TECHNICAL UNIVERSITY OF CRETE

ELECTRONIC AND COMPUTER ENGINEERING DEPARTMENT

Diploma Thesis

Spectral analysis & classification of oil products using near-IR

Konstantinos Papageorgiou

Advisor: Professor Michalis Zervakis

Examination committee: Associate Professor Nikos Pasadakis

Examination committee: Associate Professor Konstantinos Balas

Chania, December 2012

Preface

This thesis would not have been possible without the guidance and the help of several individuals, who in one way or another contributed and offered their valuable assistance in the preparation and completion of this study.

First and foremost, my utmost gratitude to my advisor, Professor M. Zervakis, who gave me the possibility to complete this thesis with his supervision, advice and guidance from the very early stage of this research, as well as the unparalleled guidance and support of my co-advisor, Associate Professor N. Pasadakis.

I am grateful for their encouragement and precious contribution throughout the elaboration of this study. I would also like to thank them for giving me the opportunity to work on this very interesting field of research.

I would also like to thank Associate Professor K. Balas, for serving in the examination committee of my diploma thesis. Moreover, I would like to thank N. Moirogiorgou, for her patience and constructive comments.

Also, I would like to thank all my friends for these great years we spent together and for many wonderful memories.

Most of all, I would like to thank my family for their enormous help, under-standing and support throughout all these years as a student.

Abstract

The goal of this study is to build predictive models for determining the percentage of aromatic concentration in a large variety of petroleum samples acquired by FTIR spectroscopy in the Mid-Infrared region of the electromagnetic spectrum. These samples are measured also with chemical analytical methods, in order to obtain some reference values of their respective concentration of aromatic hydrocarbons.

We make an effort to build a global prediction model, one that could predict the aromatic concentration of almost every petroleum product. An interactive database is created, in order to store information about samples' spectra data and the chemical analytical methods applied to them.

The analytical multivariate methods applied in this study are the *Principal Component Analysis (PCA), Principal Component Regression (PCR)* & *Partial Least Squares (PLS)* and they presented and described in full detail. These methods are tested and validated in a series of different possible datasets provided from the interactive database created.

Furthermore, a matlab-based program with Graphical User Interface (GUI) is developed, not only to meet the needs of this study, but also to make predictions concerning new samples even easier and more practical in the future.

Contents

1. INTRODUCTION

The first scientific approach carried out regarding optical measurement dates back in the second half of the precedent century [1]. Infrared (IR) spectroscopy has received much attention during the two past decades in various fields by virtue of its advantages over other analytical methods, the most salient of which is its ability to record spectra without the need of any pretreatment for any solid or liquid sample. The cost saving of IR measurements related to improved control and product quality are often achieved and can provide results significantly faster compared to traditional laboratory analysis. Furthermore, it is a non-destructive measurement technique for many chemical compounds which includes the outstanding combination of speed, accuracy, low cost and ease to use justifying why this method has proved its efficiency in several fields such as: agricultural, food/beverage industries, industrial applications, medical diagnostics /pharmaceutical and petrochemical industries.

As regards the petroleum industry, refining and production of products has to meet specific standards. Petroleum products of diesel and naphtha are important products people use in everyday life. However, the standard measurement methods for petroleum products are expensive, time consuming and complicated. For example, to measure the octane number of an oil sample, the most important property in the consumer's mind, a standard ASTM-CFR test engine is needed [2]. The instrumentation required for this needs constant maintenance, is expensive, takes about 20 minutes per sample and is not well suited for on-line applications such as blending operations.

On the other hand, regulations related to these products keep becoming stricter due to environmental and product quality issues. Environmentally, when oil leakage or other accidents happen, it is essential to identify a product as soon as possible in order to proceed with the proper following actions [3]. Over and above, there is always the case of gaining profit by illegal blending. For all of the above reasons the identification and the quality control of the products have to be quick, effective, cheap and easy.

Spectral analysis combined with chemometrics has proved its efficiency in these fields of study. "Chemometrics is a term coined in 1972, which can be defined as the chemical discipline that uses mathematical, statistical and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant chemical information by analyzing chemical data" as described in [4]. In addition, data acquired from spectrometers are sets of great number of exploitable variables for chemometrics (several hundred, even several thousands). Multivariate data analysis is considered the most robust method to treat spectra data. In [5], the advantages of multivariate calibration over univariate calibration are concentrated and can be summarized as follows: (1) Noise reduction of the data, (2) Under the implication of interferents lies significant information that can be handled, (3) Exploratory aspect and (4) Outlier control.

However, in order to build these calibration models with success and in the case of prediction of sample responses, all samples must be acquired under the same conditions, e.g. temperature and humidity control. Moreover, calibration models developed on one spectrometer can only be used for the prediction of new samples, if the instrumental response is identical to the one given by the instrument when the calibration samples were obtained [6].

1.1 INFRARED SPECTROSCOPY

Infrared spectroscopy (IR spectroscopy) is a technique based on the vibrations of the atoms in a molecule. An infrared spectrum is commonly obtained by passing infrared radiation through a sample and determining what fraction of the incident radiation is absorbed at a particular energy (or wavelength). The wavelength at which any peak in an absorption spectrum appears corresponds to the frequency of a vibration of a part of a sample molecule. The spectrometers that measure transitions between different energy levels are divided into transmittance and absorbance spectrometers. In addition, the reflectance spectrometer measures the change of the propagation direction caused by the interaction of radiation with matter

IR spectroscopy deals with the infrared region of the electromagnetic spectrum, which is light with a longer wavelength and lower frequency than visible light [7]. The infrared portion of the electromagnetic spectrum is usually divided into three regions, named for their relation to the visible spectrum: the near-infrared (14000-4000 cm^{-1} or 0.8-2.5 μm), mid-infrared (4000-400 cm^{-1} or 2.5-25 μm) and far- infrared (400-10 cm^{-1} or 25-1000 μm).

The first instrument used for recording spectral information is the *scanning monochromator,* a device that can produce monochromatic light which has found numerous useful appliances in science and in optics. A monochromator can use either the phenomenon of optical dispersion in a prism, or that of diffraction using a diffraction grating, to spatially separate the colors of light. Usually the grating or the prism is used in a reflective mode. The light enters through the hypotenuse face and is reflected back through it, being refracted twice at the same surface [8]. The total refraction, and the total dispersion, is the same as would occur if an equilateral prism were used in transmission mode.

Fourier-transform infrared (FTIR) spectroscopy is based on the idea of the interference of radiation between two beams to yield an interferogram. The latter is a signal produced as a function of the change of pathlength between the two beams. The two domains of distance and frequency are interconvertible by the mathematical method of *Fourier-transformation* [7]. However, in the infrared region of the electromagnetic spectrum FTIR method is more practical than the dispersive because the information at all frequencies is collected simultaneously, instead of dispersive where one wavelength passes through the sample at a time, improving both speed and signal to noise ratio. In addition, one more disadvantage of the dispersive method is that a dispersive measurement requires detecting much lower light than an FTIR measurement. There also are other advantages, as well as some disadvantages but nowadays all modern infrared spectrometers are FTIR instruments [9].

For each wavelength of light passing through the spectrometer, the intensity of the light passing through the reference cell is measured, as I_0. The intensity, I, of the light passing through the sample cell is also measured for that wavelength [10]. The absorbance of the light that the sample has absorbed, with symbol A, is then calculated for that wavelength, as shown at *Equation 1.1.*

$$A = \log_{10} \frac{I_0}{I} \qquad \textit{Equation 1.1}$$

The Beer-Lambert law states that there is a logarithmic dependence amongst the transmission of light through the sample's substance and the product of the absorption coefficient of the substance. The absorption coefficient can be written as a product of a molar absorptivity of the absorber, ε, the molar concentration, c, of absorbing species in the material and the distance the light travels through the material, l, as it can be seen at *Equation 1.2*

$$A = \log_{10}\frac{I_0}{I} = \varepsilon l c \qquad \text{Equation 1.2}$$

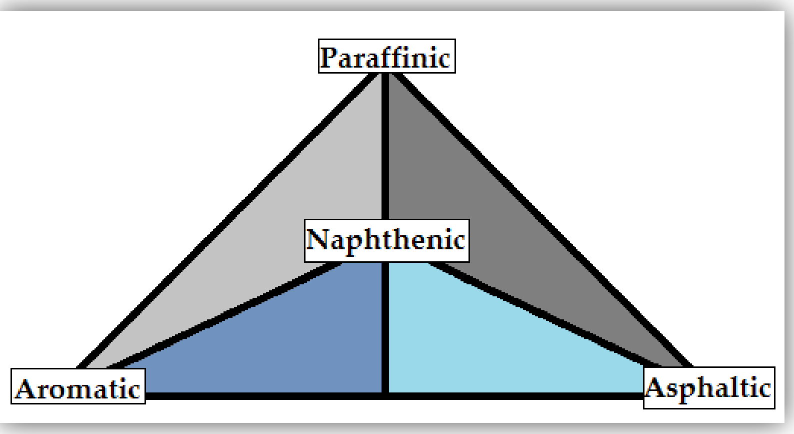
The Beer-Lambert law was first devised independently where Lambert's law stated that absorbance is directly proportional to the thickness of the sample, and Beer's law stated that absorbance is proportional to the concentration of the sample. The modern derivation of the Beer-Lambert law combines the two laws and correlates the absorbance to both, the concentration as well as the thickness (path length) of the sample, as stated at [11].

1.2 PETROLEUM COMPOSITION

Oil is not only the primary energy feedstock used but it's also the basis for the industrial development. Despite the rapid development and progress of renewable and other sources of energy, oil is, and will also be in the near future, the main source of primary energy. Moreover, oil is used in many daily-life applications, serving the needs in: transport (as fuel for diesel engines), heating (as fuel for home and industrial burners), energy (as fuel for electricity generators) and raw material (for the production of lubricants and construction materials such as asphalt and the production of materials from petrochemical industry such as polymers, fertilizers, drugs and others). Furthermore, 90% of the produced oil is used as fuel, covering half of the global energy needs, while 10% as raw material.

Oil is an extremely complex mixture of hydrocarbons and other compounds of carbon and hydrogen containing extra nitrogen, oxygen, metal atoms and sulfur. The chemical composition of oil varies strongly depending on its origin and age. Over and above, according to the mean elemental composition of oil, the element with the highest content is carbon and the next element which abundances over the others is hydrogen [12]. That indicates that oil consists mainly of hydrocarbons. The hydrocarbons found in oil are mainly alkanes, cycloalkanes and aromatics. The alkenes (olefins) are rarely found and alkynes (hydrocarbons of the acetylene series) are even rarer.

Figure 1.1

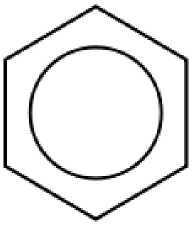


The theoretical triangle which discriminates oil classes.

The oldest and most widespread way to classify oils is to distinguish them by the content of paraffins, aromatics and bituminous constituents. Each petroleum corresponds to a point inside the triangle and it is named analogous its relevant position, as shown at ***Figure 1.1***. Usually this shape is used in conjunction with the separation of petroleum to light and heavy and thus the possible subclasses are doubled.

An aromatic hydrocarbon (or arene or even sometimes aryl hydrocarbon) is a hydrocarbon with alternating double and single bonds between carbon atoms forming rings [13]. The term "aromatic" was derived from the fact that many of the compounds have a sweet scent. The configuration of six carbon atoms in aromatic compounds is a benzene ring, after the simplest possible such hydrocarbon, benzene. Aromatic hydrocarbons can be *monocyclic* (MAH) or *polycyclic* (PAH).

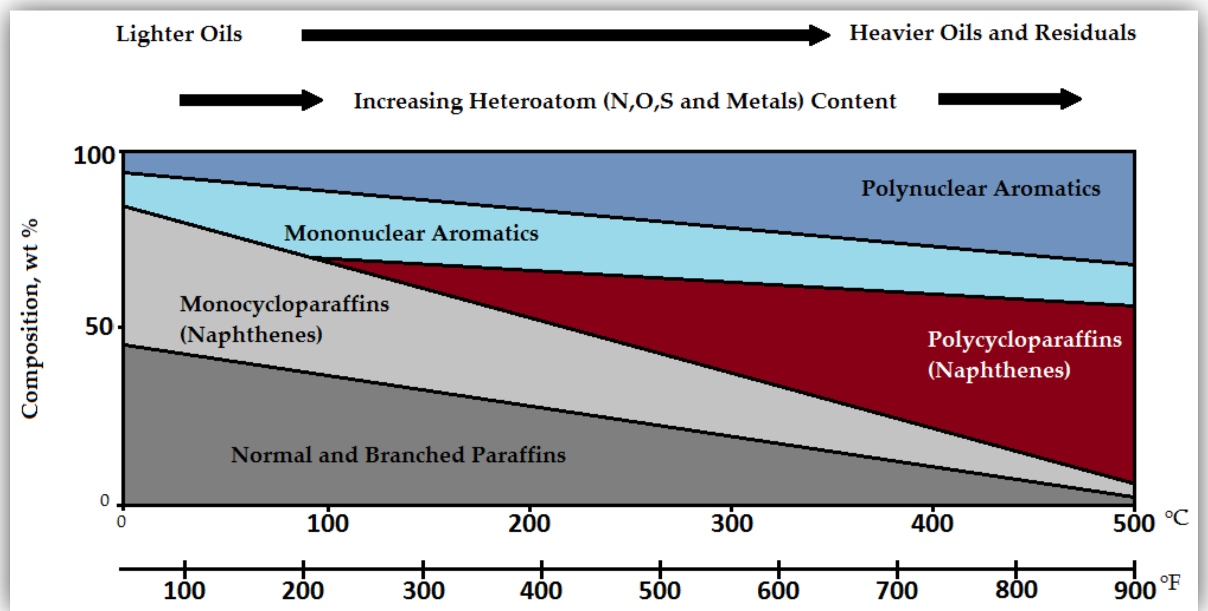
Figure 1.2



A Benzene ring C_6H_6

While the boiling point rises, the content of alkanes, monocyclic naphthenic and MAH is reduced, but the concentrations of PAH and polycyclic naphthenic increase, as shown below:

Figure 1.3



Distribution of the petroleum classes in relation with boiling point [14]

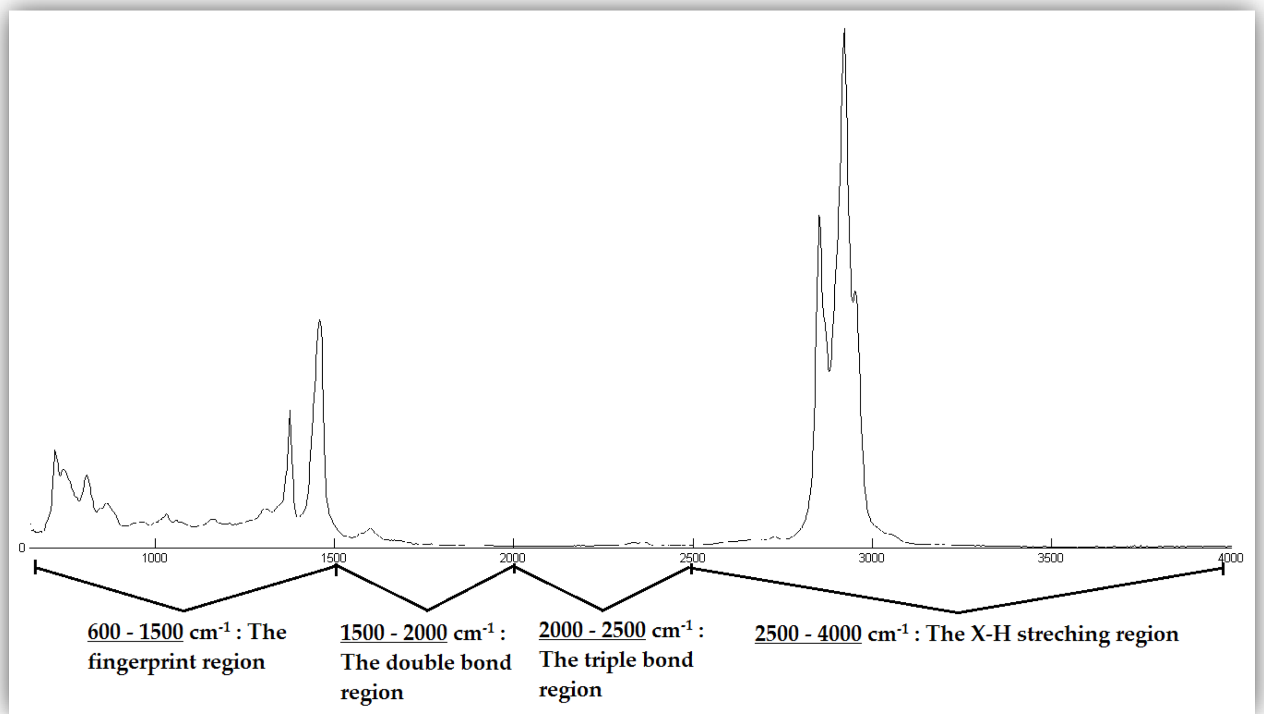
In most oils, the percentage of aromatics hydrocarbons does not overcome 15 % w/w [15]. In petroleum fractions, like gasoline, their high content is desirable because it ensures a higher octane number. On the other hand, when greasing is the targeted field, aromatics are not desired because of the high variation of viscosity with temperature. Moreover, the aromatics hydrocarbons have higher density compared to alkanes and naphthenic.

Furthermore, PAH, contained in fossil fuels or created during their production and use, are considered to be the most hazardous for human. They can be found in air, dust or even in food and they can enter human body through breath, food and skin contact. Polycyclic hydrocarbons are a hole subject to scientific research, if one considers that they are extremely dangerous contaminants with carcinogenic activity.

1.3 Aromatics in MIR

The mid-infrared spectrum (4000-400 cm^{-1}) can be approximately divided into four regions and the nature of a group frequency may generally be determined by the region in which it is located. The regions are generalized as follows: the X-H stretching region (4000-2500 cm^{-}1), the triple-bond region (2500-2000 cm^{-1}), the double-bond region (2000-1500 cm^{-1}) and the fingerprint region (1500-600 cm^{-1}).

Figure 1.4



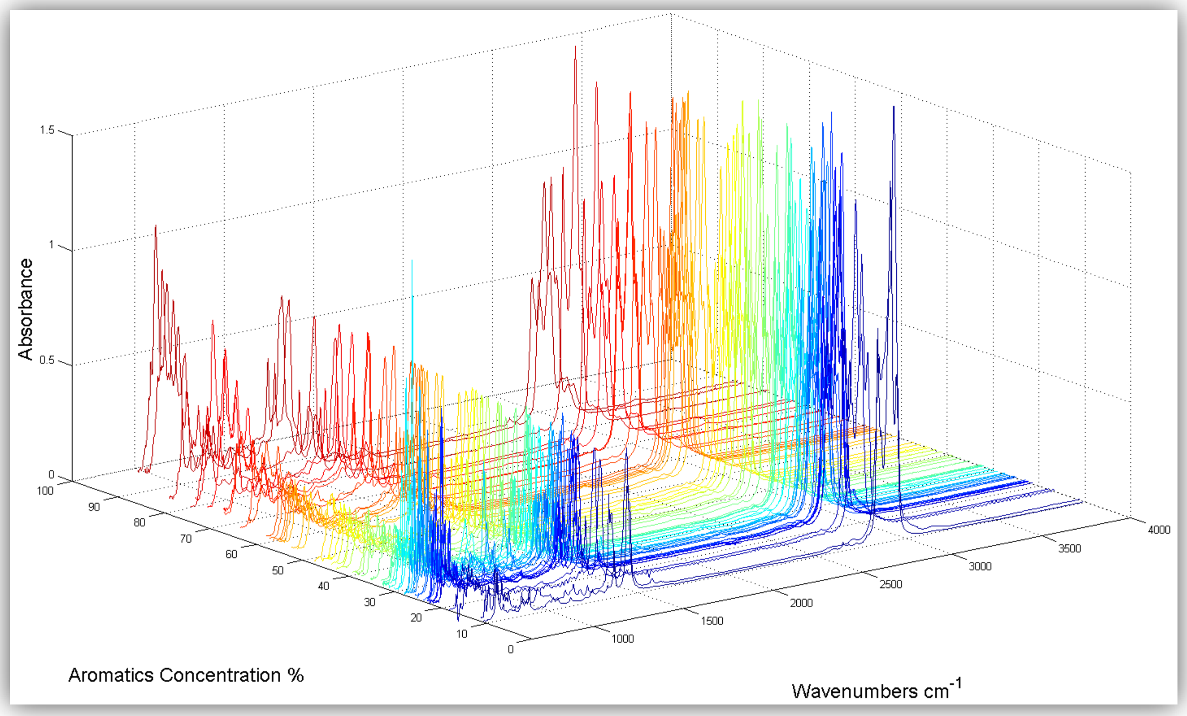
A petroleum sample. All of the four regions of the mid-infrared spectrum are shown in detail.

The fundamental vibrations in the 4000-2500 cm^{-1} region are generally due to O-H, C-H and N-H stretching [9]. O-H stretching produces a broad band that occurs in the range 3700-3600 cm^{-1}. By comparison, N-H stretching is usually observed between 3400 and 3300 cm^{-1}. This absorption is generally much sharper than O-H stretching and may, therefore, be differentiated. C-H stretching bands from aliphatic compounds occur in the range 3000-2850 cm^{-1}. If the C-H bond is adjacent to a double bond or aromatic ring, the C-H stretching wavenumber increases and absorbs between 3100 and 3000 cm^{-1}.

Triple-bond stretching absorptions fall in the 2500-2000 cm^{-1} region because of the high force constants of the bonds. C≡C bonds absorb between 2300 and 2050 cm^{-1}, while the nitrile group (C≡N) occurs between 2300 and 2200 cm^{-1}. On the other hand, the principal bands in the 2000-15000 cm^{-1} region are due to C≡C and C=O stretching. Carbonyl stretching is one of the easiest absorptions to recognize in an infrared spectrum. It is usually the most intense band in the spectrum and depending on the type of C=O bond, occurs in the 1830-1650 cm^{-1} region. C=C stretching is much weaker and occurs at around 1650 cm^{-1}, as C=N stretching does in this region but usually stronger.

It has been assumed so far that each band in an infrared spectrum can be assigned to a particular deformation of the molecule, the movement of a group of atoms, or the bending or stretching of a particular bond [7]. This is possible for many bands, particularly stretching vibrations of multiple bonds that are 'well behaved'. However, many vibrations are not so well behaved and may vary by hundreds of wavenumbers, even for similar molecules. This applies to most bending and skeletal vibrations, which absorb in the 1500-650 cm−1 region, for which small steric or electronic effects in the molecule lead to large shifts. A spectrum of a molecule may have a hundred or more absorption bands present, but there is no need to assign the vast majority. The spectrum can be regarded as a 'fingerprint' of the molecule and so this region is referred to as the *fingerprint region*.

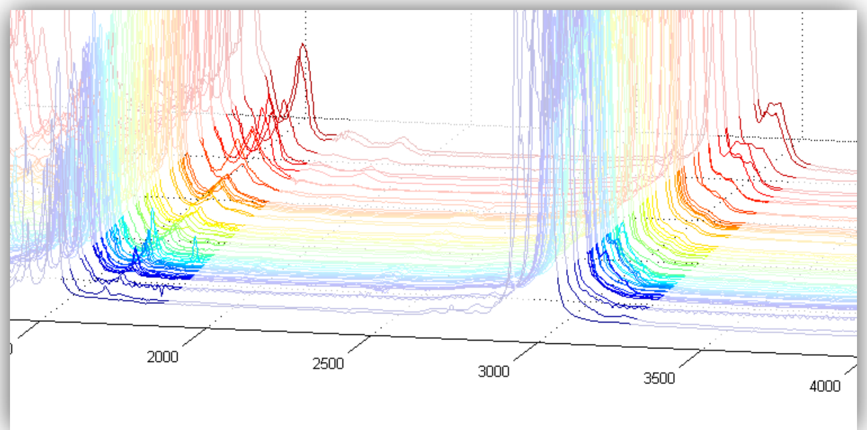
Figure 1.5



58 Petroleum Samples sorted by aromatic concentration.

These are 58 FTIR samples measured in mid-infrared range. The X-H stretching region (4000-2500 cm^{-1}), the triple-bond region (2500-2000 cm^{-1}), the double-bond region (2000-1500 cm^{-1}) and the fingerprint region (1500-600 cm^{-1}) reflect different properties but true analysis arises out of the full spectrum wavelength region. If the previous plot, at ***Figure 1.5***, is rotated and maximized, we can discriminant the regions around 1600 and 3100 cm^{-1} that the absorbance of aromatics bonds increases when the concentration percentage of aromatics increases as well. This is shown at ***Figure 1.6***.

Figure 1.6



A rotated and zoomed perspective of the ***Figure 1.5****. The regions around 1600 and 3100 cm^{-1} correspond to high absorbance of aromatics bonds.*

On the other hand, the absorptions observed in the near-infrared region (13000-4000 cm^{-1}) are overtones or combinations of the fundamental stretching bands which occur in the 3000-1700 cm–1 region. The bands involved are usually due to C-H, N-H or O-H stretching. The resulting bands in the near infrared are usually weak in intensity and the intensity generally decreases by a factor of 10 from one overtone to the next. The bands in the near infrared are often overlapped, making them less useful than the mid-infrared region for qualitative analysis. However, there are important differences between the near-infrared positions of different functional groups and these differences can often be exploited for quantitative analysis.

When MIR is compared to NIR, both spectroscopic ranges provide vibrational information, however, each one has its independent advantages and disadvantages need to be considered for quantitative analysis. The main difference between both ranges is that absorption in mid-infrared spectroscopy corresponds to fundamental bands of molecular vibrations, whereas absorptions in NIR correspond to overtones and combinations of these fundamental bands [15]. The consequence is that "absorption coefficients are much smaller in the NIR range, which allows light to better penetrate into the matter, but on the other hand the NIR spectrum is much more encumbered because of the abundance or combination and overtone bands" as described in [16]. MIR spectra show clearer and sharper spectral differences among the samples and provide greater spectral features and better spectral resolution, however, these spectra contain more spectral noise and the baselines vary. In contrast, the bands in the NIR are broad and less sensitive compared to MIR, however these bands are more reproducible. In [17] concluded that NIR has better calibration performance over MIR when it comes for the determination of the distillation property of kerosene, while in [16] MIR appears to provide better results (10-14% less prediction errors than NIR) when measuring carbon in soil.

1.4 CHEMOMETRICS

The samples' FTIR spectra data and the pre-measured aromatics' concentrations will be used as input to some multivariate analytical methods, in order to make predictive models. As regards chemometrics, there is a large variety of analytical methods for someone to use. The mathematical techniques that can be used are divided in two major categories; the pre-treatment methods and the calibration methods.

The pretreatment techniques, or the spectra pre-processing methods, are a set of mathematical procedures on spectra data before developing a calibration model. Mathematical pre-treatment of the dataset may reduce the background noise and increase the beneficial signal containing chemical information. The one that is used most commonly is a simple normalization and standardization of the data so the dataset will have centered mean 0.0 and standard deviation 1.0 which leads to a model with a reduced complexity. On the other hand, in [18] is suggested that scaling should not be used when a big part of the spectra do not contain useful information because variables that have more noise than relevant information will get the same importance as the ones with relevant signal.

Moreover, multiplicative scatter correction (MSC) and standard normal variate (SNV) are widely known and used for data pretreatment because they reduce spectral distortions due to scattering [19]. SNV centers and scales each spectrum individually, so each has a mean equal to 0 and standard deviation equal to 1, while MSC depends on the whole spectra set, making it more complex and memory consuming. SNV's effectiveness is shown in [20] & [21], when real-time classification of petroleum products and products quality control decisions need to be made respectively. Although pre-treatments are very helpful, it should be noted that there is always a tradeoff between information loss and noise reduction: when removing scattering effects, the chemical signal may also be reduced [19].

An alternative method for pre-processing spectra data is the use of derivatives, such as Savitzky-Golay derivatives which are the most widespread of all, which corrects the overlapping peaks and removes spectral base line offset and baseline slope. The calibrations resulting from applying derivatives usually require fewer variables and models are considered to be more robust. The Savitzky-Golay algorithm is used in [22] and [23] providing robust and efficient models, while in [24] Savitzky-Golay algorithm is combined with MSC. In addition, another more sophisticated method is orthogonal signal correction (OSC). OSC creates a model to remove any signal orthogonal (perpendicular in vector terminology) to the information of interest, but this method requires having previous reference data and best performance is achieved when most of the irrelevant information is actually orthogonal [19].

On the other hand, when it comes to calibration, principal component regression (PCR) and partial least square regression (PLSR) are the most widespread and successive techniques in chemometrics. PCR and PLS are both alternatives extensions of the multiple linear regression (MLS) theorem [25]. The main concept behind these methods is that the original dataset is transformed to a new less dimensional variable space, in which the main variability of the original dataset is gathered in the first components making it easier for the analyst to model the calibration. Principal component analysis (PCA) is the most widespread and successive method for analyzing and extracting useful information from a dataset and it is in fact the foundation of both PCR and PLS methods.

There are extensions of PLS like modified PLS, hybrid PLS, robust PLS, interval partial least squares (iPLS), backwards iPLS (BiPLS), forward iPLS (FiPLS), moving window partial least squares (MWPLS), (modified) changeable size window partial least squares (CSMWPLS/MCSMWPLSR) or searching combination moving window partial least squares (SCMWPLS), that can be very helpful by increasing original PLS accuracy under different set of spectra data with certain characteristics [26]. Extensions regarding the PCR method also exist, like forward selection PCR (FSPCR) or correlation PCR (CPCR) [27].

In [28], non-linear models, either parametric (polynomial-PLS) or non-parametric (genetic inside neural networks "GINN"), were tested on motor octane number prediction in comparison to classic PCR and PLS linear models using MIR spectra. Although the non-linear methods model the information in a more effective way, the overall predictions are of the same order of the linear models, which suggests that, either nonlinearities are present in the data, both PLS or PCR can effectively model the data.

However, it should be noted that, generally, there is not an optimum choice of method, but it depends on parameters such as: signal type (transmittance, reflectance), characteristics of the samples, the equipment used to record spectra and its properties, the targeted application or even the researcher's unique objective. Last but not least, it should also be noted that there is always a tradeoff between information loss and noise reduction.

1.5 ABOUT THIS STUDY

In this study, PCR and PLS are the chosen methods in order to build the predictive models. As far as the pre-processing of data, the models were tested with both normalized and original data. All the mathematical tools used in this study are described in chapters 2 and 3.

1.5.1 GOAL - CONTRIBUTION

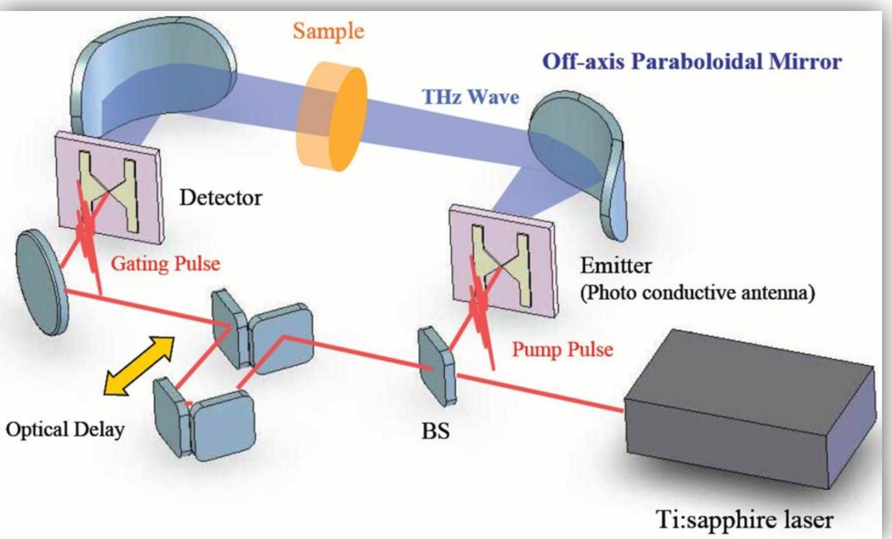
The goal of this study is to build prediction models for determining the percentage of aromatic concentration in a large variety of petroleum samples acquired by FTIR spectroscopy in the Mid-Infrared region. Fuel oils, petroleum residues and distillates, diesel blends, gazolines and a few classes which are strongly (or lightly) consisted of aromatics, like asphalts (or paraffins), are in the samples bunch.

There are many reasons why petroleum industry should be able to identify and characterize the percentage of aromatics in petroleum compounds effectively and quickly. For example, the aromatic hydrocarbon content of aviation turbine fuels affects their combustion characteristics and smoke tendencies. Aromatics constituents also increase the luminosity of the combustion flame, which can adversely affect the life of the combustion chamber. In addition, the aromatic hydrocarbon content of diesel fuels affects the cetane number and exhaust emissions. In products like gasoline, residual fuel oil, naphtha and lubricating oil, the aromatic content is a key property because the aromatic constituents influence a variety of properties including boiling range, viscosity, stability and compatibility with a variety of solutes [29].

1.5.2 SPECTRA MEASUREMENTS

The spectroscopic analysis was carried out on a Perkin-Elmer Spectrum 1000 FT-IR with a deuterated triglycine sulfate (DTGS) detector. Liquid samples were introduced using a horizontal attenuated total reflectance HATR (PIKE Technologies) cell with a zinc-selenide (ZnSe) crystal. For the solid samples (wax, asphalt) a thin film of the substance was formed on the IR cell crystal by introducing the sample dissolved in n-hexane or chloroform respectively, with subsequent evaporation of the solvent under a nitrogen stream. The spectra were acquired in absorbance mode as 20 co-added scans within the range $4000 - 650\ cm^{-1}$ at a resolution of $2\ cm^{-1}$. All the obtained spectra were digitized with a step of $2\ cm^{-1}$.

Figure 1.7



The operating principle of the spectrometer.

Figure 1.8

The Light Sensor of the instrument

Figure 1.9

FTIR Mid-Infrared Instrument of Technical University of Crete

1.5.3 AROMATIC MEASUREMENTS

Every standard analytical method for measuring aromatics, as well as the FTIR spectra measurements, took place in hydrocarbons chemistry & technology research unit at Technical University of Crete.

All the samples that took place in this study were collected from various organizations and oil companies. All of samples have been analyzed using standard methods in order to measure the aromatics percentage. These methods are: ASTM D 2549, HPLC ASTM D 7419-07, ASTM D 1319 and SARA. The term ASTM comes from the "American Society for Testing and Materials", which is an international standards organization that develops and publishes voluntary consensus technical standards for a wide range of materials, products, systems, and services [1] [30].

The chromatographic separations were carried out using an HPLC system consisting of a Waters pump model 600 and a Waters Diode Array Ultraviolet Detector (UV-DAD) model 996. Two analytical columns connected in series, Versapack NH2 4.1x300 mm from Alltech, were used at a column oven temperature of 35 oC. The solvents used as mobile phase were HPLC grade from Labscan filtered through a 20μ membrane filter. A mobile phase constant flow rate of 1 ml/min and of constant composition was used so that a Refractive Index (RI) detector could be included in the analytical system. The mobile phase mixtures were prepared beforehand to avoid any mixing problems due to the low pressure mixing system available.

1.6 CHAPTER DESCRIPTION

In the following chapter "Methods for Analysis", the main analytical multivariate method, PCA, will be described in detail, uncovering all the mathematical tools needed for the signal analysis.

In chapter 3 "Methods for Regression", the most widespread spectra regression methods will be introduced. These methods are the PCR and the PLS and by the end of the chapter it will be clear to someone how a predictive model is created from spectra data.

In chapter 4 "Proposed Methodology", we present the steps and the overall process in order to analyze our data, build predictive models and evaluate the study's results.

In chapter 5 "Results", the results obtained by each method will be shown and evaluated.

In chapter 6 "Conclusion – Further Research", the results obtained by this study, as well as future improvements will be discussed.

In the last chapter «Appendix», a matlab application with user interface which is designed for this study will be presented.

[1] For a complete list of all ASTM D international standards visit:
http://www.astm.org/Standard/alpha-lists/D.html

2. METHODS FOR ANALYSIS

Most, if not all, of the technological and other systems that generate data are multivariate, meaning that any phenomenon that we wish to study in detail depends on many variables. Thus the absorption spectrum of a material, regardless of what scientific perspective one examines it, it is extremely rare that a property of the material to be associated with only a single variable.

Chemical analysis usually consists of two steps. The first step is the calibration process. In this process two different variable groups take place, which are linked by the relation: $Y = f(X)$. The aim is to find a relation between samples X and the response variables Y. The variable group Y containing the response variables is called dependent variables, while the group X is called independent variable. The data used in this step are called *calibration set* or *training set* and the model parameters which are created are called *regression coefficients* or *sensitivities*.

In the prediction step, which will be introduced in the next chapter, the independent variables of one or more samples will be used with the regression coefficients created from calibration process in order to predict values for the dependent variables. In other words, prediction means determining Y values for new X objects, based on a previously estimated (calibrated) X-Y model, thus only relying on the new X data. The data used for this process are called *prediction set* or *test set*.

2.1 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis was introduced by Pearson (19091) and Hotelling (1933) in order to describe the change of a multivariate dataset to a new dataset where the variables are uncorrelated. Nowadays, PCA is a widespread technique that can be implemented in various areas such as face recognition, image compression and detecting patterns in multidimensional data.

PCA is data reduction method, which, for a given dataset, generates an alternative set of parameters so that the data's variability is gathered in the first parameters. Moreover, it provides a way to reduce a complicated (multidimensional) dataset to a new one with fewer dimensions, revealing a simplified structure existing under the large data volume. In essence it is a linear orthogonal transformation which alters the data to a new coordinate system so that the variance of the original data is sorted from highest to lowest. Just for this reason, the descending variance, most of the data's variance is gathered in the first parameters (or components), which means that someone can remove the last ones without losing any useful information. That is why PCA is considered as a dimension reduction method. Furthermore, the models provided from PCA analysis are the ones that PCR method is based on, as we will see in the next chapter.

2.1.1 1ST APPROACH OF PCA: EIGENVECTORS OF THE COVARIANCE

In this section we follow the mathematical proof described in [31]. Let's assume that matrix X is the original data set. The size of the X matrix is $(m * n)$, where n is the number of samples and m the number of variables. The matrix Y, which also has $(m * n)$ dimensions, is connected to matrix X through a linear relation P. Thus we have the matrix X with the initial data and the matrix Y which is another representation of X, linked by the relation:

$$PX = Y$$ *Equation 2.1*

Furthermore, we can define that p_i are the rows of P, x_i are the columns of X and y_i are the columns of Y. *Equation 2.1* is a representation of the change of basis using the matrix P that transforms X into Y. The rows of P,$\{p1, \dots, pm\}$, are the new set of basis vectors to represent the columns of X. Below is the graphical representation of the products in order to become more readily perceived.

$$PX = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} [x_1 \quad \cdots \quad x_n]$$ *Graphical Representation 2-1*

$$Y = \begin{bmatrix} p_1x_1 & \cdots & p_1x_n \\ \vdots & \ddots & \vdots \\ p_mx_1 & \cdots & p_mx_n \end{bmatrix}$$ *Graphical Representation 2-2*

We notice that every column y_i of the matrix Y has the form:

$$y_i = \begin{bmatrix} p_1x_i \\ \vdots \\ p_mx_i \end{bmatrix}$$ *Graphical Representation 2-3*

From all the above we perceive that every y_i variable is a product of the corresponding x_i with the i^{th} row of the matrix P. In other words, the j^{th} coefficient of y_i is a projection on the j^{th} row of P. It is for this reason that we can simply say that the rows of P are a new set of basis vectors representing the columns of X. Moreover, the row vectors of P, $\{p1, \dots, pm\}$, will become the principal components of X, but there is still left for us to find the criteria under P selection.

2.1.1.1 COVARIANCE MATRIX

Consider two sets of measurements with zero mean, as shown below in a simple two variable case:

$$A = \{a_1, a_2, \dots, a_n\} \; and \; B = \{b_1, b_2, \dots, b_n\} \quad ,$$

where A, B are two variables and a_i , b_i are their respective set of values. Moreover, the variance of A and B are separately defined as:

$$\sigma_A^2 = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1} \, , \sigma_B^2 = \frac{\sum_{i=1}^{n}(b_i - \bar{b})^2}{n-1},$$ *Equation 2.2*

where $\bar{a}$, $\bar{b}$ are the mean values of the A and B variables respectively. The covariance between A and B is defined as:

$$\sigma_{AB}^2 = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{n-1}$$ *Equation 2.3*

The covariance measures how much linearly correlated are the two variables. A large value of covariance means great redundancy and beyond that we can recognize that $\sigma_{AB}^2 = \sigma_A^2 \; if \; A = B$ and that $\sigma_{AB} = 0$ if and only if A and B are completely uncorrelated, because $\sigma_{AB}^2 \geq 0$.

By remodeling both A and B to corresponding row vectors

$$a = [a_1, a_2, \dots, a_n]\,, b = [b_1, b_2, \dots, b_n],$$

we are able to express the covariance as a dot product matrix computation:

$$\sigma_{AB}^2 = \frac{1}{n-1} ab^T, \qquad \text{Equation 2.4}$$

where $\frac{1}{n-1}$ is a constant for normalization.

To conclude, we may generalize all the above for many variable sets. Let $x_1 = a,\ x_2 = b$ and additional row vectors $x_3, \dots, x_m$. Now we are able to define a $(m * n)$ matrix X:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \qquad \text{Graphical Representation 2-4}$$

The covariance of matrix X is:

$$C_X = \frac{1}{n-1} XX^T, \qquad \text{Equation 2.5}$$

where the ij^{th} element of C_X is the dot product between the vector of the i^{th} variable and the vector of the j^{th} variable.

The properties of the covariance matrix C_X are in brief:

- The covariance matrix C_X is a square symmetric $(m * m)$ matrix.
- The diagonal terms of C_X are the variance of the particular variables.
- The off-diagonal terms of C_X are the covariance between different variables.

Further up, the matrix C_X contains all the correlations of all the possible pairs of variables. These correlation values reflect the noise and the redundancy in our dataset's variables. Thus, we may rule that:

- In the diagonal terms, by assumption, large (small) values correspond to interesting dynamics (or noise).
- In the off-diagonal terms large (small) values correspond to high (low) redundancy.

2.1.1.2 Solving PCA using the eigenvectors of the covariance

As described above, principal component analysis gathers the information of X based on the variables' variance. So the challenge in finding P is that we don't only wish for a new base to project the initial data, but we also want the new base to gather all the useful information in the first components, i.e. to have high variance of the variables.

After showing how the covariance matrix for X is formed, we can conclude that in order to meet the above objective of PCA, we have to transform the covariance matrix in a way so:

1. The redundancy of the variables covariance is minimized.

2. The signal corresponding to variables variance is maximized.

By definition, the covariance matrix cannot contain any negative values, so the minimum permitted value for a term in the covariance matrix is the zero value. This means that the optimized, for our case, covariance matrix should have zero values at the off-diagonal terms, that is to say the covariance values of the variables, making the covariance matrix a diagonal matrix.

Now that the rationale behind the PCA analysis is explained, we can summarize the following: We are looking for a square vector P so $Y = PX$ and $C_Y = \frac{1}{n-1} YY^T$ is a diagonal matrix. Consequently, if we replace Y from initial *Equation 2.1,* in C_Y, we get:

$$
\begin{aligned}
C_Y &= \frac{1}{n-1} YY^T && \textit{Equation 2.6}\\
&= \frac{1}{n-1} (PX)(PX)^T\\
&= \frac{1}{n-1} PXX^T P^T\\
&= \frac{1}{n-1} P(XX^T)P^T\\
C_Y &= \frac{1}{n-1} PAP^T, && \textit{Equation 2.7}
\end{aligned}
$$

where $A = XX^T$ is a symmetric matrix.

A symmetric matrix is diagonalized by an orthogonal matrix of its eigenvectors. So, for the symmetric matrix A applies:

$$A = EDE^T, \qquad \textit{Equation 2.8}$$

where D is a diagonal matrix and E is a matrix of eigenvectors of A arranged as columns.

In order to accomplish every requested requirement from PCA, we form the P matrix in way so that every row p_i is an eigenvector of XX^T. Thus, with this choice we get $P = E^T$. Substituting in *Equation 2.8* provides: $A = P^T DP$. According to this equation, and because the inverse of a square matrix is equal to its transpose ($P^{-1} = P^T$), we can complete the calculation of C_Y from *Equation 2.7.*

$$
\begin{aligned}
C_Y &= \frac{1}{n-1} PAP^T && \textit{Equation 2.7}\\
&= \frac{1}{n-1} P(P^T DP)P^T\\
&= \frac{1}{n-1} (PP^T)D(PP^T)\\
&= \frac{1}{n-1} (PP^{-1})D(PP^{-1})\\
C_Y &= \frac{1}{n-1} D && \textit{Equation 2.9}
\end{aligned}
$$

It is quite obvious that this choice of P diagonalizes the covariance matrix C_Y, which was our goal from the beginning. To sum up, from all the above processes we accomplished:

- The principal components of the initial matrix X are the eigenvectors of XX^T or the rows of P matrix.
- The i^{th} diagonal term of covariance matrix C_Y is the variance of X along p_i.

Given the above, we proved which choice has to be made for the matrix P in order to achieve the requested purpose of PCA, i.e. to diagonalize the covariance matrix C_Y.

According to all the above, given a matrix X and its dimensions $(n * m)$, to apply PCA analysis the following steps should be followed:

1) Calculate the $(m * m)$ covariance matrix C_X of the initial matrix X:

$$C_X = \frac{1}{n-1} XX^T$$

The covariance matrix C_X now captures all the possible covariance values between all the different dimensions of matrix X. Note that down the main diagonal, the covariance value is between one of the dimensions and itself. These are the variances for that dimension.

2) As mentioned above, we should find the eigenvectors and consequently the eigenvalues of the covariance matrix. Let be:

$$V: Eigenvectors\ of\ X$$

$$D: Eigenvalues\ of\ X$$

Matrix V, containing the eigenvectors of X, captures the information about the covariance values between all the different dimensions of matrix X. Matrix D contains the eigenvalues of X which correspond to the eigenvectors. V is an (m*m) matrix, while D is a diagonal $(m * m)$ matrix

3) The only thing left to do to get the principal components with decreasing order is to sort the diagonal terms of the eigenvalues matrix D.

The eigenvectors V of the covariance matrix are the principal components and the corresponding eigenvalues D represent the amount of variance explained. The eigenvalues correspond to eigenvectors, which provide a linear combination of the original dimensions.

4) Then we must sort the eigenvector matrix V in the same way the eigenvalue matrix D was sorted. By doing this, the first vector of V which corresponds to the greatest value eigenvalue from D matrix, is the first principal component of X, i.e. where the most of the variance is gathered. The second vector of V is the second principal component and so on.

The new axes that form the coordinate system as denoted by the, sorted by decreasing variance, eigenvectors V, represent the directions of decreasing variance in the dataset. Thus, the first axis, called the *first principal component* (or PC1), lies along the direction of maximum variance of the matrix X.

2.1.2 2$^{\text{ND}}$ APPROACH OF PCA: SINGULAR VALUE DECOMPOSITION

A more general solution and perhaps an easier one in computational terms for someone to find the principal components of a matrix X, is to decompose the matrix X according to the Singular Value Decomposition (SVD). In practice, as we get to see in this section, SVD is closely related to PCA and therefore often the names of the two methods are used interchangeably in the literature. Likewise the above section, we also follow here the mathematical proof found in [31].

2.1.2.1 SINGULAR VALUE DECOMPOSITION

Let X be an $(n * m)$ matrix [2]and $X^T X$ a square symmetric $(n * n)$ matrix with rank r. We define that:

- $\{\widehat{v_1}, \widehat{v_2}, \dots, \widehat{v_r}\}$ is a set of orthogonal eigenvectors with size $(m * 1)$ and $\{\lambda_1, \lambda_2, \dots, \lambda_3\}$ are the associated eigenvalues for the symmetric matrix $X^T X$.

$$(X^T X)\widehat{v_\iota} = \lambda_\iota \widehat{v_\iota}$$ *Equation 2.10*

- $\sigma_\iota = \lambda_\iota \sqrt{\widehat{v_\iota}}$ are positive real and termed the singular values.

- $\{\widehat{u_1}, \widehat{u_2}, \dots, \widehat{u_r}\}$ is the set of orthonormal $(n * 1)$ vectors defined by:

$$\widehat{u_\iota} = \frac{1}{\sigma_\iota} X \widehat{v_\iota}$$ *Equation 2.11*

- The sets of vectors $\{X\widehat{v_1}, X\widehat{v_2}, \dots, X\widehat{v_r}\}$ form an orthogonal base onto which every vector $X\widehat{v_\iota}$ has length equal to $\sqrt{\lambda_\iota}$.

Now we have all the required pieces to build the singular value decomposition. *Equation 2.11* can be written equivalently as:

$$X\widehat{v_\iota} = \widehat{u_\iota}\sigma_\iota \ ,$$ *Equation 2.12*

which consists the singular value. We can represent graphically the above equation to be more readily understood:

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \begin{bmatrix} \widehat{v_1} \\ \vdots \\ \widehat{v_m} \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \widehat{u_1} \\ \vdots \\ \widehat{u_n} \end{bmatrix}$$

Graphical Representation 2-5

Moreover, by examining this equation, we realize that multiplying the matrix X with an eigenvector of $X^T X$ equals to the result of multiplication of a scalar with another vector. The set of eigenvectors $\{\widehat{v_1}, \widehat{v_2}, \dots, \widehat{v_r}\}$ and the set of vectors $\{\widehat{u_1}, \widehat{u_2}, \dots, \widehat{u_r}\}$ are both orthogonal set of basis in an r-dimensional space. We can summarize this result for all vectors in one matrix multiplication; the matrix Σ (the matrix containing all the singular values):

$$\Sigma = \begin{bmatrix} \sigma_{\hat{1}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\hat{r}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Graphical Representation 2-6

[2] Notice that at this point we reverse by convention the matrix X from $(m * n)$ to $(n * m)$. The reason for doing so will become clearer in the next section.

In *Graphical Representation 2-6*, $\sigma_{\hat{1}}, \sigma_{\hat{2}}, \dots, \sigma_{\hat{r}}$ are sorted singular values with rank r.

Respectively we can construct the accompanying square matrices V and U:

$$V = [\widehat{v_1}, \widehat{v_2}, \dots, \widehat{v_m}]$$

$$U = [\widehat{u_1}, \widehat{u_2}, \dots, \widehat{u_n}] \quad ,$$

where we appended additional $(m - r)$ and $(n - r)$ orthonormal vectors in order to fill the matrices V and U respectively to the desired dimensions.

As already mentioned, the singular values of Σ matrix are sorted by the rank of the original matrix X and therefore entails that the matrices V and U are sorted in the same order. Every pair of $\widehat{v_\imath}\ and\ \widehat{u_\imath}$ vectors is stored at the i^{th} column of the corresponding matrix. Furthermore the corresponding singular value σ_i is located at the diagonal of Σ matrix at the ii^{th} term. The *Equation 2.13* can be formed, which is easier for someone to perceive by the *Graphical Representation 2-7*:

$$XV = U\Sigma \tag{Equation 2.13}$$

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \begin{bmatrix} v_{1,1} & \cdots & v_{1,m} \\ \vdots & \ddots & \vdots \\ v_{m,1} & \cdots & v_{m,m} \end{bmatrix} = \begin{bmatrix} u_{1,1} & \cdots & u_{1,n} \\ \vdots & \ddots & \vdots \\ u_{n,1} & \cdots & u_{n,n} \end{bmatrix} \begin{bmatrix} \sigma_{\hat{1}} & 0 & 0 & 0 & 0 & 0_{1,m} \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\hat{r}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0_{n,1} & 0 & 0 & 0 & 0 & 0_{n,m} \end{bmatrix} \tag{Graphical Representation 2-7}$$

The dimensions of V and U are $(m * m)$ and $(n * n)$ respectively. The matrix Σ is a diagonal matrix with few non-zero values along its diagonal.

Moreover, because matrix V is square, we can multiply both sides of the *Equation 2.13* with $V^{-1} = V^T$, and finally get:

$$X = U\Sigma V^T \ , \tag{Equation 2.14}$$

which is the final formula of the singular value decomposition for matrix X. This formula indicates that any arbitrary matrix can be transformed to an orthogonal matrix, a diagonal matrix and another orthogonal matrix.

The *Equation 2.14* can be modified like this:

$$\begin{aligned} X &= U\Sigma V^T && \text{Equation 2.14} \\ U^T X &= \Sigma V^T \\ U^T X &= Z \ , && \text{Equation 2.15} \end{aligned}$$

where $Z = \Sigma U^T$.

We should note here that the columns $\{\widehat{u_1}, \widehat{u_2}, \dots, \widehat{u_n}\}$ are now the rows of U^T. If we compare the Equation 2.15 to the Equation 2.1 from PCA, $PX = Y$, we can see that the vectors $\{\widehat{u_1}, \widehat{u_2}, \dots, \widehat{u_n}\}$ perform the same role as the vectors $\{\widehat{p_1}, \widehat{p_2}, \dots, \widehat{p_m}\}$, that is to say U^T is a change of basis from X to Z.

The fact that the orthonormal basis U^T (or P) transforms column vectors means that U^T is a basis that spans the columns of X. Bases that span the columns are termed the *column space* of X. Likewise we can see how the *row space* of X is defined, with SVD:

$$\begin{aligned} XV &= U\Sigma && \textit{Equation 2.13} \\ (XV)^T &= (U\Sigma)^T \\ V^TX^T &= \Sigma U^T \\ V^TX^T &= Z\ , && \textit{Equation 2.16} \end{aligned}$$

where $Z = \Sigma U^T$.

The rows of V^T (or the columns of V) are an orthonormal basis for transforming X^T into Z. Because in this case X is transposed, it means that V is an orthonormal basis that spans the rows of X. Moreover, if someone considers the equation $XV = \Sigma U$ to be a formula of the type $Xa = \kappa b$, where $\alpha = \{\widehat{v_1}, \widehat{v_2}, \dots, \widehat{v_m}\}$ and $b = \{\widehat{u_1}, \widehat{u_2}, \dots, \widehat{u_n}\}$, a can be interpreted as input and b as output. Thus, the column space formalizes the notion of what are the possible 'outputs' for any matrix, while the row space formalizes the notion of what are possible 'inputs' into an arbitrary matrix.

SVD can be helpful at numerous applications and can be analyzed with even more depth, but in order to show the relationship with PCA the above introduction is enough.

2.1.2.2 SOLVING PCA USING SVD

Let's consider the initial $(m * n)$ matrix X from PCA. We can define a new $(n * m)$ matrix Y such as[3]:

$$Y = \frac{1}{\sqrt{n-1}}X^T\ , \qquad \textit{Equation 2.17}$$

where every column of Y has zero mean.

The definition of Y becomes clear by analyzing Y^TY:

$$\begin{aligned} Y^TY &= \left(\frac{1}{\sqrt{n-1}}X^T\right)^T \left(\frac{1}{\sqrt{n-1}}X^T\right) \\ &= \frac{1}{n-1}X^{TT}X^T \\ &= \frac{1}{n-1}XX^T \\ Y^TY &= C_X && \textit{Equation 2.18} \end{aligned}$$

We now have reached the conclusion that Y^TY is equal to C_X, the covariance matrix of X. In the above section, the 1st approach of PCA by eigenvectors, we saw that the principal components of X are the eigenvectors of C_X. If we apply SVD to matrix Y, we will end up with the eigenvectors of $Y^TY = C_X$ stored at the columns of the V matrix. So, in other words, the columns of V are the principal components of X. Moreover, if we decompose Y and Y^T according to SVD formulation, as shown in *Equation 2.14*, the covariance matrix C_X can be expressed as shown in *Equation 2.20*

[3] The matrix Y has now the appropriate dimensions $(n * m)$ for solving with SVD and that is the reason for transposing X in the previous section.

$$\begin{aligned} C_X &= Y^T Y \\ &= (U\Sigma V^T)^T U\Sigma V^T \\ &= V\Sigma U^T U\Sigma V^T \\ &= V\Sigma^2 V^T \end{aligned}$$ *Equation 2.19*

2.1.3 3RD APPROACH OF PCA: NIPALS ALGORITHM

Non-linear iterative partial least squares (NIPALS), is a method to decompose a data matrix X into loadings P' and scores T. Loadings play the role of directions and scores reflect the projections in a more general formulation than PCA. As it is more general, it can be used for solving the PCA formulation. Furthermore, NIPALS can be used for solving regression formulations, as shown in next chapter. The solution presented in this section involves only the data matrix X, while in section *3.3.1*, the algorithm is expanded to handle both predictor data X and response Y matrices.

In this section we follow the mathematical proof found in [25]. According to this approach, our original matrix X with rank r can be expressed as a sum of r matrices with rank 1:

$$X = M_1 + M_2 + \cdots + M_r$$

In addition, these rank 1 matrices can be expressed as an outer product of two vectors, a score vector t_h and a loading vector p'_h:

$$X = t_1 p'_1 + t_2 p'_2 + \cdots + t_r p'_r = \sum_{i=1}^{r} t_i p_i{}'$$ *Equation 2.20*

, or equivalently:

$$X = TP'$$ *Equation 2.21*

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} = \begin{bmatrix} t_{1_1} \\ \vdots \\ t_{1_n} \end{bmatrix} [p'_{1_1} \quad \dots \quad p'_{1_m}] + \begin{bmatrix} t_{2_1} \\ \vdots \\ t_{2_n} \end{bmatrix} [p'_{2_1} \quad \dots \quad p'_{2_m}] + \cdots$$

$$\dots + \begin{bmatrix} t_{r_1} \\ \vdots \\ t_{r_n} \end{bmatrix} [p'_{r_1} \quad \dots \quad p'_{r_m}] = \begin{bmatrix} T_{1,1} & \cdots & T_{1,r} \\ \vdots & \ddots & \vdots \\ T_{n,1} & \cdots & T_{n,r} \end{bmatrix} \begin{bmatrix} P'_{1,1} & \cdots & P'_{1,m} \\ \vdots & \ddots & \vdots \\ P'_{r,1} & \cdots & P'_{r,m} \end{bmatrix}$$

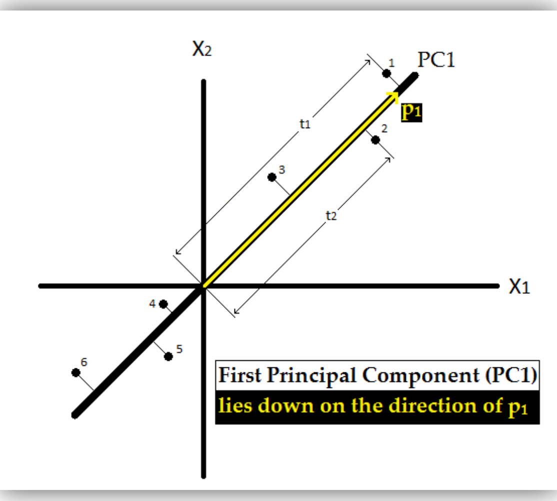
Graphical Representation 2-8

, where P' is created from p' as rows and T from t as columns.

In order to explain the meaning of score t_h and loading p'_h , an example is illustrated below in ***Figure** 2.1* & ***Figure** 2.2*. In the following example we examine a two variable case in a two-dimensional plane; extension for more variables (aka dimensions) is analogous and easy but difficult to show on paper.

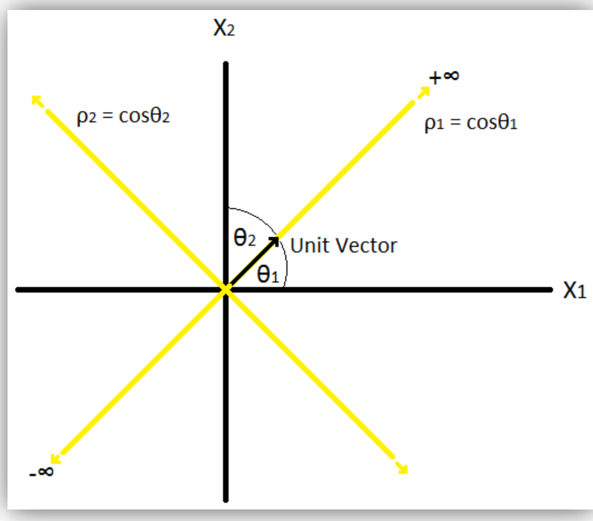
At ***Figure** 2.1* we see the 6 data points. The thick black line among them is the line of best fit, meaning that the sum of squares of x_1 and x_2 residuals in minimized. At ***Figure** 2.2*, the principal component is the yellow line, expanding from $-\infty$ to $+\infty$. In this example the loading p'_h is a $(1 * 2)$ row vector containing the terms p_1 and p_2, which are the direction cosines, or the projections of a unit vector along the principal component on the axes of the plot (yellow lines). The scores t_h is a $(n * 1)$ column vector, having for terms the coordinates of the corresponding points along the principal component line. In regards to this example, it can be easily understood why one wants the length of p'_h to be one [$\cos(\theta_1)^2 + \cos(\theta_2)^2 = \cos(\theta_1)^2 + \sin(\theta_2)^2 = 1$], while in case of more variables (or dimensions) similar rules exist.

Figure 2.1


X2
PC1
p1
t1
t2
1
2
3
4
5
6
X1
First Principal Component (PC1)
lies down on the direction of p1

The coordinates of the projected data points into the line form the score vector t_h

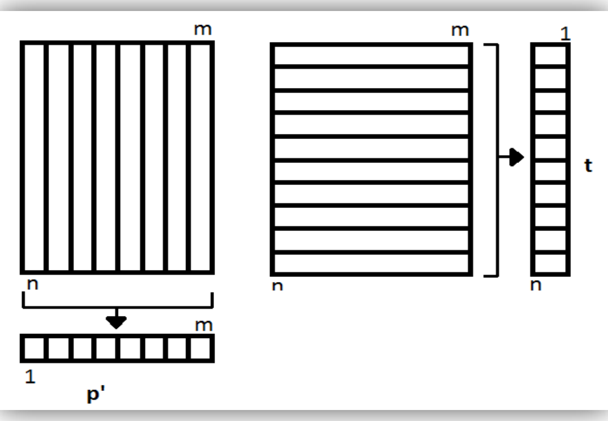
Figure 2.2



The direction of the principal component is the loading vector p_h'

Primarily, we are looking for an operator that can project the columns of X onto a single dimension and another operator which projects the rows of X onto another dimension, as it can be seen at the following figure.

Figure 2.3



Scores and loadings are created by projecting X *into vectors*

The scores and the loading are acquired by projecting X into vectors. For loadings, each column of X is represented by a scalar, meaning that each column of X is projected into an element of the vector p', while for scores each row of X is projected into an element of the vector t represented equivalently by another scalar.

Nonlinear Iterative Partial Least Squares (NIPALS) does not compute all of the t_h and p_h' together at once. The first two terms calculated are t_1 and p_1' from matrix X. Then the inner product t_1p_h' is subtracted from X and the residual E_1 is estimated. This E_1 residual will be used for the calculation of t_2 and p_2', as shown below:

$$E_1 = X - t_1 p'_1, E_2 = E_1 - t_2 p'_2 \dots.$$

$$E_h = E_{h-1} - t_h p'_h \dots E_{\tan_h(X)} = 0 = E_{\tan_h(X)-1} - t_{\tan_h(X)} p'_{\tan_h(X)} \qquad \textit{Equation 2.22}$$

In more detail, the NIPALS algorithm for PCA analysis is as follows:

1) Take a vector x_j from X and call it t_h : $t_h = x_j$
2) Compute p'_h : $p'_h = t'_h X / t'_h t_h$
3) Normalize p'_h : $p'_{h_{new}} = p'_{h_{old}} / ||p'_{h_{old}}||$
4) Compute t_h : $t_h = X p_h / p'_h p_h$
5) Compare the t_h used in step 2 with the one from step 4. If equal (within a very small predefined range), stop (the iteration has converged). If they still differ return to step 2.

At this point, it should be noted, that after the first component is calculated (meaning t_1 and p'_1), X in steps 2 and 4 has to be replaced by its residual E_1. Furthermore, the terms $t'_h t_h$ from step (2) and $p'_h p_h$ are scalars. So we can define them as:

$$t'_h t_h = s_1 \colon scalar$$

$$p'_h p_h = s_2 \colon scalar$$

Moreover, the *equation* from step (4) can be written as:

$$t_h = {}^{Xp_h}/_{s_2} \longleftrightarrow t_h{}' = \left({}^{Xp_h}/_{s_2}\right)' \longleftrightarrow t_h{}' = \frac{1}{s_2}(Xp_h)' \longleftrightarrow t_h{}' = \frac{1}{s_2} p_h{}'X'$$

Now, if we replace p_h from step (2) with $t_h{}'$ from step (4), we get:

$$\left.\begin{aligned} p'_h &= \frac{1}{s_1} t'_h X \\ t_h{}' &= \frac{1}{s_2} p_h{}'X' \end{aligned}\right\} \to p'_h = \frac{1}{s_1}\left(\frac{1}{s_2} p_h{}'X'\right) X \to p'_h = \frac{1}{S} p'_h X'X \to S p'_h = p'_h X'X$$

$$(SI_m - X'X)p_h = 0 \ , \qquad \textit{Equation 2.23}$$

where the scalars s_1 and s_2 are combined in one general scalar S. Equivalently if we replace equation from step 4 into the equation from step (2) we get:

$$(S'I_n - XX')t_h = 0 \qquad \textit{Equation 2.24}$$

The *Equation 2.23* & *Equation 2.24* are actually the equations for the eigenvectors and eigenvalues of $X'X$ and XX' respectively, where I_n is the identity matrix of size $(n * n)$ and I_m is a similar identity matrix of size $(m * m)$.

Considering all the above, we got to the point to prove that, in case of converge, the results of the NIPALS algorithm are the same as the results obtained from the approach that uses the eigenvectors. Moreover, the NIPALS algorithm is convenient when it comes to be implemented in microprocessors and it also is introductory in order to understand the next basic method of spectral analysis, PLS.

Finally, it is worth mentioning that it doesn't matter what approach of the three someone will choose to apply PCA analysis, since NIPALS will always converge, unless if two identical eigenvalues exist. In that case, it also does not matter which combination or rotation of eigenvectors someone will chose.

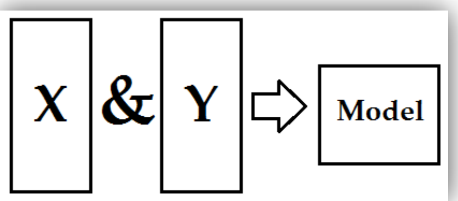
3. METHODS FOR REGRESSION

Multivariate regression, often referred as multivariate calibration, is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning [32]. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

Regression analysis concerns two matrices, X and Y. It corresponds to determining one (or several) Y variables on the basis of a well-chosen set of relevant X variables. The Y matrix consists of the *dependent variables* whilst X contains the corresponding *independent variables*.

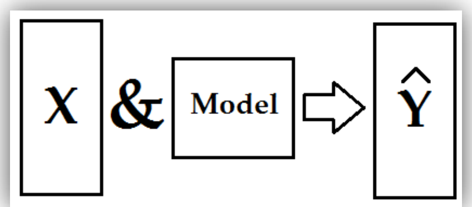
The multivariate model for (X, Y) is simply a regression between the empirical (X, Y) relations. We establish this model through multivariate calibration, as shown in ***Figure 3.1***. The statistically correct way to describe this is that we estimate the parameters of the (X, Y) regression model [33]. Thus the first stage of multivariate modeling is the calibration stage, but calibration is rarely used only for just establishing/finding a model describing the relation between X and Y. In the majority of cases we want to use the model for future prediction.

Figure 3.1



Multivariate calibration is the first step. X *and* Y *create the multivariate regression model.*

Figure 3.2



Prediction is the second step of multivariate modeling. The regression model created is now used on a new set of X *measurements in order to predict new* Y *values*

Furthermore, we wish to use the model to find new Y values from new measurements of X, i.e. to predict Y from X. Prediction is therefore the second stage of multivariate calibration. It is mandatory to start with a known set of corresponding X and Y data, in order to develop the relevant multivariate regression model. The model may then subsequently be used on new X measurements to predict new Y values.

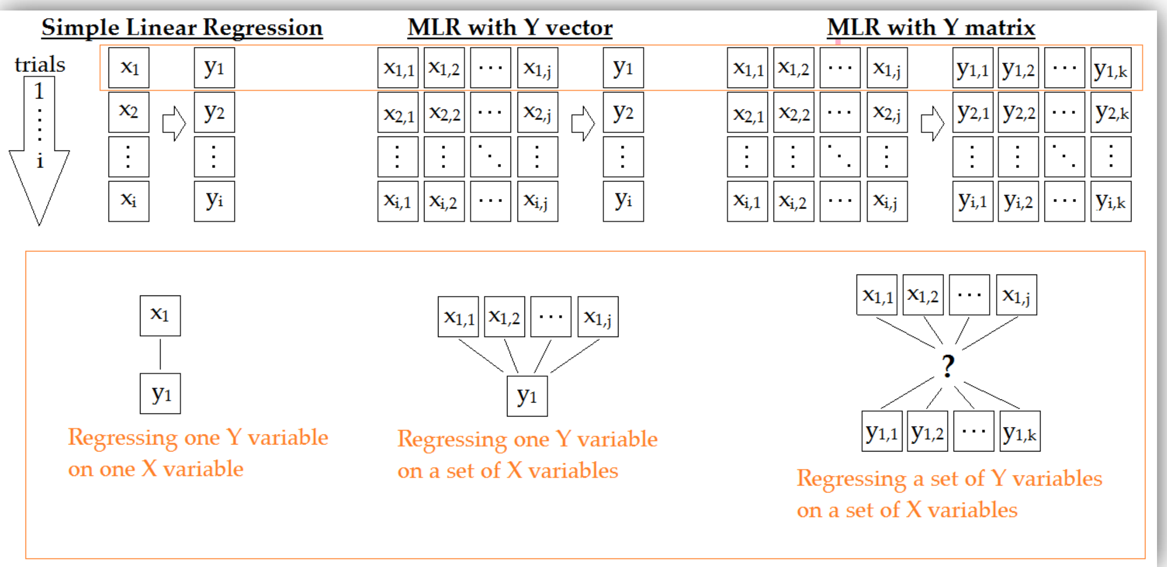
In this chapter we will focus on the two chosen regression methods, PCR and PLS, but an introductory reference to the statistical approach of the Multiple Linear Regression (MLR) has to be done in order to explain the methods of choice. These three methods have a common point in that all of them model data using a linear least squares fitting technique [34]. This means that they build linear models between the independent matrix X and the dependent matrix Y and estimate the regression coefficient matrix using ordinary least squares fitting techniques

3.1 MULTIPLE LINEAR REGRESSION

Multiple linear regression (MLR) is the classical method that combines a set of several X variables in linear combinations, which correlate as closely as possible to the corresponding single Y vector. MLR is the extension of the simple linear regression. The simple linear regression involves a single scalar predictor variable x and a single scalar response variable y, while MLR involves multiple predictor variables, denoted with a capital X. Nearly all real-world regression models involve multiple predictors, but in most of the cases the response variable y is still a scalar [35].

As mentioned above, there usually is a single response variable y which refers to a set of several X variables. We discuss this one-variable case for Y, because we only use one column vectors Y in this study, but the method can easily and readily be extended to a whole Y matrix. In ***Figure* 3.3** the simple linear regression is compared graphically to both MLR extensions, the one with single Y vector and the one with Y matrix.

Figure 3.3


Simple Linear Regression
MLR with Y vector
MLR with Y matrix
trials
1
i
x_1 y_1
x_2 y_2
x_i y_i
$x_{1,1}$ $x_{1,2}$ ⋯ $x_{1,j}$ y_1
$x_{2,1}$ $x_{2,2}$ ⋯ $x_{2,j}$ y_2
$x_{i,1}$ $x_{i,2}$ ⋯ $x_{i,j}$ y_i
$x_{1,1}$ $x_{1,2}$ ⋯ $x_{1,j}$ $y_{1,1}$ $y_{1,2}$ ⋯ $y_{1,k}$
$x_{2,1}$ $x_{2,2}$ ⋯ $x_{2,j}$ $y_{2,1}$ $y_{2,2}$ ⋯ $y_{2,k}$
$x_{i,1}$ $x_{i,2}$ ⋯ $x_{i,j}$ $y_{i,1}$ $y_{i,2}$ ⋯ $y_{i,k}$
?
Regressing one Y variable on one X variable
Regressing one Y variable on a set of X variables
Regressing a set of Y variables on a set of X variables

Comparison of three regression methods involving different number of variables[4].

Consider the model of *Equation 3.1* which is analyzed further in *Graphical Representation 3-1*

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m + \varepsilon_i$$

Equation 3.1

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \dots & X_{1,m} \\ 1 & X_{2,1} & X_{2,2} & \dots & X_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{n,m} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Graphical Representation 3-1

[4] It should be noted that all the pointers used in ***Figure* 3.3** are not the ones used in the rest of this chapter.

The variables of *Graphical Representation 3-1*are:

- Y_i is the value of the response variable in the i^{th} trial
- b_i are parameters named regression coefficients
- X_i is a row vector of m predictor variables in the i^{th} trial, and the first column is all 1's in order to include the intercept in the model
- ε_i is a random error term with mean 0 and variance σ^2

Equation 3.1 can be written if form of matrices as shown at *Equation 3.2*:

$$Y = Xb + \varepsilon \qquad \textit{Equation 3.2}$$

We wish to find the vector of regression coefficients b so that ε, the error term, is minimized. To find good estimators of the regression parameters, the least squares criterion on the squared error terms is used: find b so $\varepsilon^T\varepsilon$ is minimized. This leads to the following well known statistical estimation of b.

$$\hat{b} = (X^TX)^{-1}X^Ty \qquad \textit{Equation 3.3}$$

In addition, estimating b involves matrix inversion of (X^TX) and this may cause severe problems with MLR. If there are any collinearities[5] in X, matrix inversion may become increasingly difficult and in severe cases may not be possible at all. The $(X^TX)^{-1}$ division will also become increasingly unstable, as it will in fact correspond to dividing by zero. With intermediate to strong correlations in X, the probability of this illubehaving colliearity is overwhelming and MLR will not work in the end. To avoid this numerical instability, a standard statistical practice is to delete variables in X so as to make X become of full rank. By doing this, information is thrown away and it is almost never easy to choose amongst variables to exclude from X matrix, because we are not sure yet if a variable is useful or not (or even noise). In the worst case we may be unable to cope with the collinearities at all and have to give up. In brief MLR may fail when there is:

- Collinearity in X
- Noise, errors in X
- More variables than samples(trials) in X
- Interference amongst variables in Y

Other methods which are based on MLR, such as PCR and PLS, are able to overcome all of the above problems that may occur more or less in a dataset. PCR and PLS are bilinear projection methods which actually utilize the collinearity feature constructively and choose a solution coinciding with the variation along the solid line. This type of solution is stable with respect to collinearity.

[5] Collinearity means that the X variables are intercorrelated to a non-neglectable degree as well as that the X variables are linearly dependent to some degree [30].

3.2 PRINCIPAL COMPONENT REGRESSION

Principal component regression (PCR) is one of the most routine methods used for multivariate data analysis worldwide, especially when it comes for spectroscopic multivariate calibration. Factor analysis-based techniques, like PCR and PLS that will be introduced in *section 3.3*, are powerful multivariate statistical tools that have been successfully and widely applied to the quantitative analysis of spectroscopic data because of their ability to overcome problems common to this data such as collinearity, band overlaps and interactions [34].

PCR aims to find the factors which capture the most of the variance within the data before regression onto the response variables, whereas MLR seeks a single factor that correlates both the data and their responses. The basic idea of PCR is to use a set of orthogonal variables, thus the principal components (PCs), that are linear combinations of the original predictor variables and then regress the response variables onto the orthogonal PCs instead of the original dataset X. PCR can be thought of as a two-step procedure, first PCA is used to transform original X into a set of orthogonal and uncorrelated variables which, at the same time, maintain the variability of the original X in descending order. Then, the transformed dataset is the one used for regression with the response variables. PCR can be described as follows:

- Perform PCA to original matrix X and acquire the transformed data set $\grave{X}$.
    - In case of computing PCA with NIPALS, $\grave{X} = T$, the score vector.
    - In case of computing PCA with SVD, $\grave{X} = U\Sigma$, which came from the decomposition $X = U\Sigma V^T$.
- Put $\grave{X}$ in place of X, at MLR method from *Equation 3.2*

$$Y = \grave{X}b + \varepsilon \qquad \textit{Equation 3.4}$$

Now, the variables of X are replaced by new ones that have better properties, such as orthogonality, and also span the multidimensional space of X. In addition the inversion of $(\grave{X}^T\grave{X})$ in not a problem since the principal components are mutually orthogonal. The solution of b is analogous to *Equation 3.3*

$$\hat{b} = \left(\grave{X}^T\grave{X}\right)^{-1}\grave{X}^T y \qquad \textit{Equation 3.5}$$

Furthermore, the expanded *Equation 3.1* can be written in form of a sum like:

$$y_i = b_0 + \sum_{j=1}^{m} b_j x_{i,j} + \varepsilon_i \qquad \textit{Equation 3.6}$$

Recall that the data matrix X can also be expanded in a similar form as in the formulation of NIPALS, in *Equation 2.20* of section *2.1.3*. Thus, every x_i is decomposed, for principal components $k = \min(n, m)$[6], using scores t_h and loadings p_h, as follows:

$$x_i = \sum_{j=1}^{k} t_{i,j} p_j \quad , \qquad \textit{Equation 3.7}$$

where $k = \min(n, m)$ are all the principal components

[6] n, m are the original dimensions of X

Finally the model, for k' selected components, where $k' < k$, is written as:

$$y'_i = b'_0 + \sum_{j=1}^{k\prime} b' t_{i,j} + \varepsilon_i \qquad \textit{Equation 3.8}$$

3.2.1 CRITERIA FOR PC SELECTION

So far, not only $\grave{X}$ is uncorrelated, but it may also have smaller dimensions than the original X, according to k', the number of principal components retained. It can be seen that a sequence of models for Y are obtained, one for each choice of k'. In general, the question of which model should be used depends on the purpose of modeling and the model selection criterion used.

In case of calibrating spectra data, the predictive ability of a model is mainly the focus of modeling and the criterion most commonly used for model selection is cross validation on the root mean square error of prediction (RMSEP), as is shown at *Equation 3.9*

$$RMSEP_{k'} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - y'_i)^2}{n}} \;, \qquad \textit{Equation 3.9}$$

where:

- $RMSEP_{k'}$ is the root mean square error of prediction with k' components
- y_i is the already measured response variable of sample i
- y'_i is the estimated response variable of sample i
- n is the number of samples tested

A model is preferred if it can yield a smaller minimum prediction error and requires a smaller number of components to achieve this minimum prediction error. Therefore, the subset k' of k PCs that minimize RMSEP would be considered as best.

When prediction is the main objective, the variance of a single regression coefficient is not of interest. In this situation we are mostly interested in the variance of a predicted value and this is usually evaluated my RMSEP. Unlike the variance of estimated regression coefficients, RMSEP depends on both the new observation and the variances of principal components and not merely on variances of principal components.

As mentioned before, PCR is based on the idea of PCA which assumes that the main information of interest is contained in the first PCs of the predictor dataset with high variations. This is the reason that the criterion most often used for inclusion, or exclusion, of a PC is the magnitude of the variability it gathers. Unfortunately, this does not necessarily reflect the predictive ability of the PC, i.e. good predictive performance is only expected when PCs associated with large variations happen to be PCs with good predictive abilities. However, in spectroscopic analysis, high variance PCs can sometimes be generated by sources unrelated to prediction, but on the other hand in some situations, the high variations may be generated by sources that are not related to the variable under study. In these cases, it is natural to use only the principal components that represent relevant directions in the regression.

3.3 PARTIAL LEAST SQUARES REGRESSION

Partial least squares regression (PLSR) is also known simply as PLS [7]. PLS is used for both supervised data analysis and for regression, involving matrix X along with response matrix . PLS has seen an unparalleled application success, both in chemometrics and other fields. PLS, like PCR (and thus PCA), is a factor analysis-based method which is also an extension to MLR. Amongst other features, the PLS approach gives superior interpretation possibilities and claims to do the same job as PCR, only with fewer bilinear components [35]. The main goal of PLS usage is the building of calibration models, but this technique can also be applied for classification purposes.

PLS assumes that the information of interest about a response variable is mainly contained in the directions of the predictor space which have both large variations and high correlations with the response variables. In PCR the principal components are determined by the predictor variables. In contrast, the corresponding vectors in PLS are determined by both predictor and response variables. Correspondingly, in PLS there is an exchange of information between predictor and response variables in the process of determining these vectors, while in PCR there is no such exchange at all. The major information about a response variable is often included in the principal components with intermediate variances, especially in the case in which the ratio of the sample size to the number of predictor variables is small [36].

As in MLR, the main purpose of PLS is to build a linear model, as shown previously in *Equation 3.2*.

$$Y = Xb + \varepsilon \qquad \textit{Equation 3.2}$$

PLS produces factor scores which are linear combinations of the original predictor variables of the X matrix, so that there is no correlation between the factor score variables used in the predictive regression model. We know, from *Equation 2.21* of section *2.1.3,* that matrix X can be decomposed to loadings and scores as: $X = TP' + F$, meaning that $T = XP + F$[8]. In a similar way, matrix Y can be decomposed as:

$$\mathrm{Y} = \mathrm{UQ}' + \mathrm{E}\,, \qquad \textit{Equation 3.10}$$

where:

- Q' is a vector of regression coefficients (loadings) for Y
- U is a score matrix for Y
- E is an error term.

Thus, the score matrix for Y is $U = YQ + E$. The exchange of the information between X and Y matrices is done by the exchange of their scores, meaning that T takes the place of U in Y as: $Y = TQ' + E = XPQ' + E$. However, remember that the decomposition of X and Y is achieved when the scores and the loadings are orthogonal. Exchanging information from a score matrix to another would lead to a model with maximum covariance between responses and factor scores, but at the same time,

[7] In this study we choose to use the term PLS than PLSR, when referring to partial least squares regression.

[8] Note that we use generic variable F in many places to describe the unexplained part of the X scores for the factor model of X, while E is the residual for the factor model of Y

with the altered scores, the orthogonality for their pair loadings P would have been lost. In other words, what we want is to find a new score matrix for X, that maximizes the covariance of X and Y and at the same time keeps the loadings Y orthogonal to it.

To achieve this, we use a weight matrix W in place of the original P, which is updating, in order to keep the orthogonality of the decomposition, when a new piece of information is transferred from Y scores to X scores. So, now we have:

- U the initial score matrix of Y
- T the initial score matrix of X
- P the initial loadings matrix of X
- W the new altered loadings matrix of X

The regression method of PLS computes the factor score matrix T for an appropriate weight matrix W as:

$$T = XW + F \quad \text{Equation 3.11}$$

Then, in order to exchange information each block with the other, the scores T and U change place:

$$\mathrm{Y} = \mathrm{TQ}' + \mathrm{E} \quad \text{Equation 3.12}$$

Once the loadings Q are computed, the above regression model is equivalent to

$$\mathrm{Y} = \mathrm{XB} + \mathrm{E} \quad \text{Equation 3.13}$$

, where:

$$B = WQ' \quad \text{Equation 3.14}$$

This B matrix can now be used as a predictive regression model. As explained above, PCR and PLS differ in the approach used for extracting factor scores, meaning that PCR produces the loadings matrix P reflecting the covariance structure between the predictor variables, while PLS produces the weight matrix W reflecting the covariance structure between the predictor and response variables.

For establishing the model, PLS produces a weight matrix W for X such that: $T = XW$. The columns of W are weight vectors for the X columns producing the corresponding factor score matrix T. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. As already described above and in chapter 2, the factor loading matrix P ,which comes from the factor model:

$$X = TP' + F \quad \text{Equation 3.15}$$

, is necessary for a complete description of PLS procedures. Because PLS exchanges scores between the X block and the Y block, the order of calculations that are used to decompose X into TP' are changed, and the result is that T is no longer orthogonal. This is the main reason for the use of weights W in place of P; weights W now maintain orthogonality after every exchange of scores between X and Y blocks.

Ordinary least squares procedures for the regression of Y on T are then performed to produce Q, the loadings for Y such that: $Y = TQ' + E$. Once Q is computed, we have: $Y = XB + E$, where $B = WQ'$. That completes the prediction model.

At this point the dimensions of X, T and P matrices from *Graphical Representation 2-8* should be reminded.

- X is an $(n * m)$ matrix
- T is an $(n * r)$ scores matrix
- P is an $(m * r)$ loadings matrix, thus P' is an $(r * m)$ matrix
- r is the rank of matrix X

In addition to above matrix dimensions:

- Y is an $(n * 1)$ vector of responses
- U is an $(n * r)$ scores matrix
- W, from *Equation 3.11,* has the same $(m * r)$ dimensions as P
- Q', from *Equation 3.10,* is an $(n * 1)$ matrix, referred as loadings or weights matrix for Y
- B, from *Equation 3.13 & Equation 3.14,* is an $(m * 1)$ vector

There are two different approaches for solving PLS, the NIPALS algorithm and the SIMPLS algorithm. Their main difference is that in SIMPLS a set of weights, usually referred to as R, is calculated and it operates on the original X data to calculate the scores, while in NIPALS each weight vector w is applied to different X_i [37]. De Jong in [38] showed that it is easy to calculate the SIMPLS weights R from the NIPALS W and P as: $R = W(P'W)^{-1}$.

Moreover, in case of vector Y, meaning that there is only one response variable, the results obtained from NIPALS are the same as the results obtained by SIMPLS, as proved in [38]. However, when applied to a matrix Y, with more than one response variable, the results of the SIMPLS algorithm are different from the NIPALS. In our study, since we are focused on a single response variable, we chose to compute PLS with NIPALS.

3.3.1 SOLVING PLS: NIPALS ALGORITHM

The standard algorithm for computing PLS components (i.e., factors) is nonlinear iterative partial least squares (NIPALS). NIPALS algorithm was introduced in the previous chapter as a methodology that computes PCA on a simple data matrix X. The NIPALS algorithm presented in chapter 2 is just a simplification of the algorithm described here.

In PLS, NIPALS involves both predictor X and response Y matrices, as well as, a weight matrix W. A simplified model would consist of a regression between the scores for the X and the Y block. The PLS model can be considered as consisting of outer relations of X and Y block individually, and an inner relation linking both blocks. The outer relation for the X block is as described in *Equation 3.15* and equivalently the outer relation for the Y block is as described in *Equation 3.10.*

Assuming that X and Y matrices are mean-centered the NIPALS algorithm for PLS is as follows [39]:

1) Select a vector w from a row of X
2) Normalize it to length 1.
3) Compute a score vector $t = Xw$
4) Compute a Y-loading vector $q = Y^T t$
5) Compute a Y-score vector $u = Yq$
6) Compute a new weight vector $w_1 = X^T u$
7) Normalize w_1 to length 1
8) If $|w - w_1| < d$, where d is a very small number like 10^{-8},the converge is obtained. If not, $w = w_1$ and go back to step (2)

The scores t and u, for matrices X and Y respectively, for the first component are obtained from the above iterations. To find the next pair of (t, u) score vectors, X is adjusted for the score vector and Y is adjusted for the results obtained after the regression of Y onto t. The adjusted X is simply the residual F, while the adjusted Y is the residual E

9) Compute the loading vector $p = (X^T t)/(t^T t)$
10) Adjust X for what has been found: $F = X - tp^T$
11) Compute regression of Y onto t: $b = (Y^T t)/(t^T t)$
12) Adjust Y for what has been selected: $E = Y - tb^T$

To implement the procedure for the next component, go to step (1) and replace X and Y by their residuals. We can do these iterations until the rank of X is exhausted, since the rank of Y is not decreasing through these iterations.

3.4 PREDICTION

In summary, PLS and PCR are both methods to model a response variable when there is a large number of predictor variables, and those predictors are highly correlated or even collinear. Both methods construct new predictor variables, known as components, as linear combinations of the original predictor variables, but they construct those components in different ways. PCR creates components to explain the observed variability in the predictor variables, without considering the response variable at all. On the other hand, PLS does take the response variable into account, and therefore often leads to models that are able to fit the response variable with fewer components.

Accordingly to PCR, where a sequence of models for Y are obtained, one for each choice of k', in PLS also the number of components needed to describe the PLS model is equal to the model dimensionality. We could define k' in PLS as the maximum number of components used in PLS calibration: $h = 1,2, \dots, k'$

The number of components to be used either in PCR or PLS is a very important property of both models. Although it is possible to calculate as many components as the rank of the predictor X data matrix, not all of them are normally used. The main reasons for this are that the measured data are never noise-free, meaning that some of the smaller components will only describe noise, and, in most of the cases, the smaller components are not related with the targeted responses.

The choice of components should be made when specific criteria are met. If the number of components is too big, there is a chance for overfitting the model that would lead to poor predictive performance, as it can exaggerate minor fluctuations in the data [40].

The important part of any regression method is its use in predicting the dependent block $\hat{Y}$ from the independent block X. The main part of both PCR and PLS algorithms is the built of the B matrix. This matrix B holding the regression coefficients, is used to predict a new sample's response as:

$$\hat{y} = X_{new}B \quad , \qquad \textit{Equation 3.16}$$

where

- $\hat{y}$ is the estimation of the response variable concerning the new predictor data X_{new}
- X_{new} is the new data of which response is the one we want to predict
- B is the regression coefficients, computed via a regression method

4. Proposed Methodology

In this chapter we will annotate step by step the methodology followed in this study. As mentioned in the introduction, the goal of this study is to make a predictive model for aromatic concentration concerning petroleum products. We wish to adjust the introduced algorithms to the needs of our problem.

Furthermore, from now on, matrix X holds all the spectral information for the selected samples. X is an $(n * m)$ matrix. Each row of X corresponds to a sample, while each column of X corresponds to a variable obtained at a particular wavelength. In other words, X has n samples and each sample has m variables. In our case Y is a $(n * 1)$ row vector. Each row of Y has for term the corresponding pre-measured value of the aromatics' concentration of a sample. The spectral information of a random sample i, is stored at the i^{th} row of X, while the response of the sample, concerning the aromatics, is stored at the i^{th} row-element of Y. The size of n, m will be discussed in the next sections.

However, before even we get to apply mathematical models and algorithms to our data, we had to organize all the data we have acquired.

4.1 Experimental data organization

When this study started, we had at our disposal about 35 different petroleum samples, meaning that we had 35 spectral data one for each sample and at least one experimental measurement of aromatics concentration for each sample. There were enough cases that we had more than one experimental measurement for aromatics by different methods, such as: ASTM D 2549, HPLC ASTM D 7419-07, SARA and ASTM D 1319.

Furthermore, as time was passing by from the start of this study, we have managed to get even more samples. Finally, this study was made when we had gradually acquired 74 different petroleum samples, some of them analyzed experimentally by more than one method. This unknown and unpredictable expansion rate of our dataset gave birth to an idea; there should be a well-sorted, flexible and easy to use "library" of all the tested samples, one that could easily generate different datasets for analysis. At this point we should mention that all of the coding used for this study was made in Matlab, placing a restriction regarding what kind (or file type) should a dataset output from our library be, in order to be imported to the mathematical models designed in Matlab[9].

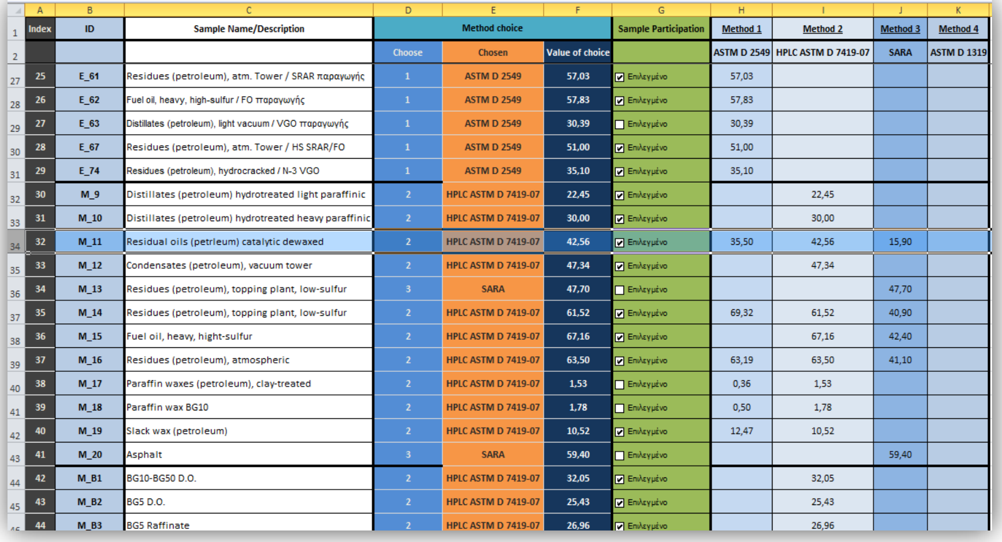
The solution to this problem, for our purposes, was to create an excel datasheet holding all the information regarding our samples. A snapshot of the database is shown in ***Figure 4.1***.

Every row of the *excel* file belongs to a sample. We may select or deselect each sample separately, in order to be part of the calculation processes. Moreover, every sample has at least one response value obtained by certain experimental method. Some samples may have more than one responses, obtained by different experimental methods. For every sample we can choose which method, aka which value, we

[9] The complete list of the samples is in the last chapter – Appendix at section *7.1.2 Complete list of samples*

want to participate in the algorithms. For a more detailed explanation of the *excel* database, see section *7.1.1* in Appendix.

Figure 4.1



	A	B	C	D	E	F	G	H	I	J	K
1	Index	ID	Sample Name/Description	Method choice			Sample Participation	Method 1	Method 2	Method 3	Method 4
2				Choose	Chosen	Value of choice		ASTM D 2549	HPLC ASTM D 7419-07	SARA	ASTM D 1319
27	25	E_61	Residues (petroleum), atm. Tower / SRAR παραγωγής	1	ASTM D 2549	57,03	☑ Επιλεγμένο	57,03			
28	26	E_62	Fuel oil, heavy, high-sulfur / FO παραγωγής	1	ASTM D 2549	57,83	☑ Επιλεγμένο	57,83			
29	27	E_63	Distillates (petroleum), light vacuum / VGO παραγωγής	1	ASTM D 2549	30,39	☐ Επιλεγμένο	30,39			
30	28	E_67	Residues (petroleum), atm. Tower / HS SRAR/FO	1	ASTM D 2549	51,00	☑ Επιλεγμένο	51,00			
31	29	E_74	Residues (petroleum), hydrocracked / N-3 VGO	1	ASTM D 2549	35,10	☑ Επιλεγμένο	35,10			
32	30	M_9	Distillates (petroleum) hydrotreated light paraffinic	2	HPLC ASTM D 7419-07	22,45	☑ Επιλεγμένο		22,45		
33	31	M_10	Distillates (petroleum) hydrotreated heavy paraffinic	2	HPLC ASTM D 7419-07	30,00	☑ Επιλεγμένο		30,00		
34	32	M_11	Residual oils (petrleum) catalytic dewaxed	2	HPLC ASTM D 7419-07	42,56	☑ Επιλεγμένο	35,50	42,56	15,90	
35	33	M_12	Condensates (petroleum), vacuum tower	2	HPLC ASTM D 7419-07	47,34	☑ Επιλεγμένο		47,34		
36	34	M_13	Residues (petroleum), topping plant, low-sulfur	3	SARA	47,70	☐ Επιλεγμένο			47,70	
37	35	M_14	Residues (petroleum), topping plant, low-sulfur	2	HPLC ASTM D 7419-07	61,52	☑ Επιλεγμένο	69,32	61,52	40,90	
38	36	M_15	Fuel oil, heavy, hight-sulfur	2	HPLC ASTM D 7419-07	67,16	☑ Επιλεγμένο		67,16	42,40	
39	37	M_16	Residues (petroleum), atmospheric	2	HPLC ASTM D 7419-07	63,50	☑ Επιλεγμένο	63,19	63,50	41,10	
40	38	M_17	Paraffin waxes (petroleum), clay-treated	2	HPLC ASTM D 7419-07	1,53	☐ Επιλεγμένο	0,36	1,53		
41	39	M_18	Paraffin wax BG10	2	HPLC ASTM D 7419-07	1,78	☐ Επιλεγμένο	0,50	1,78		
42	40	M_19	Slack wax (petroleum)	2	HPLC ASTM D 7419-07	10,52	☑ Επιλεγμένο	12,47	10,52		
43	41	M_20	Asphalt	3	SARA	59,40	☐ Επιλεγμένο			59,40	
44	42	M_B1	BG10-BG50 D.O.	2	HPLC ASTM D 7419-07	32,05	☑ Επιλεγμένο		32,05		
45	43	M_B2	BG5 D.O.	2	HPLC ASTM D 7419-07	25,43	☑ Επιλεγμένο		25,43		
46	44	M_B3	BG5 Raffinate	2	HPLC ASTM D 7419-07	26,96	☑ Επιλεγμένο		26,96		

The petroleum samples and aromatics given by standard methods

With this excel sheet we now can generate various combinations of selected samples and select which experimental value for each sample will participate to the calculations that will follow. With this method, we will try to conclude how well can the mathematical models perform among different petroleum families. We expect that datasets filled with a particular type of petroleum product, will generally perform better than datasets containing different type of samples, i.e. a dataset containing petroleum, distillates and residual oils. In addition, we will try to find an optimum model that could predict the concentration of aromatics within the whole range of petroleum products.

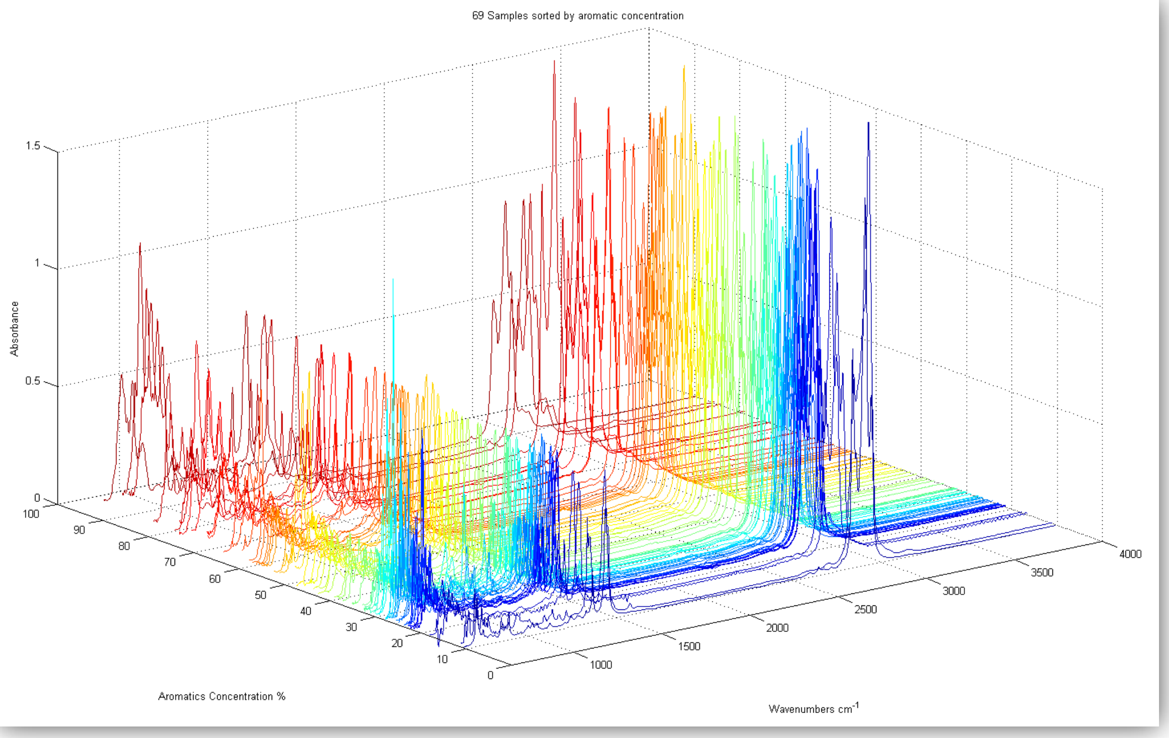
The usefulness of a tool such as a flexible multi-information database, for these types of measurements, is not fully uncovered in this study. In this study we will work with some instances of this database. Every instance of it, referred to as *scenario,* will contain a subset of the samples, and every sample will be linked with only one measurement. Practically, there are too many possible *scenarios* to be examined. Every *scenario* has its own special significance, analogous to the type of samples and the type of measurements within it.

In order to have some constant datasets to work with, four major *scenarios* were created. The first includes 69 samples of all the petroleum families and their respective response values are obtained considering all the available methods with a decreasing order from method 1 to 4. In other words, one value is obtained, which corresponds to the first available measurement from method 1 to 4. This task is generally difficult to work with analytical models, because many properties vary in these samples and also because they are generally obtained from various methods. The second *scenario* includes 30 petroleum samples which concentration values are obtained only from *ASTM D 2549* method (1). *Scenario 2* mostly includes petroleum distillates and residues but fuel oils as well as some waxes are included also. The 22 samples included in *scenario 3* are obtained by *HPLC ASTM D 7419-07* analytical method (2) and they are mainly petroleum distillates and residual oils. Finally all the samples participating in *scenario 4* are gazolines and diesel oils, all measured by the *ASTM D 1319* analytical method(4). Furthermore, 5 samples are not used in any of the *scenarios*, because they are extreme outliers for mathematical and chemical consideration. These samples are paraffin waxes with experimental measured concentrations from 1.5% to 0.2%, values that will majorly contribute to altered results.

4.1.1 Scenarios

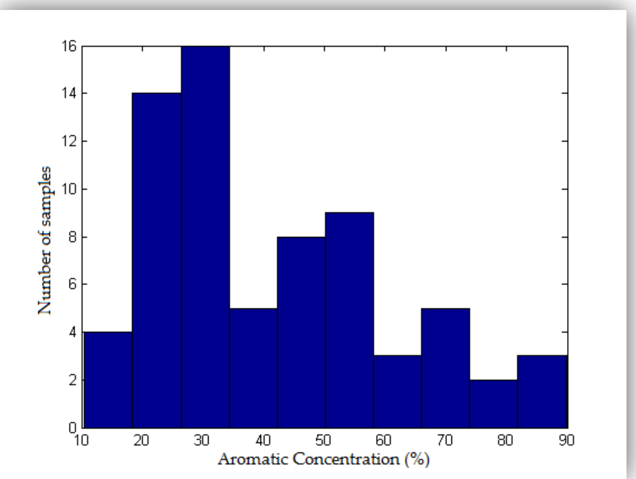
In this section, we quote, for every *scenario*, a 3D plot of the samples, sorted according to their respective response of aromatic concentration. Furthermore, for every *scenario*¸ a histogram figure is showing the range of the responses.

Figure 4.2



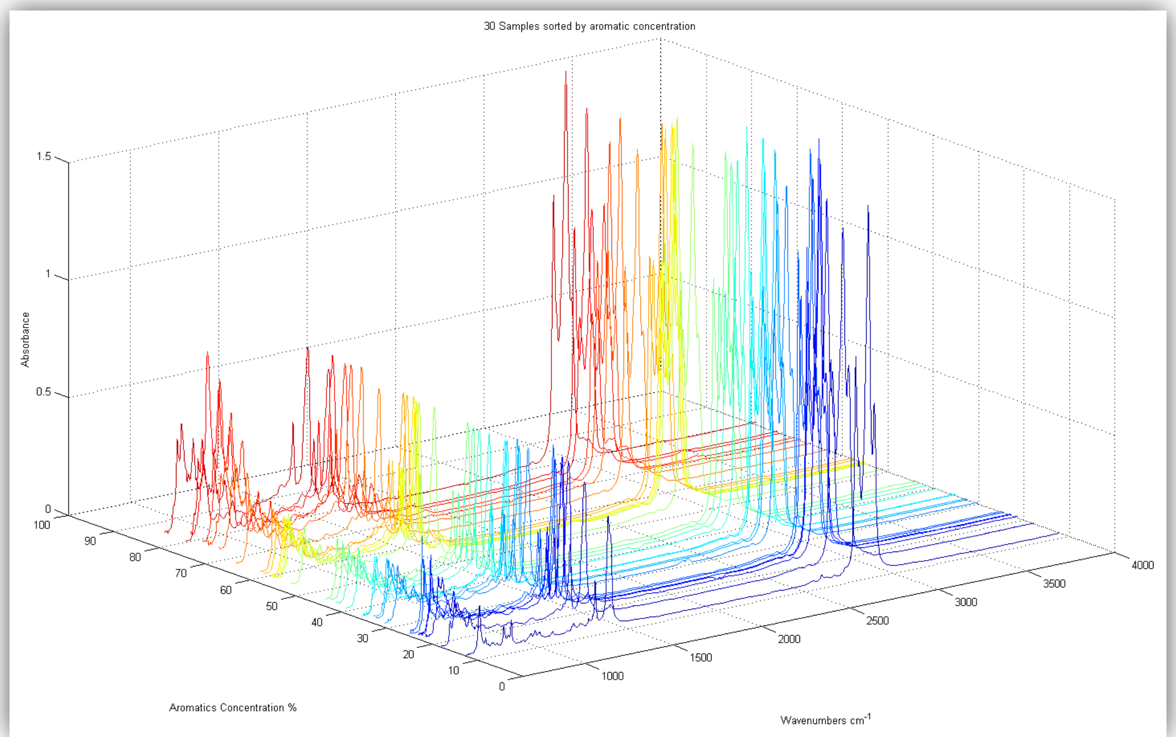
3D plot of the Mid-FTIR spectrums of scenario 1

Figure 4.3



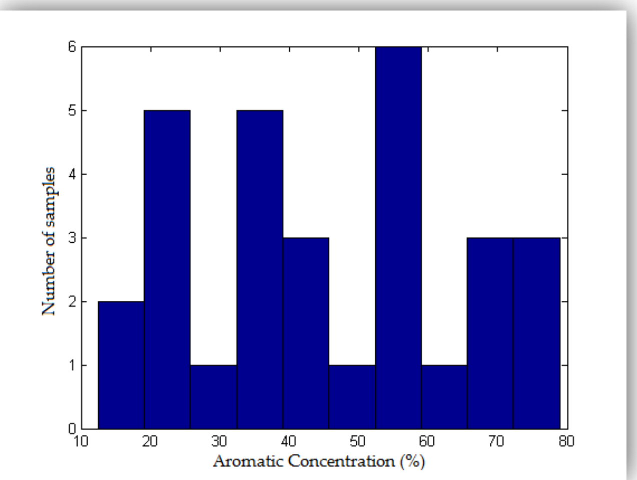
Histogram reflecting the population of the concentration measurements, for scenario 1

Figure 4.4



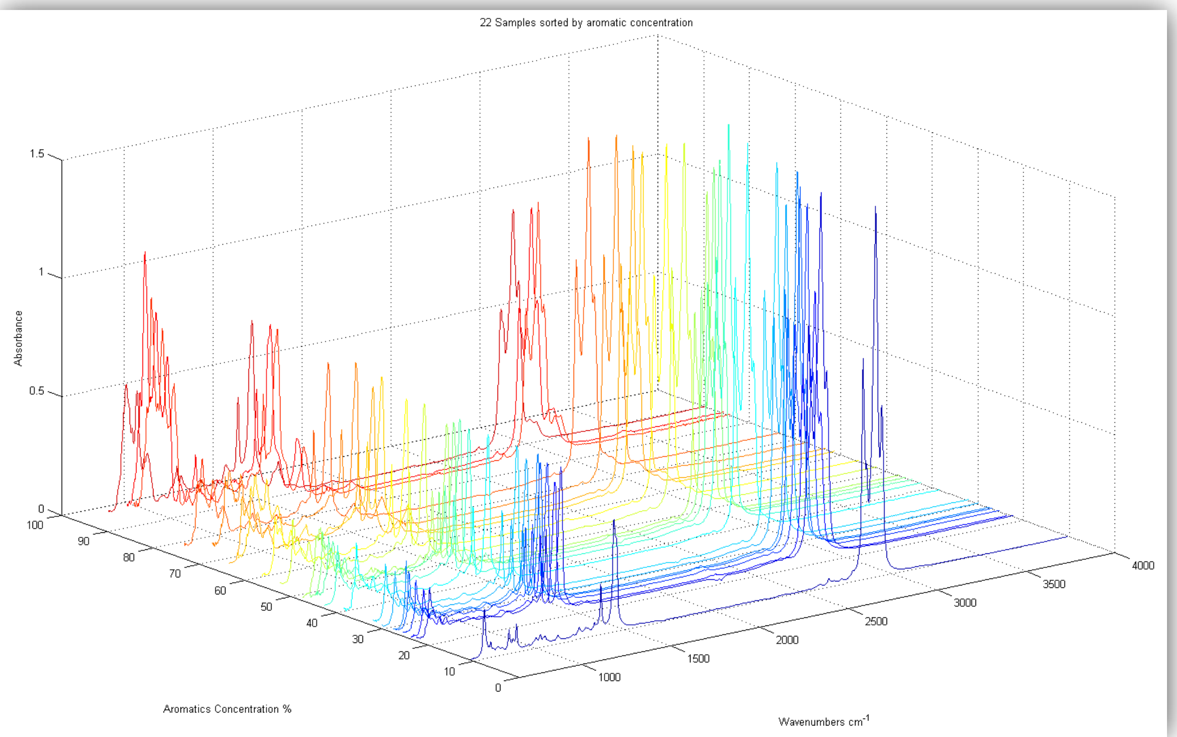
3D plot of the Mid-FTIR spectrums of scenario 2

Figure 4.5



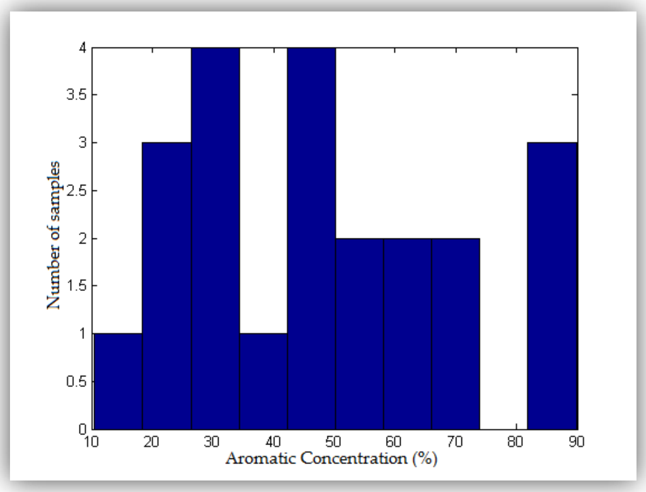
Histogram reflecting the population of the concentration measurements, for scenario 2

Figure 4.6



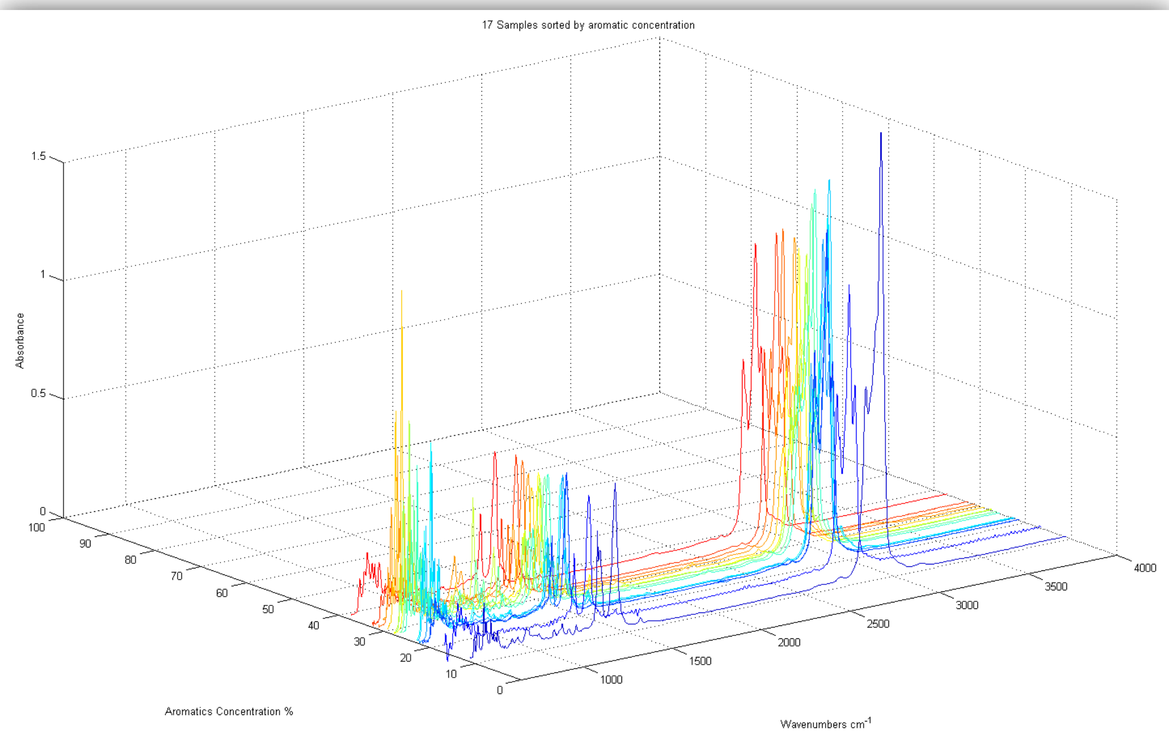
3D plot of the Mid-FTIR spectrums of scenario 3

Figure 4.7



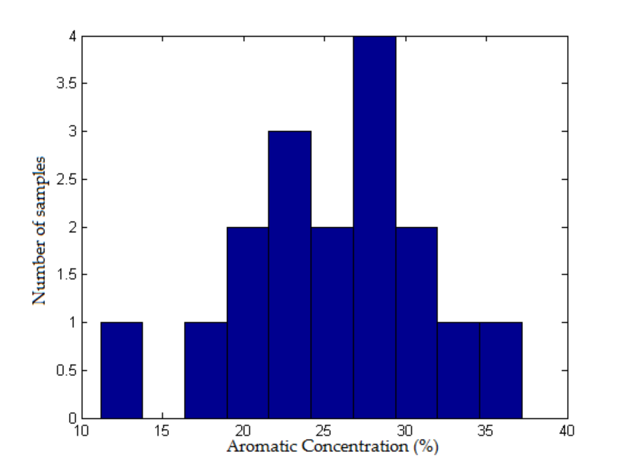
Histogram reflecting the population of the concentration measurements, for scenario 3

Figure 4.8



3D plot of the Mid-FTIR spectrums of scenario 4

Figure 4.9



Histogram reflecting the population of the concentration measurements, for scenario 4

Every generated *scenario* contains the samples that will fill the X matrix, as well as the aromatic concentration responses that will fill the Y matrix. So, in our study X and Y matrices do not have constant dimensions, but we present them in their generic form, as done in all the previous chapters.

Thus, X in an $(N * M)$ matrix, where N is the number of samples in the *scenario* loaded and M is the sample's variables. Our spectra are taken from an FTIR instrument in the mid-infrared range, from 4000 cm^{-1} to 650 cm^{-1} with an interval of 2 cm^{-1}. This produces a total of $(4000 - 650)/2 + 1$ variables, meaning that $M = 1676$. On the other hand, Y is an $(N * 1)$ vector. A representation of X and Y matrices is shown below:

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,M} \\ X_{2,1} & X_{2,2} & \dots & X_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \dots & X_{N,M} \end{bmatrix}$$

Graphical Representation 4-1

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Graphical Representation 4-2

4.2 DATA PRE-TREATMENT

Pre-treatment is the first process an analyst should do with his data. Ordinary normalization by mean centering was applied to the spectra data, but on the other hand no one can assure that normalizing the original data will give better predictive models. That is the reason why all the further mathematical calculations where performed with both original and normalized spectral data. Below there is a graphical representation of the normalization process that takes place for our matrix X.

$$From\, X\colon \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,M} \\ X_{2,1} & X_{2,2} & \dots & X_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \dots & X_{N,M} \end{bmatrix} generate\, mean\, for\, every\, row\colon \begin{bmatrix} mean_1 = \frac{\sum_{i=1}^{M} X_{1,i}}{M} \\ mean_2 = \frac{\sum_{i=1}^{M} X_{2,i}}{M} \\ \vdots \\ mean_N = \frac{\sum_{i=1}^{M} X_{N,i}}{M} \end{bmatrix}$$

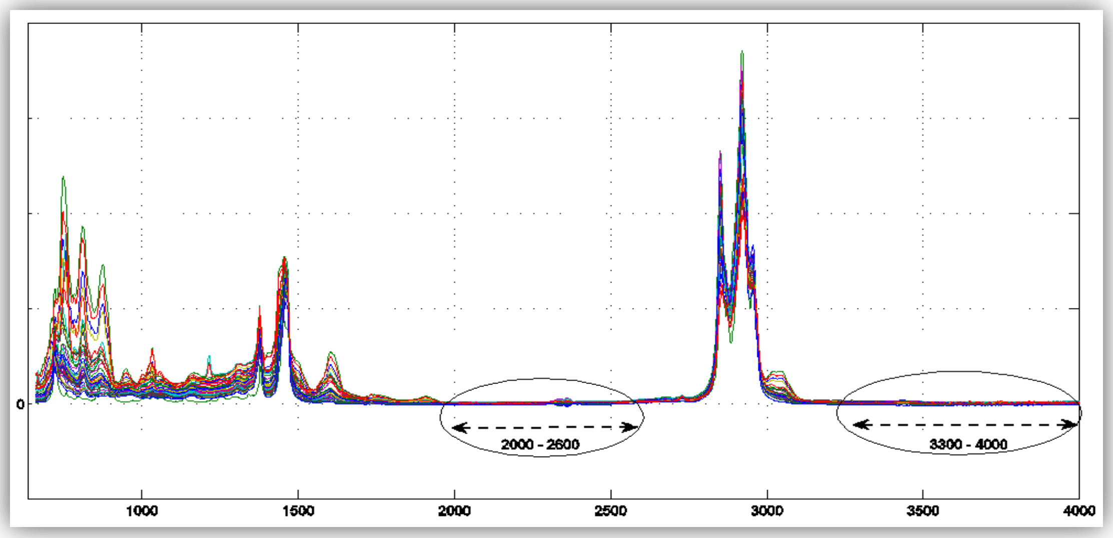
$$and\, compute\, X_{normalized} = \hat{X}\, as:$$

Graphical Representation 4-3

$$\hat{X}\colon \begin{bmatrix} \hat{X}_{1,1} = X_{1,1} - mean_1 & \hat{X}_{1,2} = X_{1,2} - mean_1 & \cdots & \hat{X}_{1,M} = X_{1,M} - mean_1 \\ \hat{X}_{2,1} = X_{2,1} - mean_2 & \hat{X}_{2,2} = X_{2,2} - mean_2 & \dots & \hat{X}_{2,M} = X_{2,M} - mean_2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{X}_{N,1} = X_{N,1} - mean_N & \hat{X}_{N,2} = X_{N,2} - mean_N & \dots & \hat{X}_{N,M} = X_{N,M} - mean_N \end{bmatrix}$$

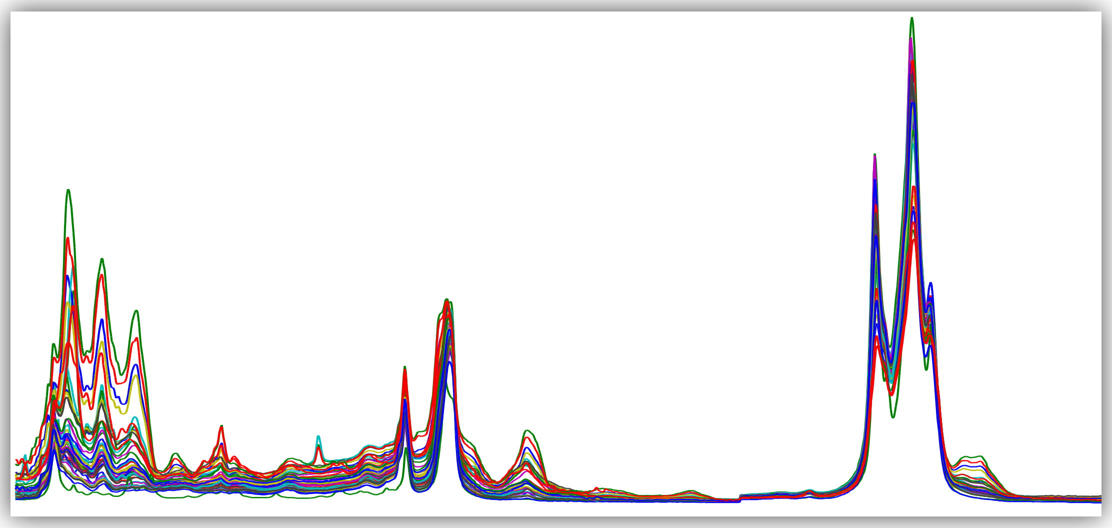
Furthermore, if we look at the original data, a two-dimensional plot will make more sense now, we will notice that there are certain spectral regions in which the spectral absorbance of the whole dataset is nearly zero. According to ***Figure 4.10***, the spectra region from 2000 cm^{-1} to 2600 cm^{-1} (approximately), as also the region from 3300 cm^{-1} to 4000 cm^{-1}, contain absorbance values which tend to zero and no significant changes of the signal can be found there. Theoretically, these two regions could reduce the coefficients based on the variance of the variables and thus they could be excluded from the total spectrum range. By removing these two regions we get a new, smaller, dataset as shown at figure 3.3.

Figure 4.10



The 2D plot for the original data stored in X . The two regions in circle contain no useful information.

Figure 4.11



The 2D plot for the reduced data. The two regions marked in circle above are now missing.

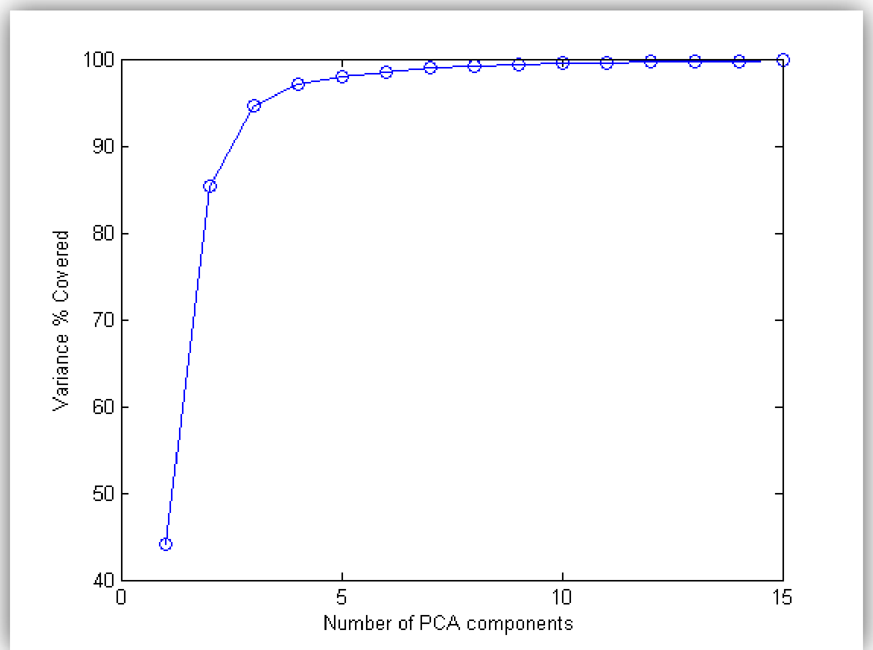
The size of the new dataset is $(N * M')$, where M' is the original multitude M of variables minus the excluded ones. The removed spectra are the spectra stored in the regions 2000-2600 cm^{-1} and 3300-4000 cm^{-1}, meaning that the spectra left in the dataset are the ones in the regions 650-2000 cm^{-1} and 2600-3300 cm^{-1}. After the variable for the 2000 cm^{-1}, the next variable corresponds to 2002 cm^{-1} due to the interval of the instrument's sampling. Therefore, the variables which are to be removed are in these regions: 2002-2598 cm^{-1} and 3302-4000 cm^{-1} and this is how one can compute M':

$$M' = M - \left(\frac{2598 - 2002}{2}\right) + 1 - \left(\frac{4000 - 3302}{2}\right) + 1 = 1027 \qquad \textit{Equation 4.1}$$

4.3 Data analysis – Prediction

The next step after the data pre-treatment is the data analysis itself. Our work was done in order to evaluate how the two most widespread methods of multivariate calibration, PCR and PLS, perform on MIR spectra data of petroleum products. In addition, the simplest and quickest way to evaluate how adequately the models describe the original data is to examine the total variance explained by each component. An example can be seen below at ***Figure 4.12*** ,where the total variance explained by each PCA component is plotted.

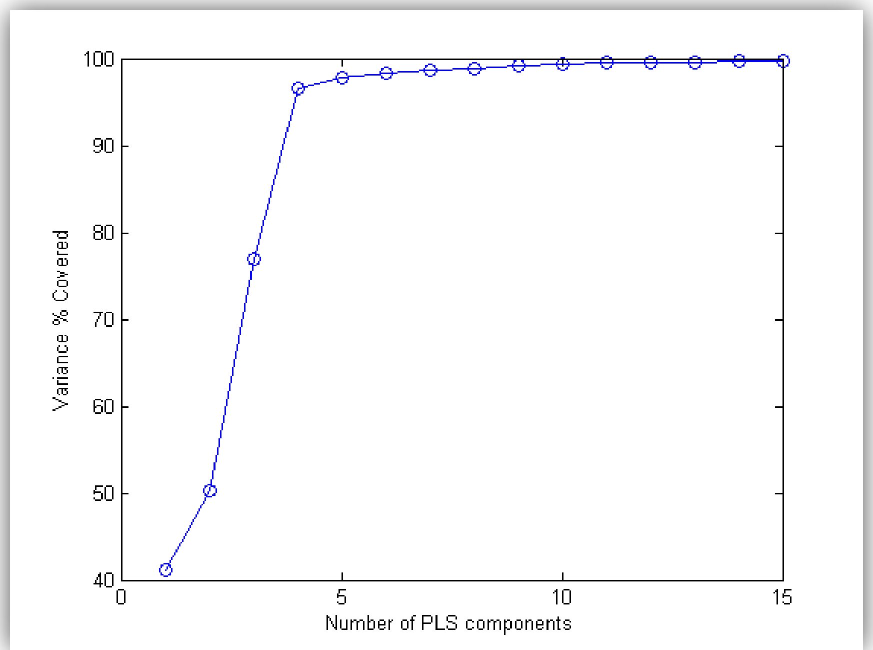
Figure 4.12



Total variance of X explained by each component of the PCA algorithm.

According to the above plot, the 99% of the variance is explained with only seven components. This means that a PCA model with seven principal components describes almost all of the information of *X*. On the other hand, PLS needs nine components to describe the same amount of information, as it can be seen at ***Figure 4.13***.

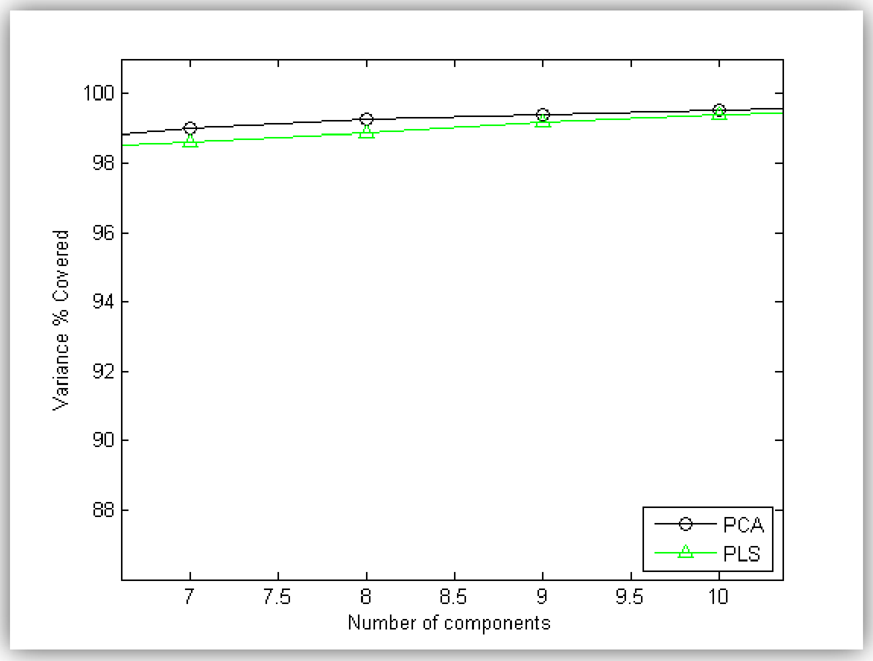
Figure 4.13



Total variance of X explained by each component of the PLS algorithm.

The difference between the above two plots is negligible, but the fact that ,in our case, PCA needs less components than PLS to describe the same piece of information can be understood by looking at the zoomed combined plot at ***Figure 4.14***.

Figure 4.14

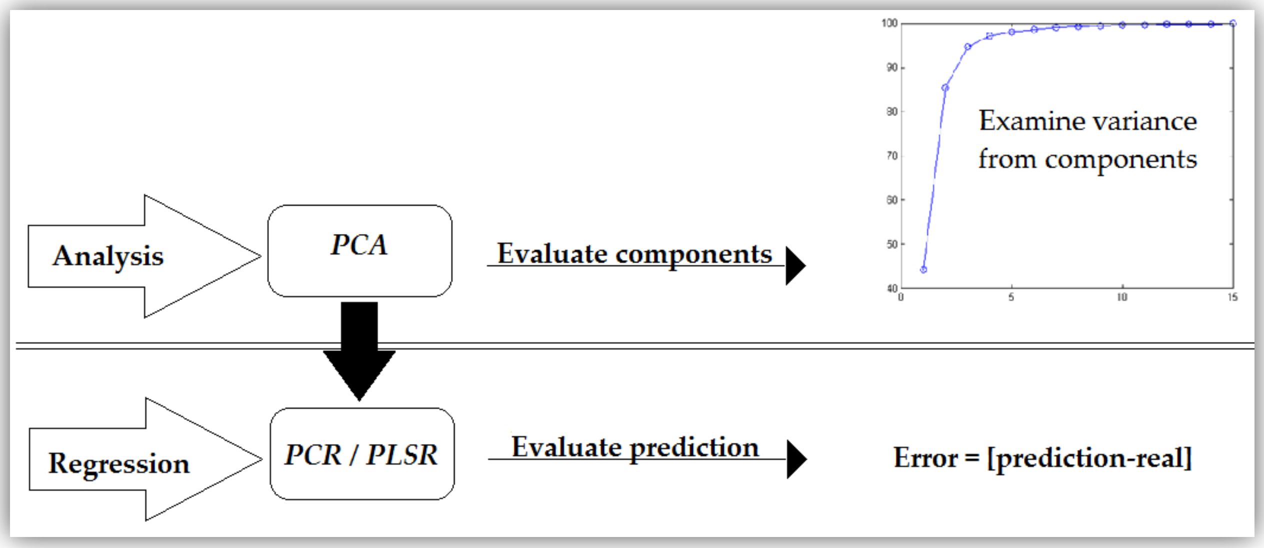


Comparison of the total variance of X explained by each component of the PCA and PLS algorithms.

In practice, we do not know if the information we are looking for (in our case, the concentration of aromatics compounds) is actually included in a model made by seven components. In our study, this hypothesis does not apply, meaning that the information we seek to model with aromatics is not found within the seven first components of PCA.

The goal of this study is to make and compare predictive models, so instead of evaluating the analysis of the samples[10], we evaluated the prediction results given by each regression method with all the possible combinations of components, as shown at ***Figure 4.15***.

Figure 4.15



Instead of evaluating the models provided from analysis, by examining each component, we evaluate the predictions' results obtained from the corresponding analytical methods.

Regarding the analysis stage, all three algorithms described under chapter *2.1* were implemented and tested. Furthermore, the chosen one to lead the PCR is SVD, as described in *2.1.2*. At the regression stage, PCR estimates the regression coefficients and predicts new responses as shown in chapter *3.2*. PLS is achieved by the NIPALS algorithm as described extensively in chapter *3.3*.

The result of the prediction is evaluated via Root Mean Square Error of Prediction (RMSEP), as shown in *Equation 4.2* below:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y_i} - y_i)^2}{N}} \;, \qquad \textit{Equation 4.2}$$

where $\hat{y_i}$ is the predicted value for sample i, y_i is the original value for sample i and N is the number of samples participated in prediction process.

[10] Remember that PCA is performed only on X matrix, while PLS is performed at both X and Y simioutanesly. PCA actually analyzes the samples themselves, while PLS has to perform the entire NIPALS regression methodology in order to acquire an analysis of X.

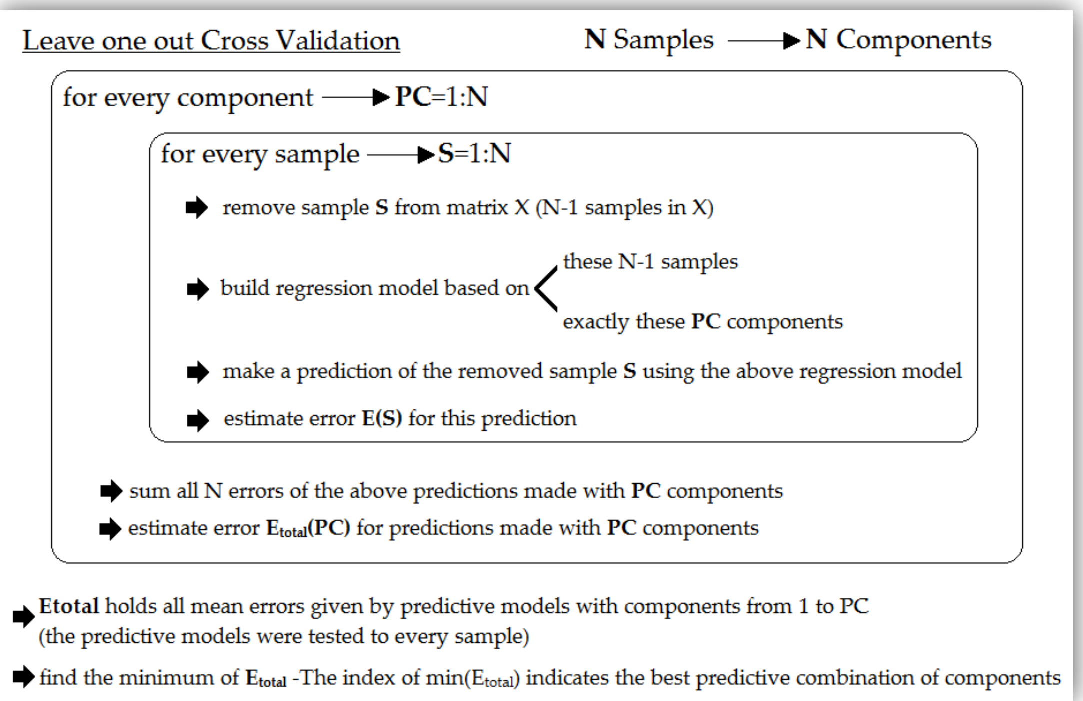
4.4 Cross validation

4.4.1 Leave one out cross validation

Cross validation methodology was used in order to evaluate the predictive models. The two different validation tests performed are *leave one out cross validation* and *5-fold cross validation.* Moreover, the validation tests can be run for any loaded set of samples from the *excel database*.

The *leave one out* validation tests are very useful at determining the optimum number of components, in order to build the best predictive model. These predictive models are made and tested twice, separately for PCR and PLSR. Thus, the true relation between the components of a dataset and their predictive ability is revealed by this process, without any need of examining each component's variance separately. The algorithm of the *leave one out validation test* is shown graphically at ***Figure 4.16***

Figure 4.16


Leave one out Cross Validation
N Samples → N Components
for every component → PC=1:N
for every sample → S=1:N
remove sample S from matrix X (N-1 samples in X)
build regression model based on
these N-1 samples
exactly these PC components
make a prediction of the removed sample S using the above regression model
estimate error E(S) for this prediction
sum all N errors of the above predictions made with PC components
estimate error E_total(PC) for predictions made with PC components
Etotal holds all mean errors given by predictive models with components from 1 to PC
(the predictive models were tested to every sample)
find the minimum of E_total -The index of min(E_total) indicates the best predictive combination of components

The process of the Leave one out cross validation test performed for any matrix X with N samples.

Each one of the N samples is removed and then is tested in a predictive model constructed without its contribution. This is done for every possible serial combination of components, starting the first time with the first one that collects the most variance (aka $PC = 1$), then the second time with the first and the second components (aka $PC = \{1\ ,2\}$) etc. For every sample, the prediction error is defined as:

$$E(s) = \sqrt{\left|value_{predicted} - value_{real}\right|^{2}}\ ,s = 1, \dots, N \qquad \textit{Equation 4.3}$$

Furthermore, the total error for every component is defined as:

$$E_{total}(pc) = \frac{1}{N}\sum_{s=1}^{N} E(s), pc = 1, \dots, N \qquad \text{Equation 4.4}$$

Once the *leave one out validation test* ends, the matrix E_{total} has N values. Each value represents the total error that corresponds to a predictive model with exactly pc components, as shown at *Graphical Representation 4-4*:

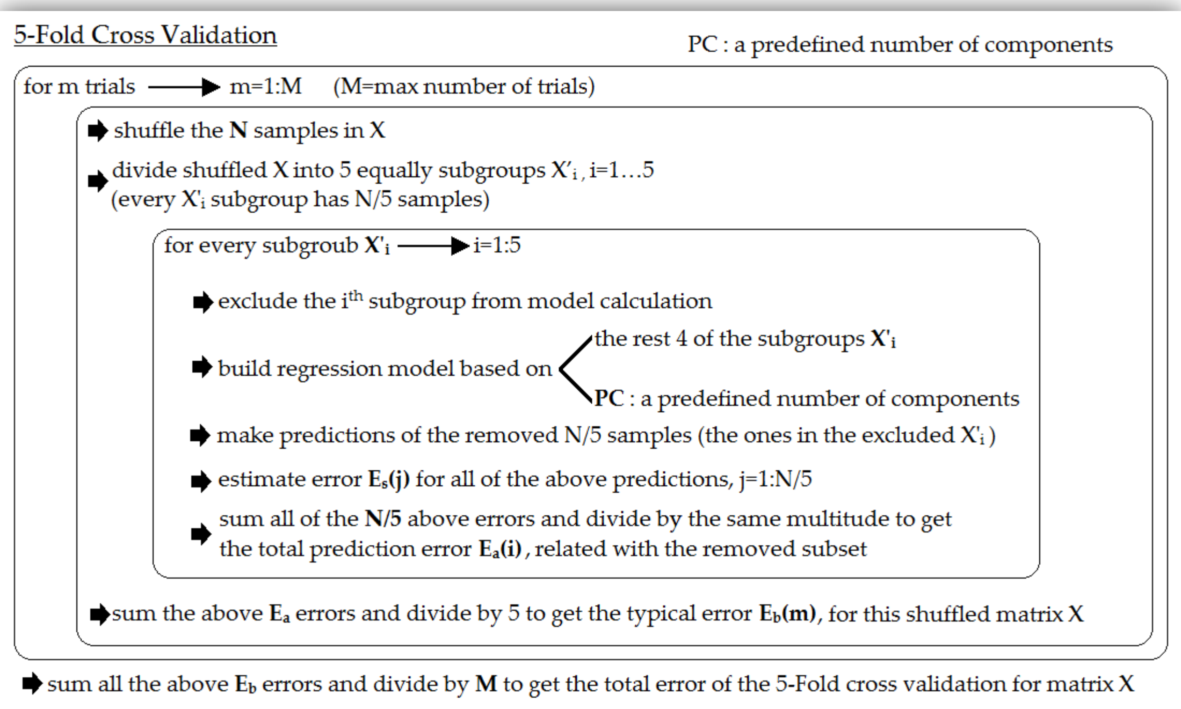
$$\begin{bmatrix} E_{total}(1) \\ E_{total}(2) \\ E_{total}(3) \\ \vdots \\ E_{total}(N) \end{bmatrix} \rightarrow \begin{bmatrix} 1^{st} component \\ \{1^{st}, 2^{nd}\} components \\ \{1^{st}, 2^{nd}, 3^{rd}\} components \\ \vdots \\ \{1^{st}, 2^{nd}, 3^{rd}, \dots, N^{th}\} components \end{bmatrix} \qquad \text{Graphical Representation 4-4}$$

, which means that the index of $\min(E_{total})$ corresponds to the ideal number of components for prediction.

4.4.2 5-FOLD CROSS VALIDATION

On the other hand, the *5-Fold cross validation tests* are useful for verification of the pre-mentioned validation test. At this test, a predefined number of components is tested in a complicated process with a dataset; we may choose the number of components to be what the *leave one out validation* method indicates as optimum. *5-Fold cross validation* also helps at discovering any systematicity, regarding high/low or unexpected prediction error, at particular samples. The graphical representation of this test is shown below at ***Figure 4.17***

Figure 4.17



```
5-Fold Cross Validation                         PC : a predefined number of components
for m trials ⟶ m=1:M    (M=max number of trials)
    ➧ shuffle the N samples in X
    ➧ divide shuffled X into 5 equally subgroups X'_i , i=1…5
      (every X'_i subgroup has N/5 samples)
        for every subgroub X'_i ⟶ i=1:5
            ➧ exclude the i^th subgroup from model calculation
            ➧ build regression model based on  < the rest 4 of the subgroups X'_i
                                                < PC : a predefined number of components
            ➧ make predictions of the removed N/5 samples (the ones in the excluded X'_i )
            ➧ estimate error E_s(j) for all of the above predictions, j=1:N/5
            ➧ sum all of the N/5 above errors and divide by the same multitude to get
              the total prediction error E_a(i), related with the removed subset
    ➧ sum the above E_a errors and divide by 5 to get the typical error E_b(m), for this shuffled matrix X
➧ sum all the above E_b errors and divide by M to get the total error of the 5-Fold cross validation for matrix X
```

*The process of the **5-Fold cross validation** test performed for any matrix X with N samples.*

This test runs for a specific number of trials, e.g. 10, 50 or even 100 times. At each trial, the tested dataset X is shuffled and then divided to five, almost equal regarding the number of samples within, subgroups. Then every subgroup is tested at a regression model built without its contribution and at the same time the overall prediction error, for all of the $^{N}/_{5}$ samples of this subgroup, is calculated. In further detail, the prediction error E_s for every sample in a tested subgroup is defined as similarly to *Equation 4.3*:

$$E_s(j) = \left|value_{predicted} - value_{real}\right|^2 , j = 1, \dots , {}^{N}/_{5} \qquad \text{Equation 4.5}$$

, the error E_a for every tested subgroup is defined as:

$$E_a(i) = \frac{5}{N}\sum_{j=1}^{N/5} E_s(j) \ , i = 1, \dots ,5 \qquad \text{Equation 4.6}$$

, the error E_b for every trial of a shuffled matrix X is defined as:

$$E_b(m) = \frac{1}{5}\sum_{i=1}^{5} E_a(i) \ \ , m = 1, \dots , M \qquad \text{Equation 4.7}$$

, the total error of the *5-fold validation test* for M trials is defined:

$$E_{total} = \frac{1}{M}\sum_{m=1}^{M} E_b(m) \qquad \text{Equation 4.8}$$

With every *data scenario* loaded we performed a different validation test. However, these validation tests need to be run once for every regression method, PCR and PLSR.

As mentioned at the previous chapters, the choice of which PCA method is used does not play any role at the produced result, so it is up to us to choose one of the three introduced at chapter 2. The PCA model, needed in order to build PCR's regression model, was computed according to the 2nd presented approach, meaning the one that involves the *Singular Vale Decomposition,* because Matlab's implemented conversion to SVD is slightly quicker than the other methods.

5. Results

In this chapter we show all the results obtained by every method, for every possible setup. One setup, for example, could be a regression model made with data from *scenario 1*, with pretreated data, without any removed spectra areas and a PCR algorithm. Thus a table showing the possible setups is shown below:

Entire Spectrum Range																Modified Spectrum Range															
No Pretreatment								Pretreated Data								No Pretreatment								Pretreated Data							
PCR				PLS				PCR				PLS				PCR				PLS				PCR				PLS			
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

Here we should note that:

- *Entire Spectrum Range* means that the whole spectrum range, from $650\ cm^{-1}$ to $4000\ cm^{-1}$, was included.
- *Modified Spectrum Range* means that only the spectrum data in regions ($650\ cm^{-1}$ to $2000\ cm^{-1}$) and ($2500\ cm^{-1}$ to $3300\ cm^{-1}$) were included.
- *No Pretreatment* means that the data were used in their original form
- *Pretreated Data* means that the data were corrected by extracting the mean out of each sample as shown in *Graphical Representation 4-3*.
- *1,2,3,4* are the four scenarios created in order to evaluate models created from samples that belong in different families of petroleum products and responses that come from different analytical methods.

Using the data from *scenario 1*, helps us estimate how a prediction model, made from a confused dataset, would correspond to predictions when a random petroleum product is tested. In addition, *scenario 2* and *scenario 3* contain measurements from *ASTM D 2549* and *HPLC ASTM D 7419-07* methods respectively, but at the same time the types of the samples vary considerably. *Scenario 4* contains gasoline and diesel oil samples that all have been measured by the *ASTM D 1319* analytical method. *Scenario 4* is a, rather appropriate, dataset containing samples that generally could lead to better predictive models, as far as the new samples are of the same petroleum family.

All of the above setups were validated via the *Leave one out Cross Validation* method. Furthermore, each test indicated an optimum component selection, which was then validated using the *5-fold Cross Validation* method. *Leave one out cross validation* results indicate the minimum *RMSEP,* which leads to the optimum number of components. For every tested sample, the prediction error is computed from a regression model made without its contribution.

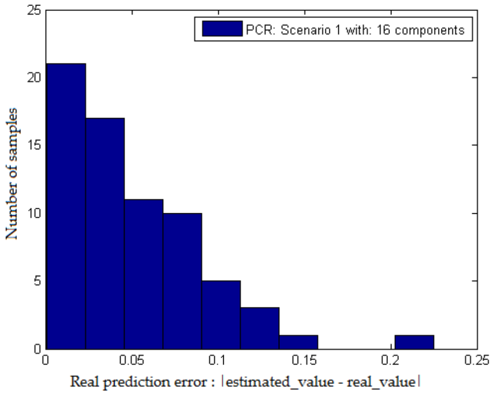
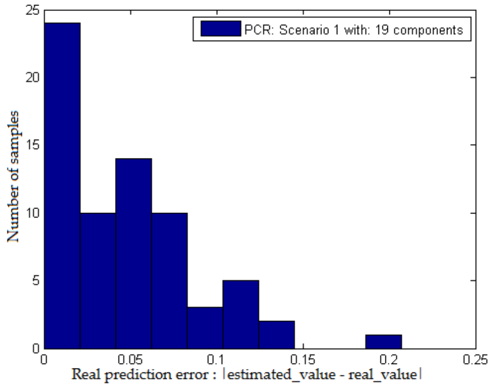
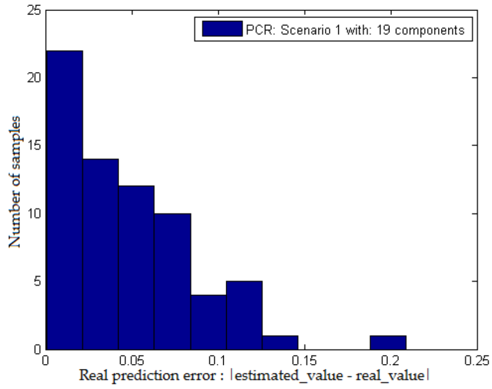
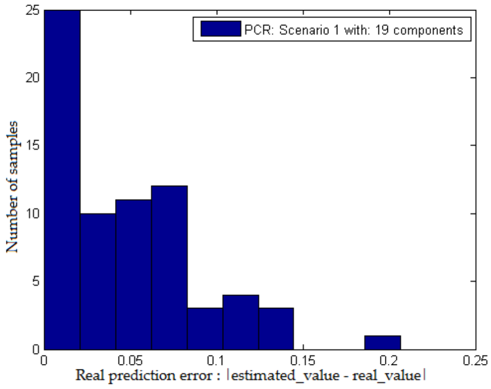
Leave one out cross validation results	**PCR testing within entire spectrum range**			
	NO pretreatment		**Pretreated Data**	
	RMSEP	**#components**	***RMSEP***	**#components**
Scenario_1	0.0076	16	0.0076	19
Scenario_2	0.0116	6	0.0113	7
Scenario_3	0.0132	8	0.0145	6
Scenario_4	0.0051	16	0.0050	16

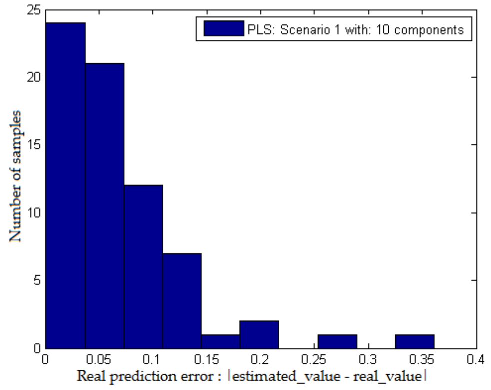
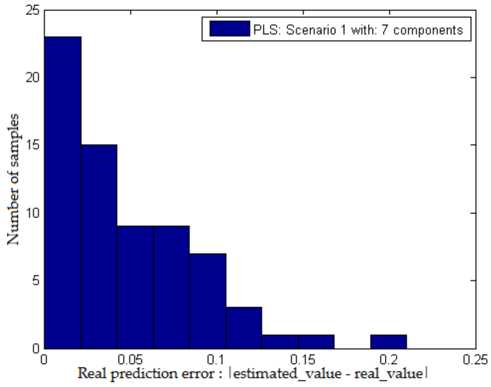
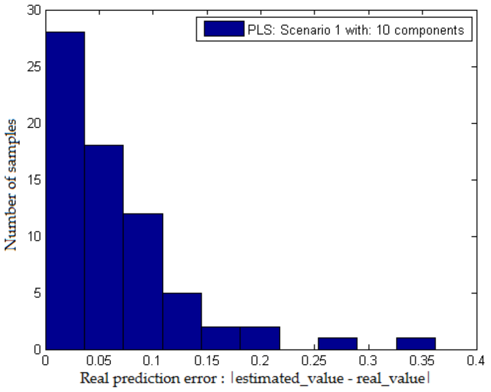
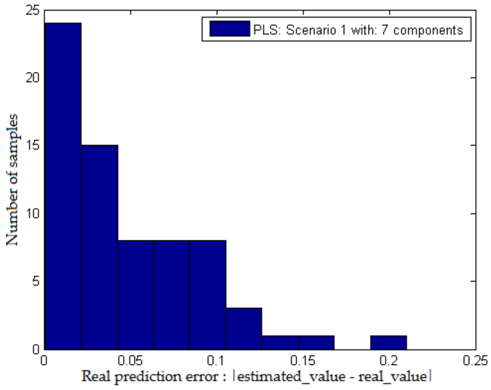
Leave one out cross validation results	**PCR testing within modified spectrum range**			
	NO pretreatment		**Pretreated Data**	
	RMSEP	**#components**	***RMSEP***	**#components**
Scenario_1	0.0074	19	0.0074	19
Scenario_2	0.0114	7	0.0111	6
Scenario_3	0.0143	8	0.0149	6
Scenario_4	0.0056	14	0.0059	14

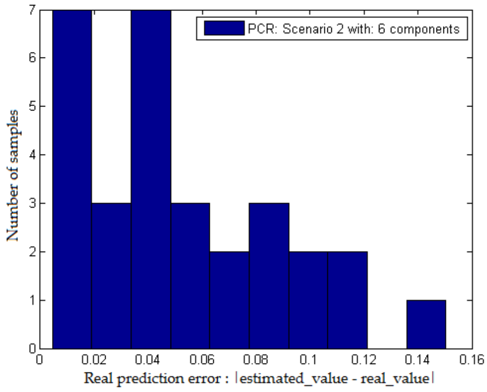
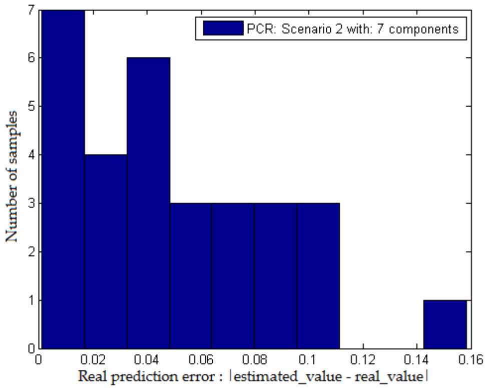
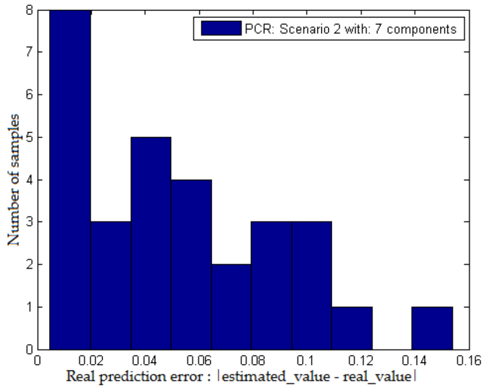
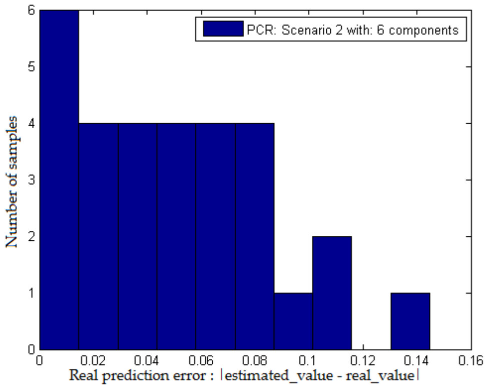
Leave one out cross validation results	**PLS testing within entire spectrum range**			
	NO pretreatment		**Pretreated Data**	
	RMSEP	**#components**	***RMSEP***	**#components**
Scenario_1	0.0111	10	0.0077	7
Scenario_2	0.0284	4	0.0121	6
Scenario_3	0.0278	3	0.0154	4
Scenario_4	0.0279	15	0.0047	9

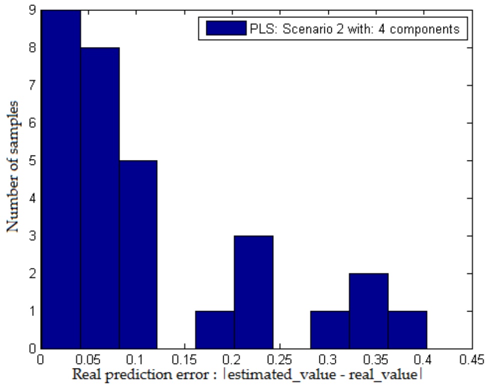
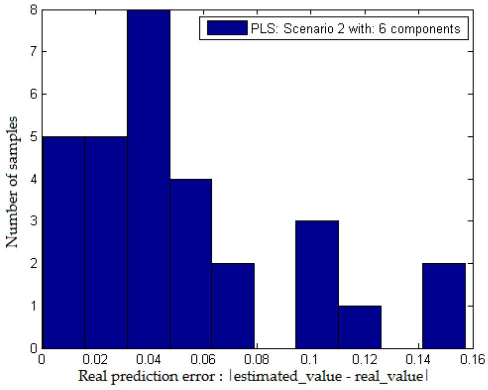
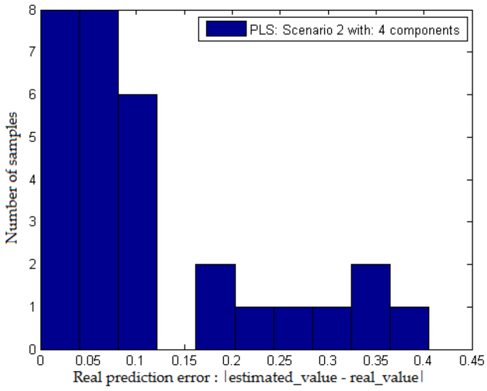
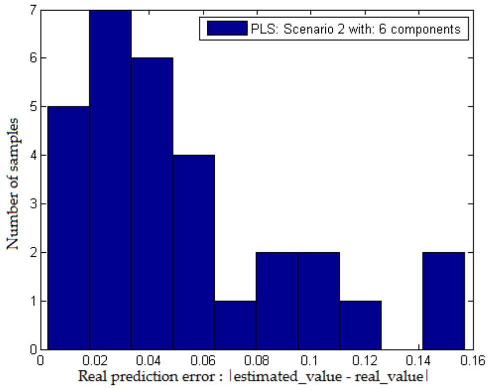
Leave one out cross validation results	**PLS testing within modified spectrum range**			
	NO pretreatment		**Pretreated Data**	
	RMSEP	**#components**	***RMSEP***	**#components**
Scenario_1	0.0109	10	0.0077	7
Scenario_2	0.0281	4	0.0120	6
Scenario_3	0.0283	3	0.0152	4
Scenario_4	0.0249	15	0.0052	8

For every setup, a histogram is presented, showing the real errors in prediction when the regression model is based on the indicated best number of components.

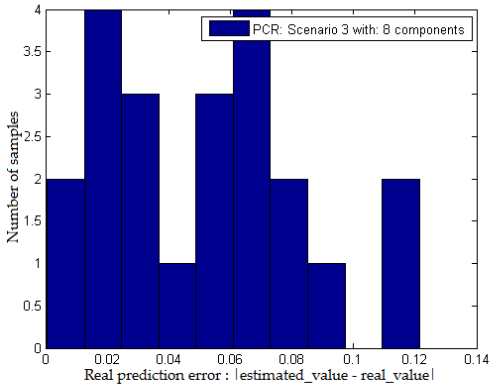
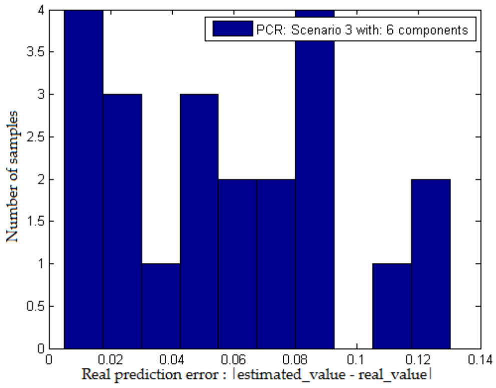
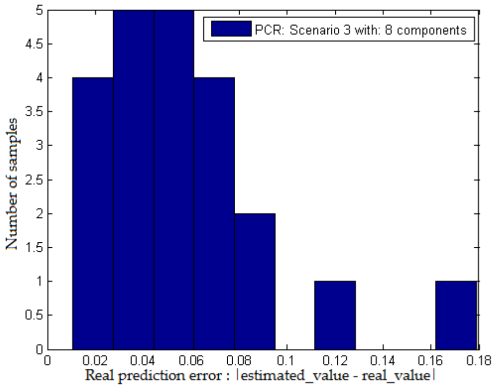
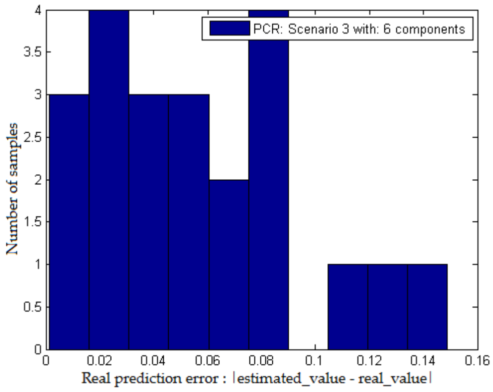
Scenario 1 : PCR		
	NO pretreatment	**Pretreated Data**
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

Scenario 1 : PLS		
	NO pretreatment	**Pretreated Data**
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

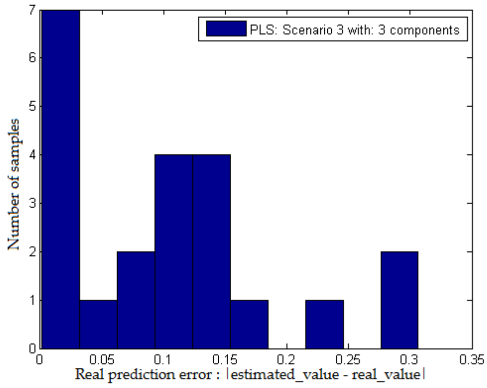
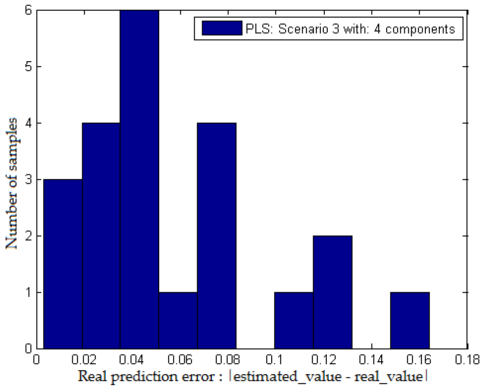
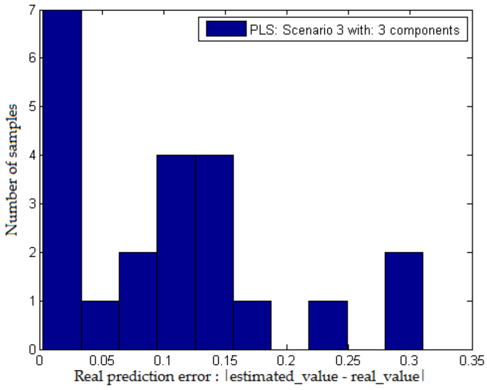
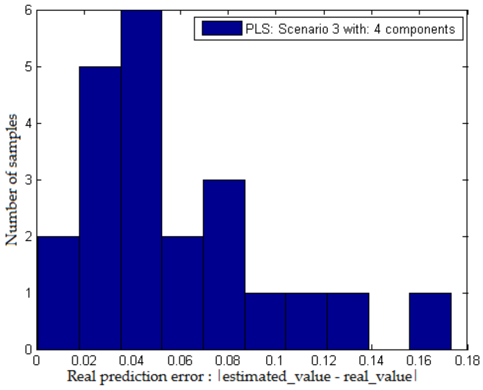
Scenario 2 **: PCR**		
	NO pretreatment	**Pretreated Data**
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

Scenario 2 **: PLS**		
	NO pretreatment	**Pretreated Data**
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

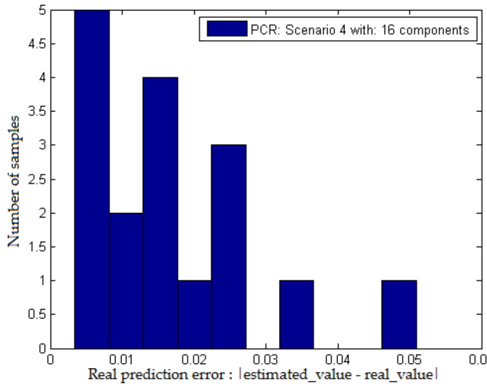
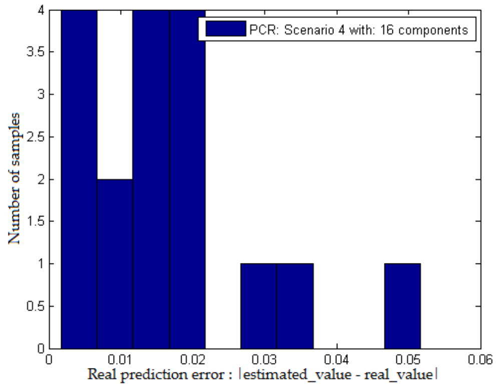
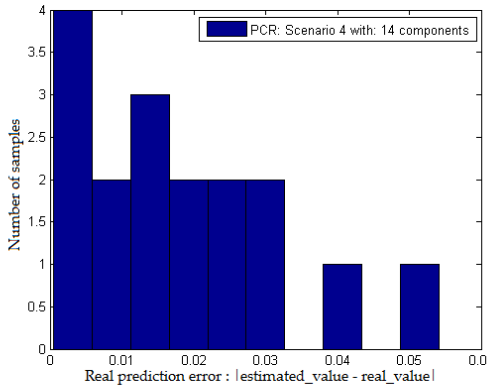
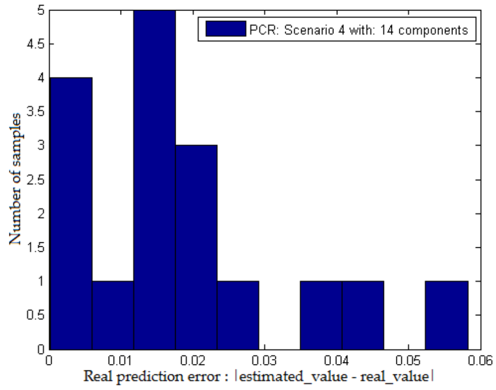
Scenario 3 : PCR

	NO pretreatment	Pretreated Data
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

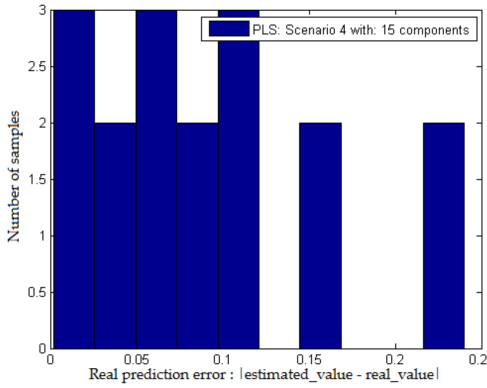
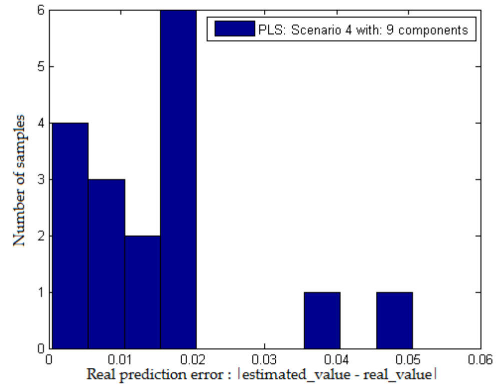
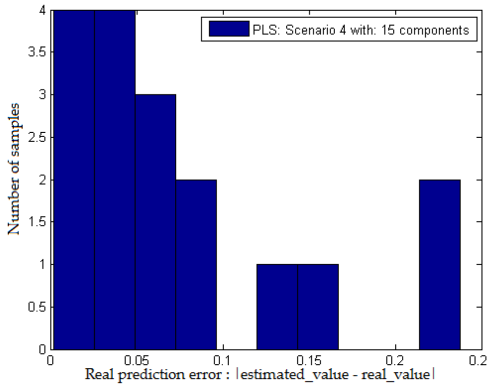
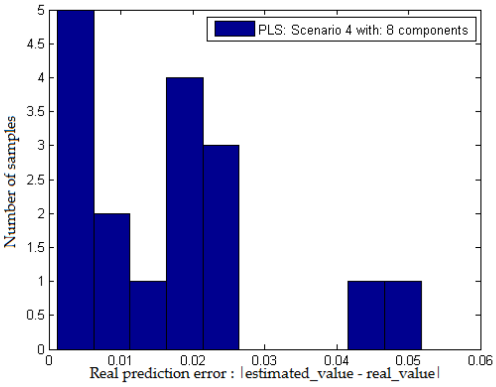
Scenario 3 : PLS

	NO pretreatment	Pretreated Data
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

Scenario 4 : PCR

	NO pretreatment	Pretreated Data
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

Scenario 4 : PLS

	NO pretreatment	Pretreated Data
FULL SPECTRUM RANGE		
MODIFIED SPECTRUM RANGE		

Every optimum component selection as indicated from *leave one out cross validation* tests above, was then validated with *5-fold cross validation.* For every scenario, a *5-fold cross validation test* was performed, with 100 iterations. This means that each dataset examined was randomly shuffled 100 times, each time providing 5 different *training sets* and 5 different *test sets.* The results are shown below:

	PCR			
	Full spectrum range		**Modified spectrum range**	
	NO Pretreatment	**Pretreatment**	**NO Pretreatment**	**Pretreatment**
Scenario 1	0.0185	0.0188	0.0186	0.0191
Scenario 2	0.0370	0.0302	0.0341	0.0298
Scenario 3	0.0321	0.0314	0.3644	0.3054
Scenario 4	0.0097	0.0098	0.0107	0.0094
	PLS			
	Full spectrum range		**Modified spectrum range**	
	NO Pretreatment	**Pretreatment**	**NO Pretreatment**	**Pretreatment**
Scenario 1	0.0344	0.0297	0.0298	0.0297
Scenario 2	0.0357	0.0305	0.1172	0.0294
Scenario 3	0.0298	0.0289	0.1079	0.0284
Scenario 4	0.0380	0.0084	0.0356	0.0087

5.1 RESULT EXPLANATION-DISCUSSION

PCR predictive models with pretreated data do not differ very much from the models made without pretreatment. In addition, whether the spectrum range of the dataset is full or not, in PCR predictive models there are minor differences, concerning the *RMSEP* and the optimum number of components.

On the other hand, PLS provides predictions models with slightly better *RMSEP* and at the same time with significant fewer components than PCR. Furthermore, removing some of the spectra improves even more the results as it can be seen from the histogram figures. However, it is clear that PLS cannot create efficient models when the data are not pretreated. Concerning this study, the predictive models of PLS are entitled to be declared better than the ones of PCR, because they provide smaller RMSEP with less components.

Scenario 1 performs decently for the majority of the samples within it. Regarding PCR, *scenario 1* requires more components from *scenarios 2 & 3*; however, it provides smaller *RMSEP*. In PLS with pretreated data, *scenario 1* provides smaller *RMSEP* than *scenarios 2 & 3*, but the number of components does not differ as much as differs in PCR.

Scenario 2 and *scenario 3* have the largest predictions errors with the smallest number of components. That is reasonable because the type of samples participated in these *scenarios* are very different amongst them. Comparing to *scenario 1*, which also has various sample types, *scenarios 2 & 3* suffer that maximum *RMSEP* due to the total number of samples involved. Recall that *scenario 1* contains 69 samples while *scenarios 2 & 3* contain 30 and 22 samples respectively.

Scenario 4 provides models with minimum *RMSEP* for both methods. Furthermore, concerning PLS, according to the histograms, almost every sample is predicted with a real error below 2%, providing the best regression/predictions results in our study. That is something we expected, since this scenario contains spectra measurements for only gazolines and diesel oils and the responses are derived from one particular analytical method.

6. Conclusion – Further Research

Our study showed that PLS can provide more robust prediction models than PCR, when the target response is the concentration of aromatics hydrocarbons. An effort to make a global prediction model, which can predict the aromatics' concentration of almost any petroleum product, has been made. In addition to that, a tool has been designed in order to be easier for an analyst to examine all of the samples, as well as the analytical methods themselves.

It is mandatory to understand that the efficiency of a prediction model is analogous to the initial dataset and to the sequence of the mathematical tools applied to it. The golden ratio between the right dataset and the right mathematical tools is also different when different characteristics are examined. In other words, the best regression model cannot be defined without exploring all the different combinations of either analytical methods or chemometrics methods.

Finding a general prediction model is a complicated thing. As we have seen by the results, pretreating the data does not always provide better results, as in case of PCR. Moreover, PCR predictive models seem to be less effective when the data are pretreated. Other pretreatments methods applied to data may enhance the regression models based on PCR, but we cannot be sure if a pretreatment method can be as effective in PCR, as in PLS, or vice versa.

Due to the successive measurements of numerous data *scenarios* that took place in this study, we concluded that larger datasets don't always provide better predictive models. A dataset having small number of samples, that are of the same type, may produce better predictive models when the targeted unknown samples are of the same origin.

Regarding the removal of spectra areas, this study has shown that there is no significant difference whether the spectrum range is full or not. However, a further research could be done, considering selecting small spectra regions, generally the ones that aromatics' concentration is highly associated, as described in chapter 1.3 -Aromatics in MIR.

Furthermore, a variety of different extensions of PLS methods, as described in chapter 1.4, could be implemented for this tool. However, there are many methods, besides PCR and PLS, that could also be implemented, such as the Regularized Discriminant Analysis (RDA), Soft Independent Modeling of Class Analogy (SIMCA), K-Nearest Neighbor (KNN), Multilayer Perception (MLP), Support Vector Machines (SVM), Probabilistic Neural Network (PNN), Genetic Inside Neural Network (GINN), Multi-Layer Feed forward Neural Network (MLF), Linear Discriminant Analysis (LDA) and Bayesian Classifiers [41].

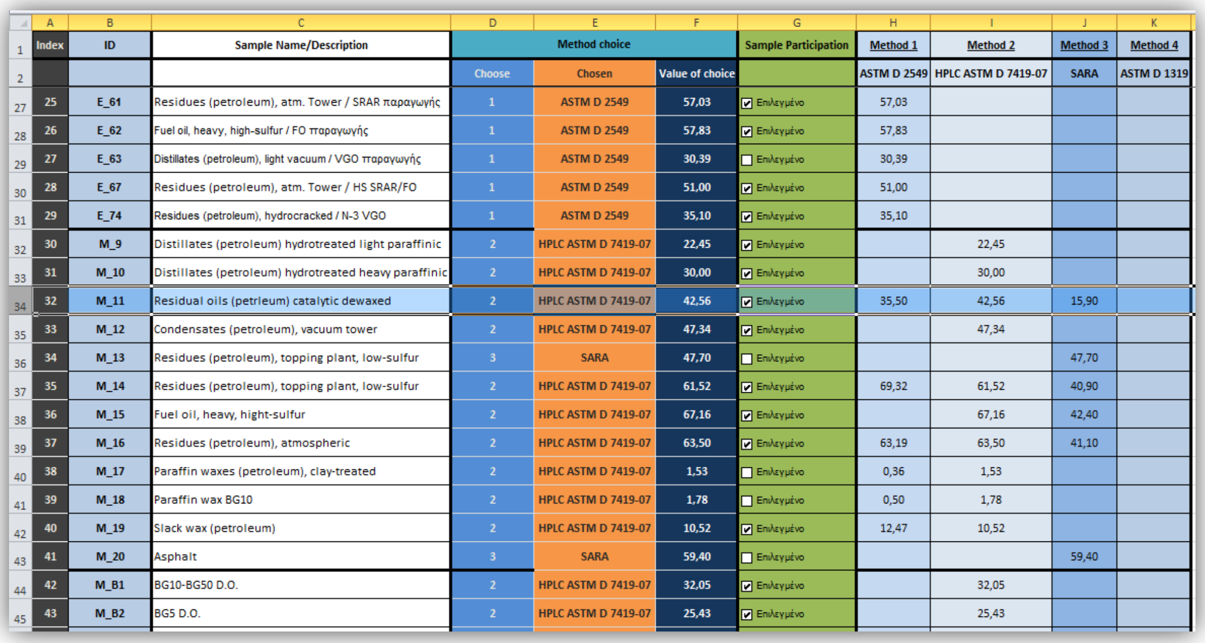
Moreover, if a tool, such the one presented in this study, is developed and standardized for certain types of samples or for certain types of measurements, it could provide a tool for evaluating future experimental measurements of certain responses, in cases where human operators are involved.

7. Appendix

7.1 Samples

7.1.1 The excel database

Figure 7.1



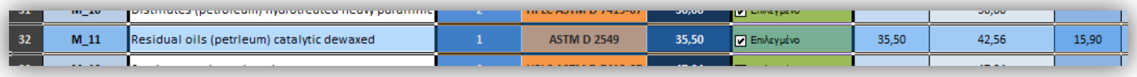
	A	B	C	D	E	F	G	H	I	J	K
1	Index	ID	Sample Name/Description	Method choice			Sample Participation	Method 1	Method 2	Method 3	Method 4
2				Choose	Chosen	Value of choice		ASTM D 2549	HPLC ASTM D 7419-07	SARA	ASTM D 1319
27	25	E_61	Residues (petroleum), atm. Tower / SRAR παραγωγής	1	ASTM D 2549	57,03	☑ Επιλεγμένο	57,03			
28	26	E_62	Fuel oil, heavy, high-sulfur / FO παραγωγής	1	ASTM D 2549	57,83	☑ Επιλεγμένο	57,83			
29	27	E_63	Distillates (petroleum), light vacuum / VGO παραγωγής	1	ASTM D 2549	30,39	☐ Επιλεγμένο	30,39			
30	28	E_67	Residues (petroleum), atm. Tower / HS SRAR/FO	1	ASTM D 2549	51,00	☑ Επιλεγμένο	51,00			
31	29	E_74	Residues (petroleum), hydrocracked / N-3 VGO	1	ASTM D 2549	35,10	☑ Επιλεγμένο	35,10			
32	30	M_9	Distillates (petroleum) hydrotreated light paraffinic	2	HPLC ASTM D 7419-07	22,45	☑ Επιλεγμένο		22,45		
33	31	M_10	Distillates (petroleum) hydrotreated heavy paraffinic	2	HPLC ASTM D 7419-07	30,00	☑ Επιλεγμένο		30,00		
34	32	M_11	Residual oils (petrleum) catalytic dewaxed	2	HPLC ASTM D 7419-07	42,56	☑ Επιλεγμένο	35,50	42,56	15,90	
35	33	M_12	Condensates (petroleum), vacuum tower	2	HPLC ASTM D 7419-07	47,34	☑ Επιλεγμένο		47,34		
36	34	M_13	Residues (petroleum), topping plant, low-sulfur	3	SARA	47,70	☐ Επιλεγμένο			47,70	
37	35	M_14	Residues (petroleum), topping plant, low-sulfur	2	HPLC ASTM D 7419-07	61,52	☑ Επιλεγμένο	69,32	61,52	40,90	
38	36	M_15	Fuel oil, heavy, hight-sulfur	2	HPLC ASTM D 7419-07	67,16	☑ Επιλεγμένο		67,16	42,40	
39	37	M_16	Residues (petroleum), atmospheric	2	HPLC ASTM D 7419-07	63,50	☑ Επιλεγμένο	63,19	63,50	41,10	
40	38	M_17	Paraffin waxes (petroleum), clay-treated	2	HPLC ASTM D 7419-07	1,53	☐ Επιλεγμένο	0,36	1,53		
41	39	M_18	Paraffin wax BG10	2	HPLC ASTM D 7419-07	1,78	☐ Επιλεγμένο	0,50	1,78		
42	40	M_19	Slack wax (petroleum)	2	HPLC ASTM D 7419-07	10,52	☑ Επιλεγμένο	12,47	10,52		
43	41	M_20	Asphalt	3	SARA	59,40	☐ Επιλεγμένο			59,40	
44	42	M_B1	BG10-BG50 D.O.	2	HPLC ASTM D 7419-07	32,05	☑ Επιλεγμένο		32,05		
45	43	M_B2	BG5 D.O.	2	HPLC ASTM D 7419-07	25,43	☑ Επιλεγμένο		25,43		

A part of the excel database

Note that there are until now 72 samples. Each row corresponds to a sample. Every sample has:

- Column *ID*: A unique identifier
- Column *Sample Name/Description:* A name or a short description about its content
- Column *Method choice:*
  - Column *Choose:* Editable field, with the number of the chosen method
  - Column *Chosen:* Name of the chosen method
  - Column *Value of choice:* The value of the chosen method
- Column *Sample Participation:* Editable field, with a mark for selecting/unselecting a sample
- Columns *Method 1 - Method 4:* Sample's value for each method[11]

Figure 7.2



32	M_11	Residual oils (petrleum) catalytic dewaxed	1	ASTM D 2549	35,50	☑ Επιλεγμένο	35,50	42,56	15,90

[11] There are samples which are tested with only one method, as there are others which are tested with more than one.

The highlighted sample of **Figure 7.1***, with a different chosen method*

7.1.2 COMPLETE LIST OF SAMPLES

1	Fuel oil, residual / Bunker No3	38	Paraffin waxes
2	Asphalt	39	Paraffin wax
3	Distillates (petroleum), vacuum	40	Slack wax (petroleum)
4	Fuel oil, residual	41	Asphalt
5	Residues (petroleum)	42	Deasphalted oil BG10-BG50
6	Fuel oil	43	Deasphalted oil BG5
7	Distillates (petroleum), light vacuum / VGO	44	BG5 Raffinate
8	Residues (petroleum), catalytic cracking	45	BG10-BG50 Raffinate
9	Residues (petroleum)	46	Fuel oil
10	Gas oils (petroleum), hydrotreated vacuum / VGO	47	BG10-BG50 Extract
11	Residues (petroleum), vacuum	48	BG5 Extract
12	LVGO	49	BG5 Distillate
13	Gas oils (petroleum), heavy vacuum / HVGO	50	BG10-BG50 Distillate
14	Distillates (petroleum), light vacuum	51	Wax residue
15	Paraffin waxes and Hydrocarbon waxes	52	LVGO-HVGO
16	Fuel oil, no. 4	53	VGO
17	Gas oils (petroleum), heavy vacuum / HVGO	54	Fuel oil
18	Distillates (petroleum), light vacuum	55	Fuel oil
19	Gas oils (petroleum), heavy vacuum / HVGO	56	CLO
20	Distillates (petroleum), light vacuum	57	Gasoline
21	Residues (petroleum), catalytic cracking	58	Gasoline
22	Fuel oil, residual / Atm	59	GASOIL
23	Residues (petroleum)/ Atm	60	GASOLINE
24	Fuel oil, residual	61	GASOLINE
25	Residues (petroleum)/ atm	62	Diesel
26	Fuel oil, heavy	63	Jet Fuel
27	Distillates (petroleum), light vacuum / VGO	64	Diesel
28	Residues (petroleum)/ Atm	65	Diesel
29	Residues (petroleum), hydrocracked / VGO	66	Diesel
30	Distillates (petroleum) hydrotreated light paraffinic	67	Gasoline
31	Distillates (petroleum) hydrotreated heavy paraffinic	68	Diesel
32	Residual oils (petroleum) catalytic dewaxed	69	Diesel
33	Condensates (petroleum), vacuum	70	Gasoline
34	Residues (petroleum)	71	Gasoline
35	Residues (petroleum)	72	Gasoline
36	Fuel oil, heavy	73	Gasoline
37	Residues (petroleum)/ Atm	74	Diesel

7.1.3 Scenarios

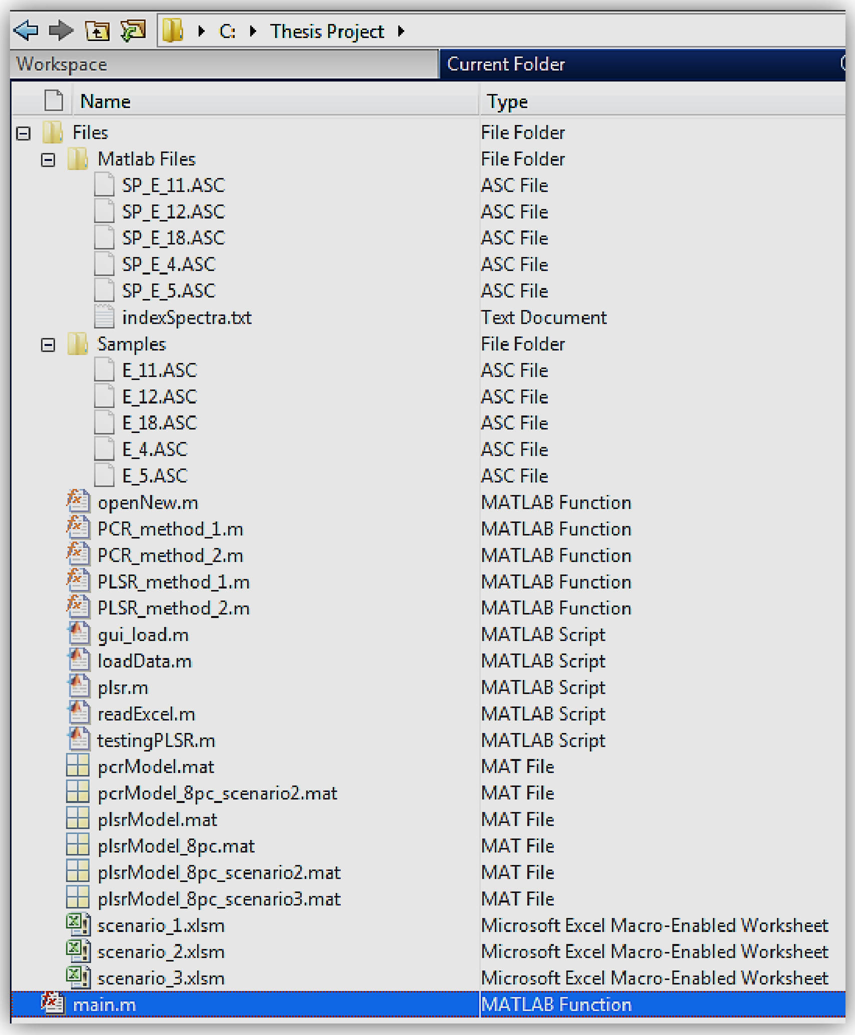
Each scenario contains the following samples by *ID*.

Scenario 1							Scenario 2			Scenario 3			Scenario 4	
1	2	3	4	5	6	7	1	3	4	30	31	32	57	58
8	9	10	11	12	13	14	5	6	7	33	35	36	60	61
16	17	18	19	20	21	22	8	10	12	37	40	42	62	63
23	24	25	26	27	28	29	13	14	16	43	44	45	64	65
30	31	32	33	34	35	36	17	18	19	46	47	48	66	67
37	40	41	42	43	44	45	20	21	22	49	50	52	68	69
46	47	48	49	50	52	53	24	25	26	53	54	55	70	71
54	55	56	57	58	60	61	27	28	29	56			72	73
62	63	64	65	66	67	68	32	35	37				74	
69	70	71	72	73	74		40	45	52					

7.2 Application

7.2.1 Structure

Figure 7.3



Name	Type
Files	File Folder
Matlab Files	File Folder
SP_E_11.ASC	ASC File
SP_E_12.ASC	ASC File
SP_E_18.ASC	ASC File
SP_E_4.ASC	ASC File
SP_E_5.ASC	ASC File
indexSpectra.txt	Text Document
Samples	File Folder
E_11.ASC	ASC File
E_12.ASC	ASC File
E_18.ASC	ASC File
E_4.ASC	ASC File
E_5.ASC	ASC File
openNew.m	MATLAB Function
PCR_method_1.m	MATLAB Function
PCR_method_2.m	MATLAB Function
PLSR_method_1.m	MATLAB Function
PLSR_method_2.m	MATLAB Function
gui_load.m	MATLAB Script
loadData.m	MATLAB Script
plsr.m	MATLAB Script
readExcel.m	MATLAB Script
testingPLSR.m	MATLAB Script
pcrModel.mat	MAT File
pcrModel_8pc_scenario2.mat	MAT File
plsrModel.mat	MAT File
plsrModel_8pc.mat	MAT File
plsrModel_8pc_scenario2.mat	MAT File
plsrModel_8pc_scenario3.mat	MAT File
scenario_1.xlsm	Microsoft Excel Macro-Enabled Worksheet
scenario_2.xlsm	Microsoft Excel Macro-Enabled Worksheet
scenario_3.xlsm	Microsoft Excel Macro-Enabled Worksheet
main.m	MATLAB Function

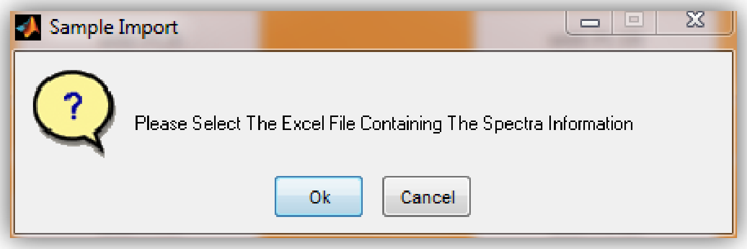
Files structure

The designed application consists of various matlab files shown at ***Figure** 7.3*. The Graphical User Interface (GUI) starts from the file *main.m*. The rest of the functions, scripts, excel files and saved models files are located at the *Files* folder. In *Samples* folder, are located all the files obtained by instruments and they are converted automatically to files containing only variables, stored at *Matlab Files* folder[12].

7.2.2 GUI

The application has a simple GUI. From the main window, shown at ***Figure** 7.6*, one can open an excel file by clicking the *Open Excel File* button. A confirmation window will then appear.

Figure 7.4



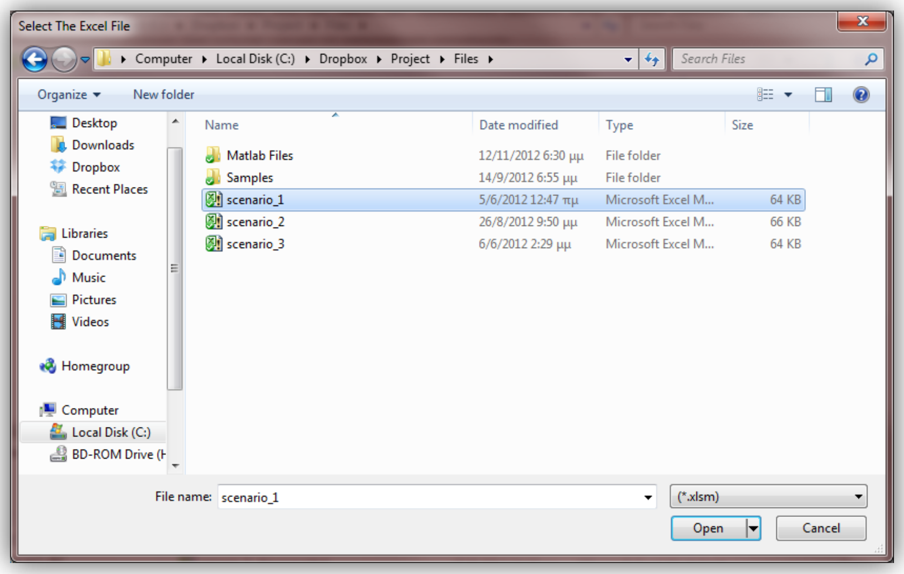
Sample Import

Please Select The Excel File Containing The Spectra Information

*Confirmation dialog box after clicking Open Excel File b*utton

After clicking *Ok* then a pre-saved excel file with a data scenario can be selected as in ***Figure** 7.5*.

Figure 7.5



Select The Excel File

Computer ▸ Local Disk (C:) ▸ Dropbox ▸ Project ▸ Files ▸

Name	Date modified	Type	Size
Matlab Files	12/11/2012 6:30 μμ	File folder	
Samples	14/9/2012 6:55 μμ	File folder	
scenario_1	5/6/2012 12:47 πμ	Microsoft Excel M...	64 KB
scenario_2	26/8/2012 9:50 μμ	Microsoft Excel M...	66 KB
scenario_3	6/6/2012 2:29 μμ	Microsoft Excel M...	64 KB

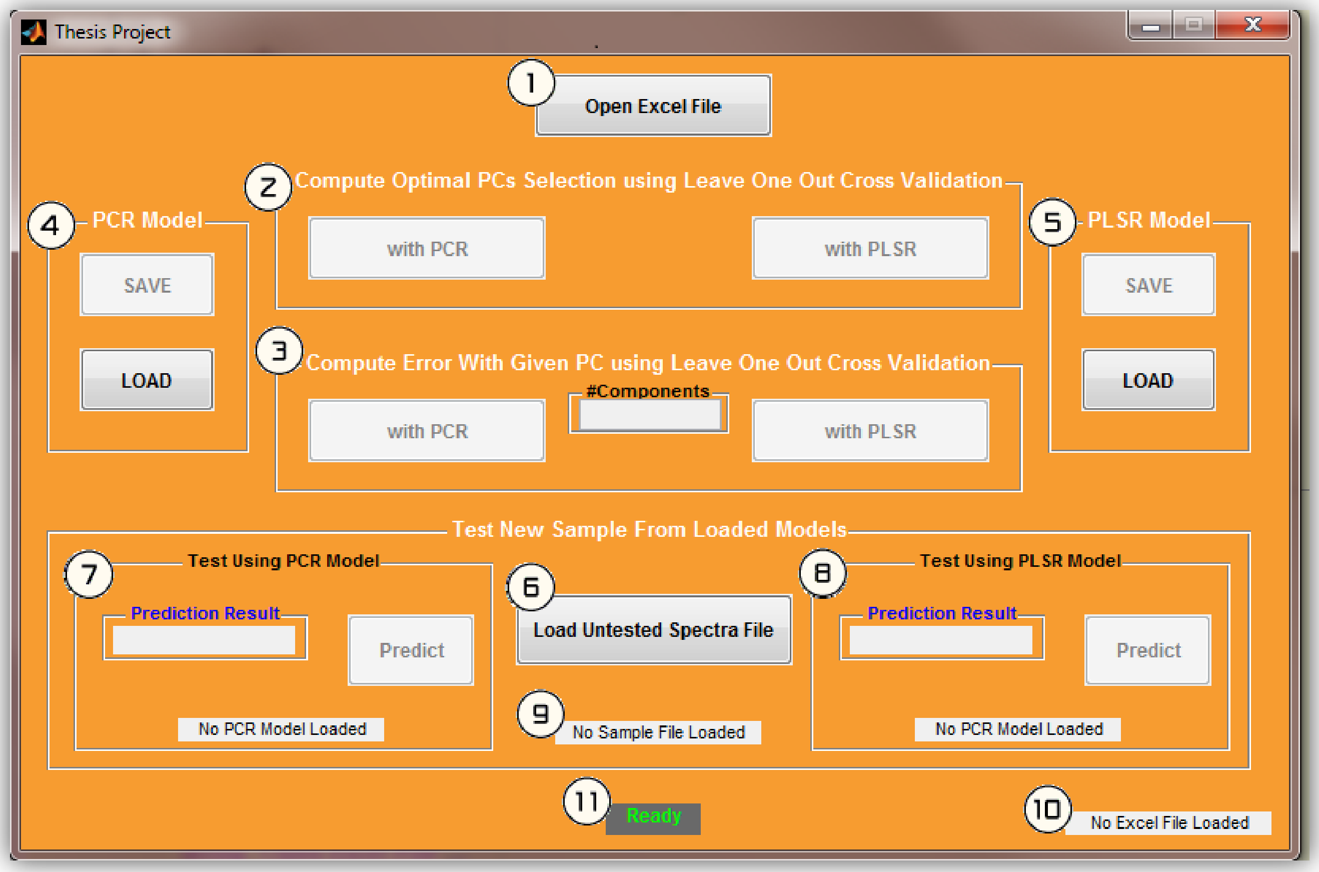
File name: scenario_1

(*.xlsm)

Selection of excel file (scenario)

[12] In ***Figure** 7.3* the majority of the files in folders *Samples* and *Matlab Files* are removed just for the screenshot.

Figure 7.6



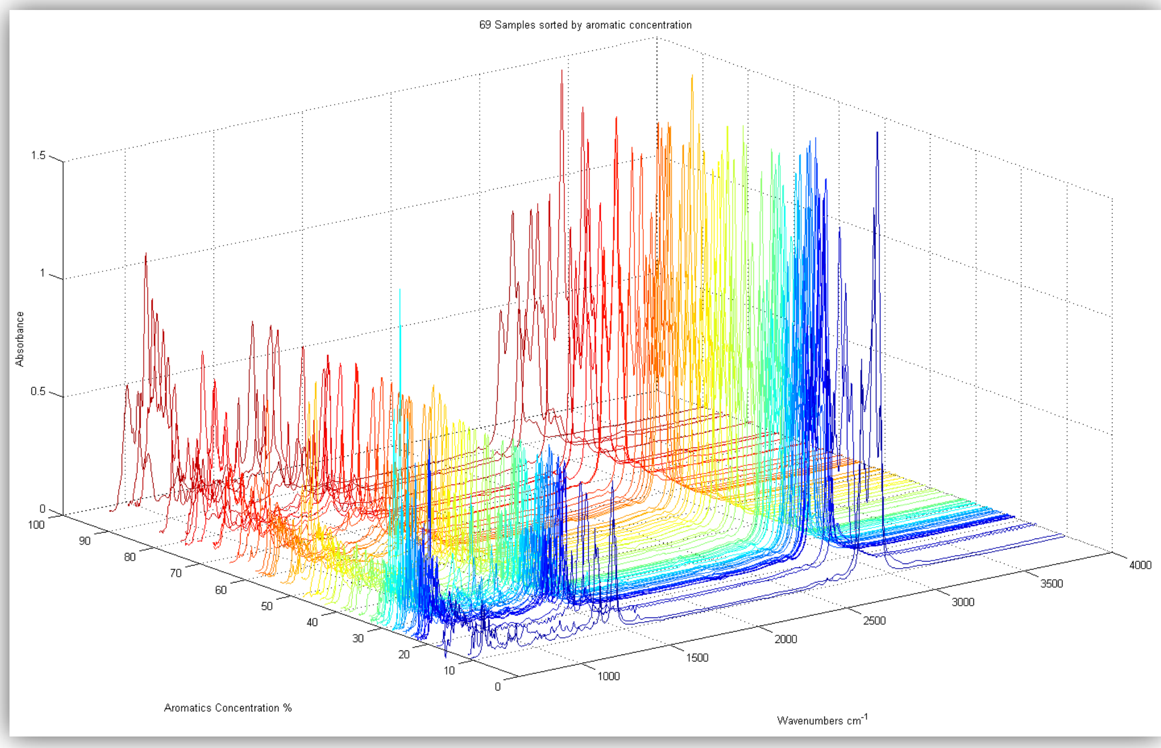
The Graphical User Interface

Labels 1-11 description:

1) *Open Excel File* button for opening an excel file with a data scenario
2) Once a spectra scenario is loaded, this area is for finding the best combination of components for the data of the loaded scenario via *Leave one out cross validation*
3) Once a spectra scenario is loaded, this area is for benchmarking, via *Leave one out cross validation,* with a specified component number.
4) From here a pre-saved PCR regression model can be loaded, and an already computed one, from (2) or (3), can be saved.
5) From here a pre-saved PLS regression model can be loaded, and an already computed one, from (2) or (3), can be saved.
6) *Load Untested Spectra File* button for loading a simple spectrum data for prediction
7) *Test Using PCR Model* is the area for testing/predicting with a PCR model
8) *Test Using PLSR Model* is the area for testing/predicting with a PLSR model
9) Indicator for the test sample loaded
10) Indicator for the excel file loaded
11) Indicator for when matlab is working or idle. When a process is running the indicator is changing from Ready to Busy .

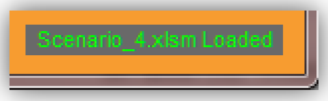
After a scenario is loaded, a plot, containing all the samples of the scenario, is opened for a quick observation and verification of the scenario.

Figure 7.7



After a scenario is loaded, the indicator (10) is displaying the selected file, as shown below:

Figure 7.8

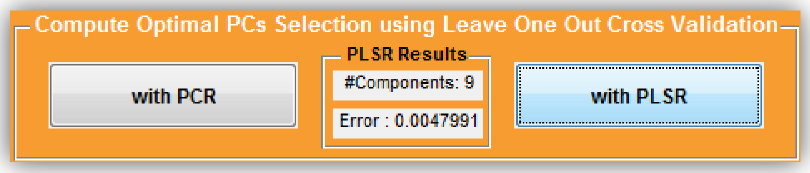


Scenario_4.xlsm Loaded

Indicator for the excel file loaded

Furthermore, the buttons in areas (2) and (3) are active. In area (2) when clicking one of two buttons (*with PCR* or *with PLSR*), after a short time, a small *result* area is shown indicating the minimum *RMSEP* and the optimum component number, as shown in ***Figure 7.9***

Figure 7.9



Compute Optimal PCs Selection using Leave One Out Cross Validation

PLSR Results

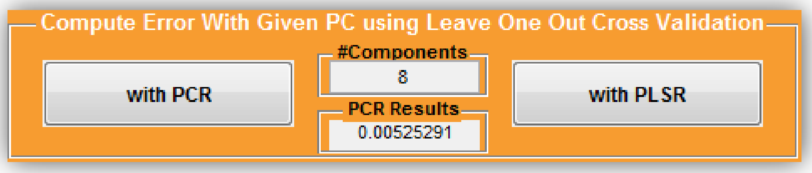
#Components: 9

Error : 0.0047991

Area (2). Find optimal component number, as indicated by the minimum RMSEP

In area (3), when clicking one of two buttons (*with PCR* or *with PLSR*), after entering a desired component number, a small result area is shown indicating the corresponding *RMSEP*, as shown :

Figure 7.10


Compute Error With Given PC using Leave One Out Cross Validation
with PCR
with PLSR

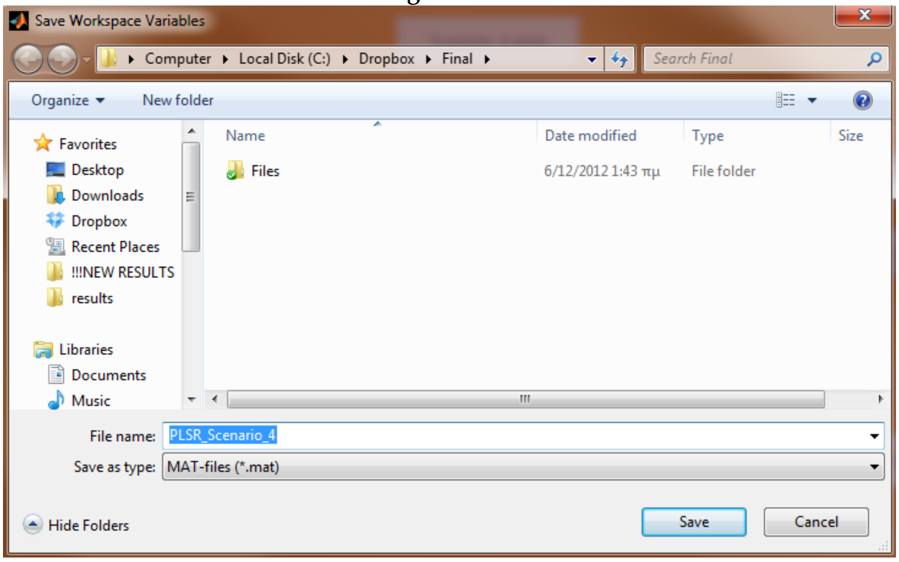
#Components: 8

PCR Results: 0.00525291

Area (3), compute the RMSEP for given components

Having compute a regression model, either from area (2), where the best model is found, or from area (3), where a model with targeted number of components is found, buttons *SAVE* in areas (4) and (5) are active. By clicking one of these, the respective model can be saved, as shown in ***Figure 7.11***

Figure 7.11



File name: PLSR_Scenario_4

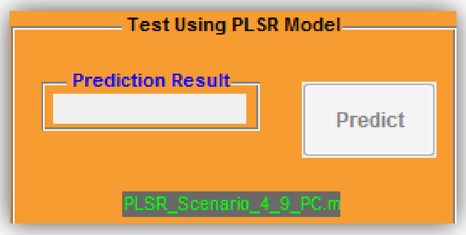
Save as type: MAT-files (*.mat)

Saving the last computed model

Accordingly, by clicking buttons *Load* from areas (2) and (3), a dialog box appears in order to select a pre-saved model for loading, as shown in ***Figure 7.13***.

After a model is loaded the file indicators in areas (7) and (8), whether a PCR or PLS model is loaded respectively, are displaying the name of the loaded model as shown below:

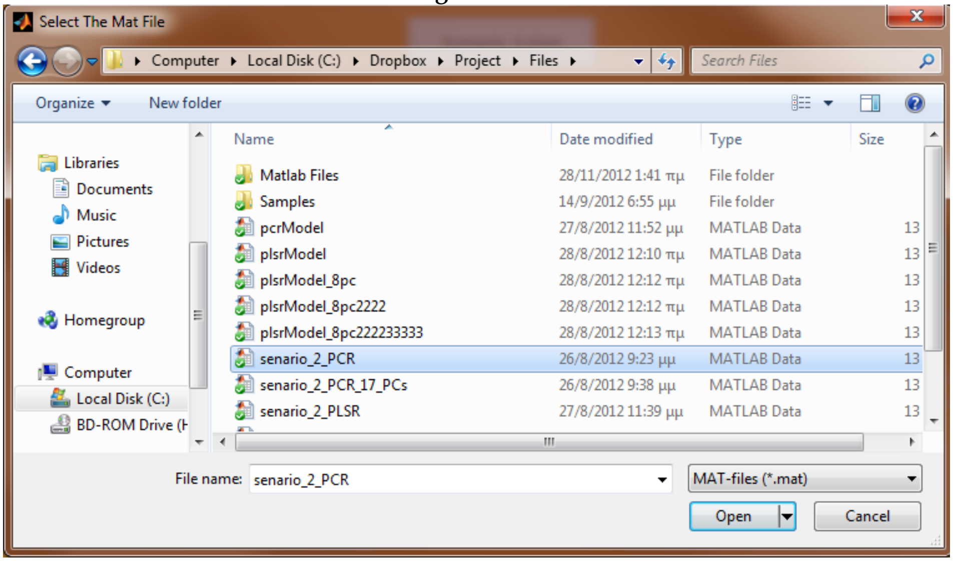
Figure 7.12



PLSR_Scenario_4_9_PC.m

The loaded model indicator

Figure 7.13



Name	Date modified	Type	Size
Matlab Files	28/11/2012 1:41 πμ	File folder	
Samples	14/9/2012 6:55 μμ	File folder	
pcrModel	27/8/2012 11:52 μμ	MATLAB Data	13
plsrModel	28/8/2012 12:10 πμ	MATLAB Data	13
plsrModel_8pc	28/8/2012 12:12 πμ	MATLAB Data	13
plsrModel_8pc2222	28/8/2012 12:12 πμ	MATLAB Data	13
plsrModel_8pc222233333	28/8/2012 12:13 πμ	MATLAB Data	13
senario_2_PCR	26/8/2012 9:23 μμ	MATLAB Data	13
senario_2_PCR_17_PCs	26/8/2012 9:38 μμ	MATLAB Data	13
senario_2_PLSR	27/8/2012 11:39 μμ	MATLAB Data	13

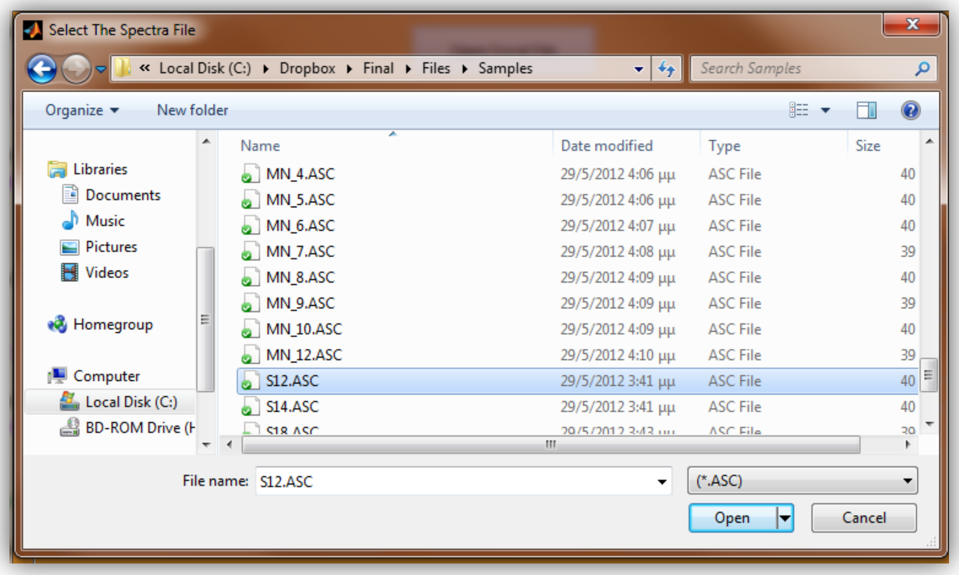
File name: senario_2_PCR

MAT-files (*.mat)

Loading a pre-saved prediction model

To predict a new sample's response, click the *Load Untested Spectra File* button (5). A dialog box for selecting the sample to load is appeared, as shown below:

Figure 7.14



Name	Date modified	Type	Size
MN_4.ASC	29/5/2012 4:06 μμ	ASC File	40
MN_5.ASC	29/5/2012 4:06 μμ	ASC File	40
MN_6.ASC	29/5/2012 4:07 μμ	ASC File	40
MN_7.ASC	29/5/2012 4:08 μμ	ASC File	39
MN_8.ASC	29/5/2012 4:09 μμ	ASC File	40
MN_9.ASC	29/5/2012 4:09 μμ	ASC File	39
MN_10.ASC	29/5/2012 4:09 μμ	ASC File	40
MN_12.ASC	29/5/2012 4:10 μμ	ASC File	39
S12.ASC	29/5/2012 3:41 μμ	ASC File	40
S14.ASC	29/5/2012 3:41 μμ	ASC File	40
S18.ASC	29/5/2012 3:43 μμ	ASC File	39

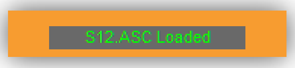
File name: S12.ASC

(*.ASC)

Selecting new sample for prediction

After the new sample is loaded the corresponding indicator (9) is displaying the name of the loaded sample, as shown below:

Figure 7.15

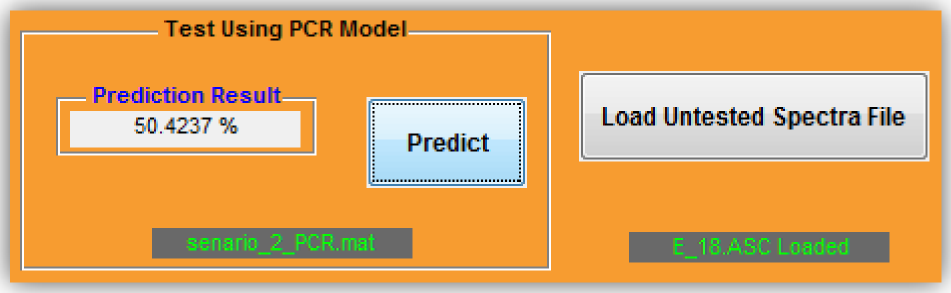


S12.ASC Loaded

Loaded sample indicator

Whenever a test sample and a model are loaded, buttons *Predict* in areas (7) and (8) are active. By clicking one of this buttons, a prediction for the loaded sample is made, correspondingly to the loaded model. The prediction result is displayed in the corresponding area, as shown below:

Figure 7.16



Prediction Result: 50.4237 %

senario_2_PCR.mat

E_18.ASC Loaded

The prediction result for the loaded sample, using the loaded prediction model

8. BIBLIOGRAPHY

1. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. M. R. Maleki, A. M. Mouazen, H. Ramon, J. De. Baerdemaeker. Biosystems Engineering (2007) 96(3), 427-433. Elsevier Ltd
2. Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties. Xin Bao, Liankui Dai. Fuel 88 (2009) 1216-1222. Elsevier Ltd
3. Rapid identification of petroleum products by near-infrared spectroscopy. Hoeil Chung, Hyuk-Jim Choi, Min-Sik Ku. Bulll. Korean Chem. Soc 1999, Vol. 20, No 9
4. Chemometrics: a textbook. D.L. Massart, B.G.M. Vandeginste, S.M. Deming, Y. Michotte, L. Kaufman. Elsevier Ltd
5. Multivariate calibration. What is in chemometrics for the analytical chemist. Rasmus Bro. . Analytica Chimica Acta 500 (2003) 185-194. Elsevier Ltd
6. Standardization of near-infrared spectrometric instruments. E. Bouveresse, C. Hartmann, D. L. Massart, I. R. last, K. A. Prebble. Anal. Chem. 1996, 68, 982-990
7. Infrared spectroscopy: fundamentals and applications. Barbara H. Stuart. John Wiley & Sons, Ltd
8. Monochromator. Retrieved from Wikipedia.org : http://en.wikipedia.org/wiki/Monochromator
9. Near-infrared spectroscopy. Retrieved from Wikipedia.org : http://en.wikipedia.org/wiki/Near-infrared_spectroscopy
10. The Beer-Lambert Law. Jim Clark 2007. Retrieved from : http://www.chemguide.co.uk/analysis/uvvisible/beerlambert.html
11. Beer-Lambert Law. Retrieved from Wikipedia.org : http://en.wikipedia.org/wiki/Beer%27s_law
12. Mineral fuels. N. Pasadakis. Technical University of Crete, Chania 2009
13. Aromatic Hydrocarbon. Retrieved from Wikipedia.org : http://en.wikipedia.org/wiki/Aromatic_hydrocarbon
14. The chemistry and technology of petroleum Second edition, revised and expanded. James G. Speight.ISBN: 0-8247-8481-2
15. Near-infrared Technology in the Agricultural and Food Industries. Williams, P., Norris, K. Am. Assoc. Cereal Chem., St. Paul, MN , 1987.
16. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock is soils – Critical review and research perspectives. Veronique Bellon-Maurel, Alex McBratney. Soil Biology and Biochemistry vol.43, Issue 7, July 2011, 1398-1410. Elsevier Ltd
17. Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene. Hoeil Chung, Min-Sik Ku, Joon-Sik Lee. Vibrational Spectroscopy vol.20, issue 2, August 1999, 155-163. Elsevier Ltd
18. Partial least squares methods for spectral analysis. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. D. M. Haaland, E. V Thomas, Analytical Chemistry 60 (1988):1193-1202
19. A tutorial on Near Infrared spectroscopy and its calibration. Lidia Esteve Agelet, Charles R. Hurburgh, Jr., Critical Reviews in Analytical Chemistry. 40:246-260, 2010
20. Real-time classification of petroleum products using near-infrared spectra, Minjin Kim, Young-Hak Lee, Chonghum Han. Computers and Chemical Engineering 24 (2000) 513-517. Elsevier Ltd
21. Quality control decisions with near infrared data, M.S Sanchez, E. Mertran, L.A. Sarabia, M.C. Prtiz, M. Blanco, J. Coello. Chemometrics and Intelligent Laboratory Systems 53 (2000) 69-80. Elsevier Ltd

22. Suppressing the temperature effect in near infrared spectroscopy by using orthogonal signal correction. Marcel Blanco, Damarih Valdes. J. Near Infrared Spectroscopy 14, 155-160 (2006).NIR Publications
23. Determination of physic-chemical parameters for bitumens using near infrared spectroscopy, M. Blanco, S. Maspoch, I. Villarroya, X. Peralta, J.M. Gonzalez, J. Torres. Analytica Chimica ACta 434 (2001) 133-141. Elsevier Ltd
24. Monitoring of a batch organic synthesis by near-infrared spectroscopy: modeling and interpretation of three-way data. Paul Geladi, Jennie Forsstrom. Journal of chemometrics 2002, 16: 329-338. John Wiley & Sons, Ltd
25. Partial least squares regression: a tutorial. Paul Geladi, Bruce R. Kowalski. Analytica Chimica Acta, 185 (1986) 1-17. Elsevier Ltd
26. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods in biodiesel data. Roman M. Baladin, Sergey V. Smirnov. Analytica Chimica Acta 692 (2011) 63-72. Elsevier Ltd
27. Evaluation of principal component selection methods to form a global prediction model by principal component regression. Yu-Long Xie, John H. Kalivas. Analytica Chimica Acta 348 (1997) 19-27. Elsevier Ltd
28. Applicability of high-absorbance MIR spectroscopy in industrial quality control of reformed gasolines. Jose M. Andrade, Maria S. Sanchez, Luis A. Sarabia. Chemometrics and Intelligent Laboratory Systems vol.46,, Issue 1, 15 February 1999, 41-55. Elsevier Ltd
29. Handbook of petroleum product analysis. James G. Speight. Wiley Interscience
30. ASTM International. Retrieved from Wikipedia.org: http://en.wikipedia.org/wiki/ASTM_International
31. A tutorial on principal component analysis. Jonathon Shlens. Systems Neurobiology Laboratory, University of California at San Diego. 10/12/2005 Version 2.
32. Regression analysis. Retrieved from Wikipedia.org : http://en.wikipedia.org/wiki/Regression_analysis
33. Multivariate Data Analysis in practice. Kim H. Esbensen, HiT/TF. Camo 4th edition
34. Variables selection methods in near-infrared spectroscopy. Zou Xiaobo, Zhao Jiewen, Malcolm J.W. Povey, Mei Holmes, Mao Hanpin. Analytica chimica acta 667 (2010) 14-32. Elsevier Ltd
35. Linear regression. . Retrieved from Wikipedia.org : http://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_regression
36. A correlation principal component regression analysis of NIR data. Jianguo Sun. Journal of chemometrics, vol. 9,21-29 (1995). John Wiley & Sons, Ltd
37. Different kinds of PLS weights, loadings, and what to look at?. Barry and the eigenvector staff. Retrieved from eigenvector.com: http://www.eigenvector.com/evriblog/?p=86
38. SIMPLS: an alternative approach to partial least squares regression. Sijmen de Jong. Chemometrics and intelligent lagoratory systems, 18 (1993) 251-263. Elsevier Ltd
39. Centre of Advanced Data Analysis. Retrieved from predict.ws/: http://www.predict.ws/H_principle/SvanteHarald.htm
40. Overfitting. Retrieved from Wikipedia.org: http://en.wikipedia.org/wiki/Overfitting
41. Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines. Roman M. Balabin, Ravilya Z. Safieva, Ekaterina I. Lomakina. Microchemical Journal vol 98 (2011), 121-128. Elsevier Ltd.