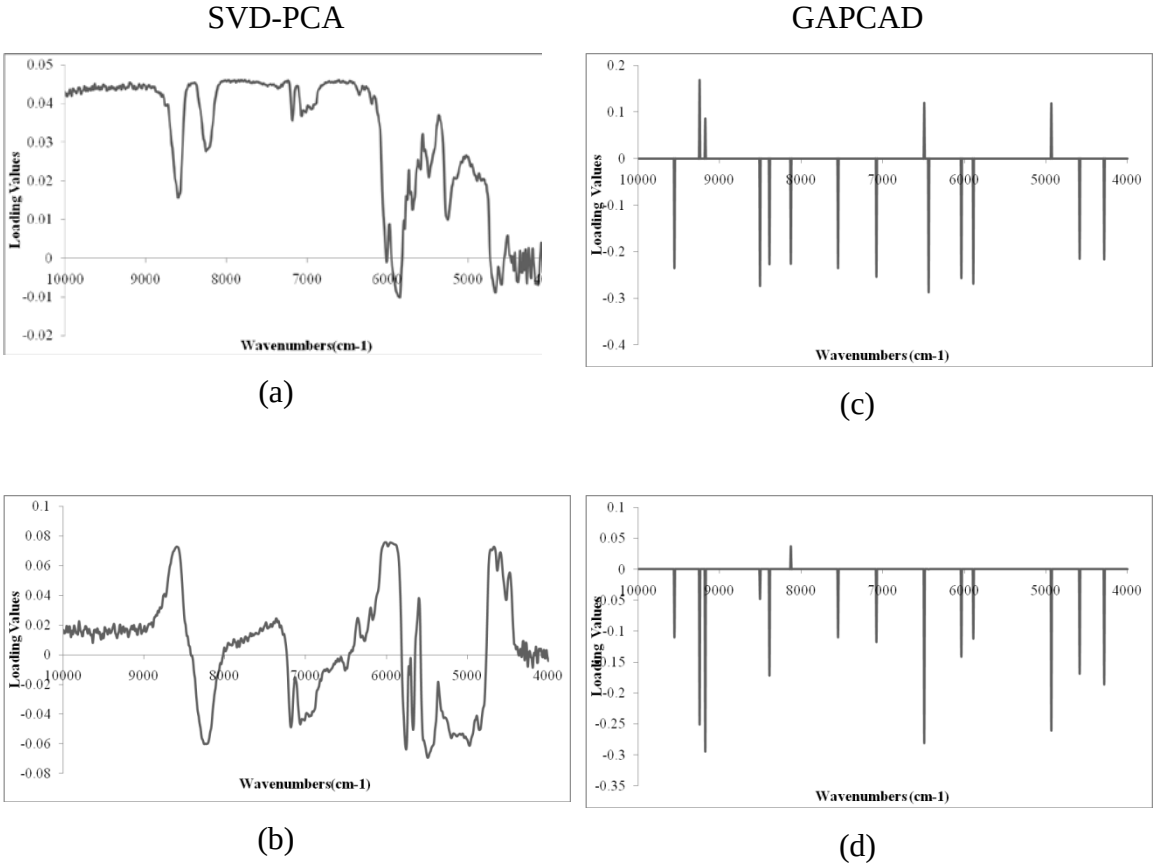


Figure 7.72 Loading plots of NIR spectral data of the olive oil samples (autoscaled data) obtained from the calculation of SVD-PCA a) loading of PC1 b) loading of PC2, and GAPCAD c) loading of PC1, d) loading of PC2.

On the other side, GAPCAD was selected only 25 wavenumbers that have the most contribution in the classification of vegetable oil samples. These wavenumbers are the 9550, 9240, 9170, 8500, 8386, 8124, 7546, 7075, 6489, cm^{-1} and the weight of these wavenumbers changes according to the scores and loadings matrices of PCA. Figure 7.72.c and Figure 7.72.d show the weight of loading for each wavenumbers. According to the scores of GAPCAD, the classification of olive oil samples generally is

along on PC2. Therefore the loading matrix of PC2 has the larger contribution than the loadings of PC1 on the classification of vegetable oils. GAPCAD almost selected the maxima exist in the NIR spectra. As it mentioned before these maxima depends on the amount of the triglyceride, *cis*-unsaturation, saturated and *trans*-fatty acids.

The GAPCAD algorithm was performed 100 times to obtain the frequency of selected wavenumbers and then the frequency of these selected wavenumbers was plotted against the wavenumbers of NIR region. Figure 7.73 shows the frequency of selected wavenumbers after performing 100 times. It is clearly seen that, the maxima around 8000, 7000, and 6000 – 5000 cm^{-1} are the most important region and wavenumbers in the classification of vegetable oils.

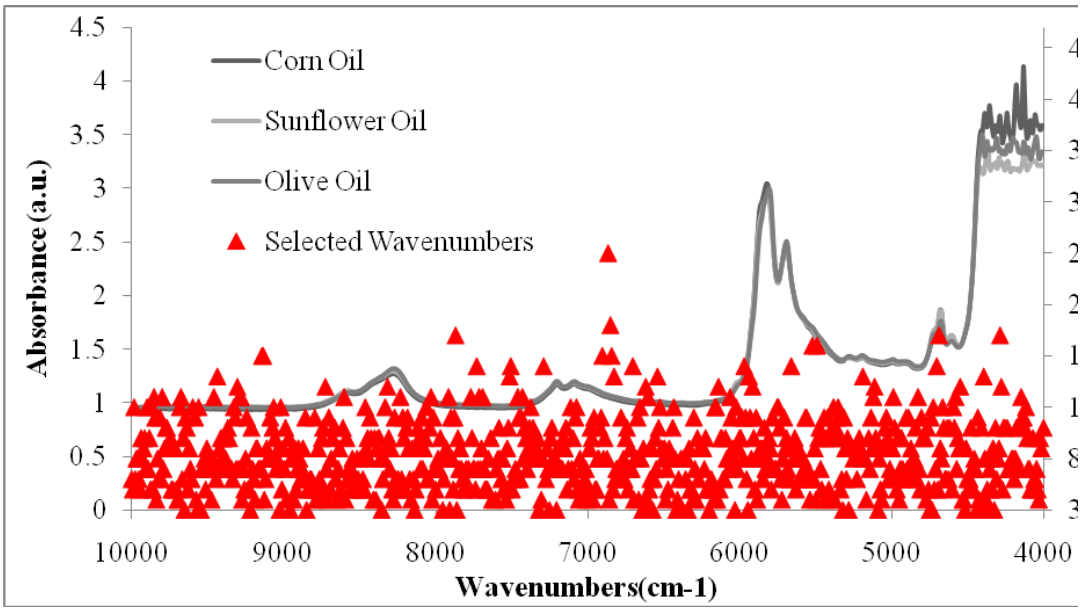


Figure 7.73. Frequency of selected wavelengths after 100 runs of GAPCAD in the examination of NIR spectral data

CHAPTER 8

CONCLUSION

Two different developed genetic algorithm based classification and clustering methods were performed using the spectral data. NIR, FTIR and fluorescence spectroscopic measurements of olive oils and vegetable oils were examined in both developed methods (GAPCAD, and GADA) and the results were compared to SVD-PCA and SIMCA methods.

Two different classifications were aimed in the studies of vegetable oils and olive oils. Vegetable oils were classified according to the pulp and seed oils whereas olive oil samples were classified based on their free acidity values. Olive oil samples were also classified based on the geographical region of olive oil samples using the data matrix of their chemical properties.

In the classification studies, the main advantage of using genetic algorithm based classification techniques instead of principal component analysis or SIMCA is that the output loadings are more interpretable, since the working principle of genetic algorithm based methods is laid on the selection of wavelengths which contains the necessary the information for the separation of olive oil samples.

REFERENCES

Al-Alawi, A., Von de Voort, F.R., Sedman, J. 2004. New FTIR method for the determination of FFA in oils. *Journal of the American Oil Chemists' Society* 81(5):441-446.

Allam, M. A., Hamed, S. F. 2007. Application of FTIR spectroscopy in the assessment of olive oil adulteration. *Journal of Applied Sciences Research* 3(2):102-108.

Aparicio, R., Aparicio-Ruiz, R. 2000. Authentication of Vegetable Oils by Chromatographic Techiques. *Journal of Chromatography A* 881: 93-104.

Beebe, K.R., Randy, J., 1998. *Chemometrics- a practical guide*. John Wiley & Sons, USA.

Bendini, A., Cerretani, L., Di Virgilio, F., Belloni, P., Carbognin, M.B., Lercker, G. 2007. Preliminary evaluation of the application of the FTIR spectroscopy to control the geographic origin and quality of virgin olive oils. *Journal of Food Quality* 30:424-437.

Berrueta, L.A., Alonso-Salces, R.M., Héberger, K. 2007. Supervised pattern recognition in food analysis. *Journal of Chromatography A* 1158:196-214.

Bertran, E., Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., Montoliu, I. 1999. Determination of olive oil free fatty acid by Fourier transform infrared spectroscopy. *Journal of the American Oil Chemists' Society* 76(5):611-616.

Brereton R.G. 2000. Introduction to multivariate calibration in analytical chemistry. *The Analyst* 125:2125-2154.

Burns, D.A., Ciurczak E.W., 2001. *Handbook of Near Infrared Analysis 3rd edition* Practical Spectroscopy Series Volume 35, Marcel Dekker Inc., USA.

Caponio, F., Bilancia, M.T., Pasqualone, A., Sikorska, E., Gomes, T. Influence of the Exposure to Light on Extra Virgin Olive Oil Quality During Storage. *European Food Research and Technology* 221:92-98.

Cheikhousman, R., Zude, M., Jouan-Rimbaud D., Bouveresse, Léger, C.L., Rutledge, D.N., Birlouez-Aragon, I. 2005. Fluorescence spectroscopy for monitoring deterioration of extra virgin olive during heating. *Analytical and Bioanalytical Chemistry* 382:1438-1443.

Christy, A. A., Kasemsumran, S., Du, Y., Ozaki, Y. 2004. The Detection and Quantification of Adulteration in Olive Oil by Near-Infrared Spectroscopy and Chemometrics. *Analytical Sciences* 20:935-940.

Cong, P. and Li, T. 1994. Numeric genetic algorithm part I. theory, algorithm and simulated experiments. *Analytica Chimica Acta* 293:191-203.

Daz, T.G., Merns, I.D., Correa, C.A., Roldn, B., and Cceres, M.I.R., 2003. Simultaneous fluorometric determination of chlorophylls a and b and pheophytins a and b in olive oil by partial least-squares calibration. *Journal of Agricultural and Food Chemistry* 15(24):6934-6940

De Weijer, A.P., Lucasius, T.C.B., Buydens, L., Kateman G., Heuvel, H.M., and Mannee, H., 1994. Curve Fitting Using Natural Computation. *Analytical Chemistry* 66:23-31.

Di Bela, G., Maisano, R., La Pera, L., Lo Turco, V., Salvo, F., Dugo, G. 2007 Statistical Characterization of Sicilian Olive Oils from the Peloritana and Maghrebian Zones According to the Fatty Acid Profile. *Journal of Agricultural and Food Chemistry* 55: 6568-6574.

Downey, G., McIntyre, P., Davies, A.N. 2002. Detecting and Quantifying Sunflower Oil Adulteration in Extra Virgin Olive Oils from the Eastern Mediterranean by Visible and Near-Infrared Spectroscopy. *Journal of Agricultural and Food Chemistry* 50:5520-5525.

Downey, G., McIntyre, P., Davies, A.N. 2003. Geographic Classification of Extra Virgin Olive Oils From the Eastern Mediterranean by Chemometric Analysis of Visible and Near-Infrared Spectroscopic Data. *Applied Spectroscopy* 57:158-163.

Dupuy, N., Duponchel, L., Huvenne, J.P., Sombret, B., Legrand, P. 1996. Classification of edible fats and oils by principle component analysis of Fourier transform infrared spectra. *Food Chemistry* 57(2):245-251.

E. Sikorska, E., Romaniuk, A., Khmelinskii, I.V., Herance, R., Bourdelande, J.L., Sikorski, M., and Koziol, J. 2004. Characterization of edible oils Using Total Luminescence Spectroscopy. *Journal of Fluorescence*, 14(1):25-35

E.K. Kemsley, E.K., 2001. A hybrid classification method: discrete canonical variate analysis using a genetic algorithm. *Chemometrics and Intelligent Laboratory Systems* 55:39–51.

Excitation-emission fluorescence spectroscopy combined with three-way methods of analysis as a complementary technique for olive oil characterization. *Journal of Agricultural and Food Chemistry* 53(24):9319-9328.

F. Guimet,F., Ferré, J., Boqué, R., Rius, X. 2004. Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation – emission matrix fluorescence spectroscopy. *Analytica Chimica Acta* 515:75-85

Firestone, D., 2005. Olive Oil. *Bailey's Industrial Oil and Fat Products (edited by F. Shahidi), Sixth Edition,* John Wiley Inc. 303–331.

Fontain, E. 1992. The problem of atom-to-atom mapping. An application of genetic algorithms. *Analytica Chimica Acta* 265:227-232.

Galtier, O., Dupuy, N., Le Dreau, Y., Ollivier, D., Pinatel, C., Kiser, J., Artaud, J. 2007. Geographic Origins and Compositions of Virgin Olive Oils Determined by Chemometric Analysis of NIR Spectra. *Analytica Chinica Acta* 595:136-144.

Guillén, M. D., Cabo, N. 1997. Characterization of edible oils and lard by fourier transform infrared spectroscopy. Relationships between composition and frequency of concrete bands in the fingerprint region. *Journal of the American Oil Chemists' Society* 74(10):1281-1286.

Guillén, M.D., Cabo, N. Characterization of Edible Oils and Lard by Fourier Transform Infrared Spectroscopy. Relationships between Composition and Frequency of Concrete Bands in the Fingerprint Region. *Journal of the American Oil Chemists' Society* 74:1281-1286.

Guimet, F., Boqué, R., and Ferré 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(22):6673-6679

Guimet, F., Boqué, R., and Ferré J., 2006. Application of non-negative matrix factorization combined with Fisher's linear discriminant analysis for classification of olive oil excitation-emission fluorescence spectra. *Chemometrics and Intelligent Laboratory Systems* 81(1):94-106.

Gunstone, F.D., Vegetable Oils. *Bailey's Industrial Oil and Fat Products (edited by F. Shahidi), Sixth Edition*, John Wiley Inc. 213-267.

Gurdeniz, G., Tokatli, F., Ozen, B. 2007. Differentiation of mixtures of monovariatel olive oils by mid-infrared spectroscopy and chemometrics. *European Journal of Lipid Science and Technology* 109:1194-1202.

Hanagandi, V., and Nikolaou, M., 1998. A hybrid approach to global optimization using a clustering algorithm in a genetic search framework. *Computers Chemical Engineering* 22(12):1913-1925.

Handschuh,S., Wagener, M., and Gasteiger, J., 1998. Superposition of Three-Dimensional Chemical Structures Allowing for Conformational Flexibility by a Hybrid Method. *Journal of Chemical Information and Modeling* 38:220-232.

Hibbert, D.B., 1993. Genetic algorithms in chemistry. *Chemometrics and Intelligent Laboratory Systems* 19:277-293

Hibbert, D.B., Hybrid Genetic Algorithms. 2003. *Data Handling in Science and Technology (Leardi, R. edited)*, Elsevier publishing 23:55-68

Holman, R.T., Edmondson, R. 1956. Near-Infrared Spectra of Fatty Acids and Some Related Substances. *Analytical Chemistry* 28:1533-1538.

Hourant, P., Baeten, V., Morales, M.T., Meurens, M. 2000. Oil and Fat Classification by Selected Bands of Near-Infrared Spectroscopy. *Applied Spectroscopy* 54:1168-1174.

http://www.oliveoillife.com/en/market/html/76.html (accessed May 22, 2009)

http://www.internationaloliveoil.org/downloads/NORMAEN1.pdf (accessed May 22, 2009)

http://cesonoma.ucdavis.edu/hortic/pdf/iocc_standards_purity_grade.pdf (accessed May 22,.2009)

Ingle, J.D., Crouch, S.R., 1988. *Spectrochemical Analysis*. Prentice-Hall Inc. New Jersey.

Iñón, F.A., Garrigueus, J.M., Garrigues, S., Molina, A., De la Guardia, M. 2003. Selection of calibration set samples in determination of olive oil acidity by partial least squares-attenuated total reflectance-Fourier transform infrared spectroscopy. *Analytica Chimica Acta* 489:59-75.

Introduction to Fourier Transform Infrared Spectrometry Thermo Nicolet Corporation 2001.*http://mmrc.caltech.edu/FTIR/FTIRintro.pdf* (accessed May, 22, 2009).

Jee, M., 2002. *Oils and Fats Authentication*. Blackwell publishing.

Karaman, İ., 2008. Prediction of extractives and lignin contents of anatolian black pine (pinus nigra arnold. Var pallasiana) and turkish pine (pinus brutia ten.) Trees using infrared spectroscopy and multivariate calibration. *İzmir Institute of Technology of Ms Thesis.*

Kyriakidis, N.B. and Skarkalis, P. 2000. Fluorescence Spectra Measurement of Olive Oil and Other Vegetable Oils. *Journal of Association of Official Analytical Chemists International* 83(6):1435-1439.

Lakowicz, J.R., 1999. *Principles of Fluorescence Spectroscopy 2nd Edition*. Kluwer Academic Plenu Publishers.

Laszlo, M., and Mukherjee, S., 2007. A genetic algorithm that exchanges neighboring centers for k-means clustering. *Pattern Recognition Letters* 28:2359-2366.

Leardi, R., González A.L., 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent laboratory systems* 41:195-207.

Lucasius, C.B., and Kateman, G., 1991. Genetic algorithms for large-scale optimization in chemometrics: an application. *Trends in Analytical Chemistry* 10(8):254-261.

Lucasius, C.B., and Kateman, G., 1993. Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemometrics and Intelligent Laboratory Systems* 19:1-33.

Luke, B.T., 2003. Genetic algorithms and beyond. *Data Handling in Science and Technology (Leardi, R. edited),* Elsevier publishing 23:1-54.

M. Bockisch, M., 1998. *Fats and oils handbook.* American Oil Chemists' Society.

M. Forina, M., Oliveri, P., Casale, L.M., 2008. Class-modeling techniques, classic, and new, for old and new problems. *Chemometrics Intelligent Laboratory Systems* 93:132-148.

M.A. Grompone, M.A., Sunflower oil. *Bailey's Industrial Oil and Fat Products (edited by F. Shahidi), Sixth Edition,* John Wiley Inc. 655-730.

Maesschalck, R.D., Candolfi, A., Massart, D.L., Heuerding S., 1999. Decision criteria for soft independent modeling of class analogy applied to near infrared data. *Chemometrics Intelligent Laboratory Systems* 47:65-77.

Maesschalck,R.D., Joan-Rimbaud, D., Massart, D.L., 2000. The Mahalanobis Distance. *Chemometrics Intelligent Laboratory Systems* 50:1-18.

Maggio, R.M., Kaufman, T.S., Del Carlo, M., Cerretani, L., Bendini, A., Cichelli, A., Compagnone, D. 2009. Monitoring fatty acid composition in virgin olive oil by Fourier transformed infrared spectroscopy coupled with partial least squares. *Food Chemistry* 114:1549-1554.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J. 1998. *Part A and B: Handbook of chemometrics and qualimetrics*. Elsevier.

Maulik, U., and Bandyopadhyay, S., 2000. Genetic algorithm based clustering technique. *Pattern recognition* 33:1455-1465.

Micklander, E. 2003. Quantitative, qualitative and exploratory analysis of food using spectroscopy and chemometrics. *The Royal Veterinary and Agricultural University of PhD.*

Moreau, R.B., 2005.Corn Oil. *Bailey's Industrial Oil and Fat Products (edited by F. Shahidi), Sixth Edition,* John Wiley Inc. 149-172.

Navas, M.J., and Jiménez, A.M. 2007. Chemiluminescent Methods in Olive Analysis *Journal of American Oil Chemists' Society* 8: 405-411.

Otto, M., 1999. *Chemometrics Statistics and Computer Application in Analytical Chemistry.* John Wiley & Sons.

Ozdemir, D. and Williams, R.R. 1999. Multi-instrument calibration with genetic regression in UV-visible spectroscopy. *Applied Spectroscopy* 53:210-217(8).

Ozdemir, D., Mosley, R.M., and Williams, R.R. 1998a. Hybrid calibration models an alternative to calibration transfer. *Applied Spectroscopy* 52:599-603(5).

Ozdemir, D., Mosley, R.M., and Williams, R.R. 1998b. Effect of wavelength drift on single- and multi-instrument calibration using genetic regression. *Applied Spectroscopy* 52: 1203-1209(7).

Paradkar, R.P. and Williams, R.R. 1996. Genetic regression as a calibration technique for solid-phase extraction of dithizone-metal chelates. *Applied Spectroscopy* 50:753-758(6).

Paradkar, R.P. and Williams, R.R. 1997. Correcting fluctuating baselines and spectral overlap with genetic regression. *Applied Spectroscopy* 51:92-100(9).

Patra, D., and Mishra, A.K., 2002. Recent Developments in Multicomponent Synchronous Fluorescence Scan Analysis. *Trends in Analytical Chemistry* 21(2):787-798

Pontes, M.J.C., Galvão, R.K.H., Araújo, M.C.U., Moreira, P.N.T., Neto, P.D.P., José, G.E., Saldanha, T.C.B., 2005. The successive projections algorithm for spectral variable selection in classification problems. *Chemometrics Intelligent Laboratory Systems* 78:11-18.

Poulli, K.I., Mousdis, G.A., Georgiou C.A., 2005. Classification of Edible and Lampante virgin olive oil based on synchronous fluorescence and total luminescence spectroscopy. *Analytica Chimica Acta*, 542:151-156

Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D., Kuhn, L.A., 1997. Predicting Conserved Water-mediated and Polar Ligand Interactions in Proteins Using a K-nearest-neighbors Genetic Algorithm. *Journal of Molecular Biology* 265:445–464.

Reynés, C., de Souza, S., Sabatier, R., Figuéres, G., Vidal, B., 2006. Selection of discriminant wavelength intervals in NIR spectrometry with genetic algorithms. *Journal of Chemometrics* 20:136-145

Rinnan, A. 2004. Application of PARAFAC on spectral data. *The Royal Veterinary and Agricultural University of PhD.*

Sacchi, R., Mannina, L., Piero, P.F. 1998. Characterization of Italian Extra Virgin Oils Using H-NMR Spectroscopy. *Journal of Agricultural and Food Chemistry* 46:3947-3951.

Sádecká, J., and Tóthová, J., 2007. Fluorescence spectroscopy and Chemometrics in the Food Classification – a review. *Czech Journal of Food Science* 25(4):159-173.

Sato, T. 1994. Application of Principal-Component Analysis on Near-Infrared Specroscopic Date of Vegetable Oils for their Classification. *Journal of the American Chemists' Society* 71:293-298.

Sato, T., Kawano, S., Iwamoto, M. 1991. Near Infrared Spectral Pattern of Fatty Acid Analysis From Fats and Oils. *Journal of the American Chemists'* Society 68:827-833.

Schoefs, B. 2002. Chlorophyll and Carotenoid analysis in food products. Properties of the pigments and methods of analysis. *Food Science and Technology* 13:361-371

Scrimgeour, C., 2005. Chemistry of Fatty Acids. *Bailey's Industrial Oil and Fat Products (edited by F. Shahidi), Sixth Edition,* John Wiley Inc. 1-45.

Sikorska, E., Górecki,T., Khmelinskii, I.V., Herance, R., Sikorski, M., and Koziol J. 2005. Classification of edible oils using synchronous scanning fluorescence spectroscopy. *Food Chemistry* 89:217 – 225

Sikorska,E., Gliszczyka-Wigo, A., Khmelinskii, I.V., and Sikorski M., 2005. Synchronous Fluorescence Spectroscopy of Edible Vegetable Oils. Quantification of Tocopherols. *Journal of Agricultural and Food* C*hemistry* 53(8):6988-6994

Skoog, D.A., Holler, F.J., Nieman, T.A. 1998. *Principles of instrumental analysis – fifth edition*. Philadelphia: Saunders College Publishing, Harcourt Brace College Publishers.

Smith, B.C. 1996. *Fundamentals of Fourier transform infrared spectroscopy*. New York: CRC Press.

Smith, B.M., and Gemperline, P.J., 2000. Wavelength selection and optimization of pattern recognition methods using the genetic algorithm. *Analytica Chimica Acta* 423:167 -177

Stuart, B., 2004, *Infrared Spectroscopy. Fundamentals and Applications*. John Wiley &Sons, West Sussex, England.

Szyk, E., Szydowska-Czerniak, A., Kowalczyk-Marzec, A. 2005. NIR Spectroscopy and Partial Least-Squares Regression for Determination of Natural-Tocopherol in Vegetable Oil. *Journal of Agricultural and Food Chemistry* 53:6980-6987.

Tapp, H. S., Defernez, M., 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.

Tay, A., Singh, R.K., Krishnan, S.S., Gore, J.P. 2002. Authentication of olive oil adulterated with vegetable oils using Fourier transform infrared spectroscopy. *Lebensm.-Wiss. U.-Technology*. 35:99-103.

UNCTAD based on data from the *Report on the proceedings of the 86th session of the International Oil Council* June 2002

Valuer, B., 2001. *Molecular Fluorescence. Principles and Applications*., Wiley-VCH, Weinheim.

Velasco, L., Peréz-Vich, B., Fernandez-Martinez, J.M. 1998. A Rapid and Simple Approach to İdentify Different Sunflower Oil Types by Means of Near-Infrared Reflactance Spectroscopy. *Journal of American Oil Chemists' Society* 75:1883-1888.

Vivo-Truyols, G., Torres-Lapasio, J.R., Garrido-Frenich, A., Garcia-Alvarez-Coque, M.C., 2001a. A hybrid genetic algorithm with local search II. Continuous variables: multibatch peak deconvolution *Chemometrics and Intelligent Laboratory Systems* 59:107–120.

Vivo-Truyols, G., Torres-Lapasio, J.R., Garrido-Frenich, A., Garcia-Alvarez-Coque, M.C., 2001b. A hybrid genetic algorithm with local search: I. Discrete variables: optimisation of complementary mobile phases. *Chemometrics and Intelligent Laboratory Systems* 59:89–106

Vlachos, N., Skopelitis, Y., Psaroudaki, M., Konstantinidou, V., Chatzilazarou, A., Tegou, E. 2006. Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta* 573-574:459-465.

Vo-Dinh, T. 1978. Multicomponent Analysis by Synchronous Luminescence Spectrometry. *Analytical Chemistry* 50(3):396- 401

Yang, H., and Irudayaraj, J. 2001. Comparison of Near –Infrared, Fourier Transform-Infrared, and Fourier Transform-Raman Methods for determining olive pomace oil adulteration in extra virgin olive oil. *Journal of the American Oil Chemists' Society* 78(9):889-895.

Yoshida, H., and Kimito Funatsu, K., 1997. Optimization of the Inner Relation Function of QPLS Using Genetic Algorithm. *Journal of Chemical Information and Modeling* 37:1115-1121.

VITA

BETÜL ÖZTÜRK

e-mail: betulchem@hotmail.com

She was born on 4th August 1976 in Duisburg, Germany. She received her B.Sc. degree from Ege University Faculty of Science Department of Chemistry in 1998. She completed senior project with Prof. Gürel Nişli as an advisor. The title of the project was "Method Development for Recovery of Silver from Used Roentgen Bath Solutions". She achieved her M.Sc. degree from Izmir Institute of Technology Department of Chemistry in 2003 with the project titled as "Monitoring the Esterification Reactions of Carboxylic Acids with Alcohols Using Near Infrared Spectroscopy and Multivariate Calibration Methods" under the supervision of Asst.Prof. Durmuş Özdemir.

She has been working as a research assistant Izmir Institute of Technology Department of Chemistry since starting of M.Sc. studies. She has been working as a teaching assistant in the laboratories; General Chemistry I–II, Analytical Chemistry I–II, Instrumental Analysis, Inorganic Chemistry. She has been working as a research assistant in the laboratories; Chemometrics, Instrumental Analysis and Analytical Chemistry. She has been responsible for the recitation hours of the courses; General Chemistry I–II, Instrumental Analysis, Biostatistics, and Experimental Design and Optimization. She has experience in spectroscopic and chromatographic techniques.

She participated in six international and eight national conferences. She has three publications:

Öztürk B., Arıkan A., Özdemir D.; "Olive Oil Adulteration with Sunflower and Corn Oil Using Molecular Fluorescence Spectroscopy: Determination of Olive Oil Adulteration with Sunflower and Corn Oil"; to be published in the book "Olives and Olive Oil in Health and Disease Prevention" edited by Preedy V.R., Watson R.R., Elsevier Inc.; 2009.

Öztürk B., Özdemir D.; "Near Infrared Spectroscopic Determination of Olive Oil Adulteration with Sunflower and Corn Oil", Journal of Food and Drug Analysis, 2007, 15, 40–47.

Öztürk B., Özdemir D.; "Genetic Multivariate Calibration Methods for Near Infrared (NIR) Spectroscopic Determination of Complex Mixtures", Turkish Journal of Chemistry, 2004, 28, 497–514.