Monitoring the Esterification Reactions Of Carboxylic Acids With Alcohols Using Near–Infrared Spectroscopy And Multivariate Calibration Methods

By

Betül Öztürk

A Dissertation Submitted to the
Graduate School in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF CHEMISTRY

Department: Chemistry
Major: Chemistry

İzmir Institute of Technology
İzmir, Turkey

September 2003

ACKNOWLEDGEMENTS

I wish to sincere thanks to my research advisor Dr. Durmuş ÖZDEMIR, for his support and guidance in the completion of my master research program. His encouragement and patience are greatly appreciated. I would also thank my committee members, Dr. Figen KOSEBALABAN TOKATLI and Dr. Ahmet Emin EROĞLU, for their help in editing and refining this manuscript.

I would also like to acknowledge those who provided financial support for my research Izmir Institute of Technology.

Finally I would like to thank my family and my friends for their love, patience, and support.

ABSTRACT

Simultaneous determination of mixtures of alcohols, acids, esters, and water using near-infrared spectroscopy (NIR) and four different multivariate calibration methods were realized. The four multivariate calibration methods were Genetic Inverse Least Squares (GILS), Genetic Regression (GR), Principle Component Regression (PCR), and Partial Least Squares (PLS). Four different esterification reactions were investigated. These are methyl acetate, ethyl acetate, propyl acetate, and butyl acetate. The sample set contains 40 ternary mixtures of these esterification processes. Duplicate measurements were done for each sample and 80 NIR absorbance spectra that cover the range from 4000 to 10000 cm^{-1} were obtained. Of these 80 spectra, 50 were used as calibration set, 30 were reserved for the prediction purposes. Several calibration methods were built for each component of each esterification reactions. Standard error of calibration (SEC) and standard error of prediction (SEP) were calculated for each calibration model and comparison of these four multivariate calibration methods was done.

ÖZ

Yakın-Enfraruj Spektroskopisi (NIR) ve dört farkli çok değişkenli kalibrasyon metotları kullanılarak alkol, asit, ester ve su karışımlarının eş zamanlı tayinleri yapılmıştır. Bunlar, Genetik Ters En küçük Kareler Yöntemi (GILS), Genetik Regresyon (GR), Temel Bileşenler Regresyonu (PCR) ve Kısmi En Küçük Kareler Yöntemi (PLS) kullanılan dört çok değişkenli kalibrasyon metodudur. Dört değişik esterleşme reaksiyonu bu amaçla incelenmiştir. Bu esterleşme reaksiyonları metilasetat, etil asetat, propilasetat ve bütilasetat oluşum reaksiyonlarıdır. Örnek seti 40 adet dörtlü karışım içermektedir. Her bir örnek iki kez ölçülmüş ve 4000 – 10000 cm^{-1} aralığında 80 adet yakın-enfraruj absorbans spektrumları elde edilmiştir. Bu 80 spektrumdan 50 tanesi kalibrasyon seti için kullanılmış; 30 tanesi test amacı ile ayrılmıştır. Her bir esterleşme reaksiyonundaki her bir bileşen için bir çok kalibrasyon modeli oluşturulmuştur. Her bir model için standard kalibrasyon hatası (SEC) ve standard tahmin hatası (SEP) hesaplanmış ve kullanılan dört çok değişkenli kalibrasyon metodu karşılaştırılmışıtır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The analytical chemists need measuring in order to quantify different matter in our daily life. In chemical analysis it is often difficult to find ideal measurements. The challenge today is how to detect smaller and smaller analyte signals and separate the net signals from background and noises. Due to the phenomena in the samples themselves, the data may be affected by the chemical or physical interferences. Traditional analytical method contains a chain in which samples are brought to the laboratory for separation, extraction, preparation and finally determination [1] (Figure 1.1).

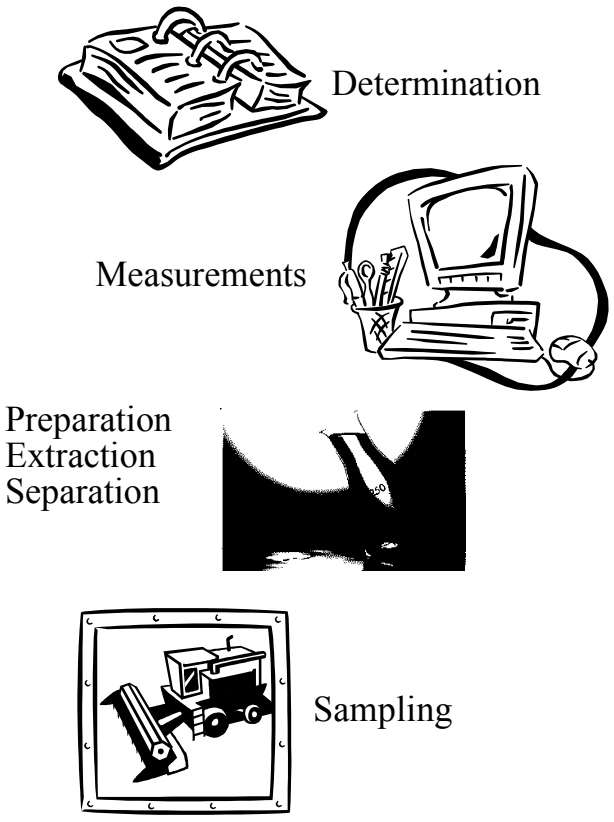

Figure 1.1.The analytical chain.

While this chain is still employed, the modern era of automation and computer power has resulted in tremendous improvements. This development has given new phenomena to process monitoring by using fiber optic sensors etc… On-line, at-line, in-line monitoring… In the last years, industrial analytical chemists have mostly used on-line monitoring measurements.

These improvements are helping the analytical chemist in various steps of this chain. Minimizing the sample preparation, and the determination time are the few examples of these improvements. Because during the process we do not have any time for cleaning or standardizing our samples in the laboratory

Monitoring usually contains long analysis terms and correlations. Working with process monitoring, enormous data are obtained from the measurement results. In practice an alternative approach is often needed to use and interpret all the information stored in our database. One of them is the use of models based on statistical principles. The data processing and modeling can be quickly done at the same time using these principles by the help of modern powerful computers. Despite this quick analysis of the measurement results, there are some limitations in the model building. In model construction in order to find the best model requires a lot of trials that cause cost and time. In addition, the error approach is necessary and very important to analytical chemist to find out the best model. In spite of those disadvantages models are extremely valuable and gives enormous information about figures of merit for the user. An attempt to explain the value of establishing mathematical relationships is given by the Figure 1.2. A model calculating the slopes of the lines between the black and white stripes would quickly reveal that the lines really are parallel or not, only by visual inspection. *Optimally a model can be used for objective numerical interpretations and precise predictions over time, with the only requirement of fast inexpensive, and non-hazardous measurements as model input data.*

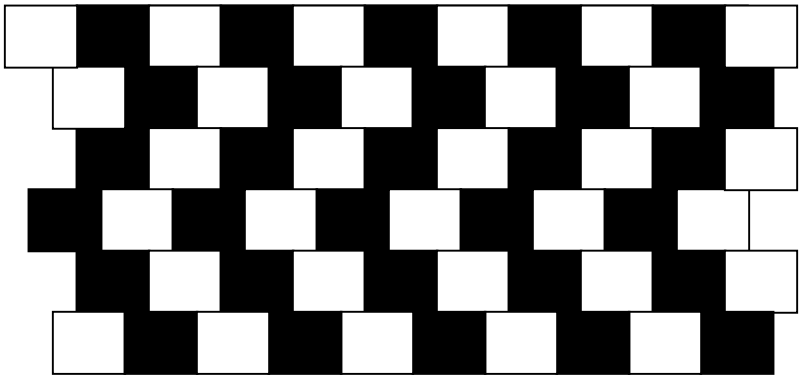

Figure 1.2.Are the horizontal lines are parallel or do they slope?

Nowadays, for these and many other reasons modern monitoring schemes that are based on models are used. Especially Near-Infrared Spectrometers are the most widely used instruments for these analyzing systems.

Aim of the study:

The aim of this project is to investigate the possibility of analyzing the complex mixtures of carboxylic acids, alcohols, esters and water by using near infrared spectroscopy and multivariate calibration methods (GR, GILS, PCR, and PLS) and also to investigate the possibility of developing calibration models for the solutions of these compounds. Due to this reason, the performance of these calibration methods in developing these models were investigated and they were compared with each other. Later on, it will be searched if it were possible to use the obtained models to measure these compounds in their real process (ex; in industrial process)

CHAPTER 2
NEAR-INFRARED SPECTROSCOPY

2.1 DESCRIPTION OF THE INFRARED REGION

In the electromagnetic spectrum, infrared region has a wide wavelength range compared to the other regions. It starts from around 780 nm and ends up to the 1 x 10^6 nm. Because of that, this region is divided into three groups according to the requirements that change from instruments and the applications. Table 1.1 shows the wavelength range of these different infrared regions. [1]

Table 2.1. The corresponding wavelengths of the IR regions

Name of the region	**Wavelength range (nm)**	**Wavelength number (cm^{-1})**
Near–Infrared (NIR)	780–2500	12.800-4000
Mid–Infrared (MIR)	2500–50.000	4000-200
Far–Infrared (FIR)	50.000–1 x 10^6	200–10

Generally mid–infrared (MIR) is used for both qualitative and quantitative analysis in analytical chemistry. Especially organic molecules are widely detected in this region; since each functional group in the molecule has unique information in this region, which is called fingerprint. Also each molecular group has sharp absorption bands. Organometallic or inorganic molecules are qualitatively analyzed in the far–infrared (FIR) region due to the metallic band in these compounds. However near–infrared (NIR) spectrometry is used for quantitative analysis of complex mixtures by the help of computers and mathematical methods that are based on statistics.

In infrared spectroscopy the constituents or the molecules can be detected as: Light from a spectrometer is directed to strike a complex mixture consisting of one or more types of molecules. Molecules absorb the radiation of the energy to be excited it to the vibrational or rotational states. During this motion, a change in the dipole moment is required for the selection of the absorption bands. Therefore diatomic molecules such as H_2, N_2, and O_2 cannot absorb the IR radiation. Table 2.2 shows the molecular interactions associated with the infrared regions. [2]

Table 2.2. The molecular interactions associated with the infrared regions.

Name of region	Characteristics Measured
NIR	Overtone & combination bands of fundamental molecular vibrations
MIR	Fundamental molecular vibrations
FIR	Molecular rotations

2.2 THEORY OF NEAR-INFRARED SPECTROSCOPY

The near-infrared region composes around 780 nm to 2500 nm. In this region, as shown above, the absorption bands are due to the overtones (780 to 1800 nm) and combination (1800 to 2500 nm) bands of fundamental mid-IR molecular vibrations. [2-5] The energy transition occurs between the ground state and the second or third excited vibrational states.

Harmonic oscillator can explain the vibrations in infrared spectroscopy. A disturbance of one atom along the axis of the spring results in a vibration called a simple harmonic motion. In Figure 2.1, the vibration of a single mass attached to a spring that is hung from an immovable object is shown.

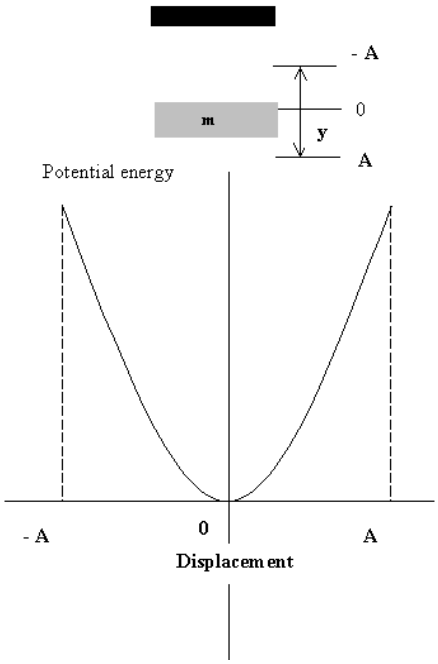

Figure 2.1. Simple harmonic motion.

Using Hooke's Law, the restoring force F is proportional to the displacement, which is shown by y.

$$F = -ky \tag{2.1}$$

where k is a force constant. It depends on the bond order.

By the help of the second laws of Newton the natural frequency for the diatomic molecule as:

$$\nu = {}^{1}\!\diagup\!_{2\pi}\sqrt{\frac{k}{\mu}} \tag{2.2}$$

where k is a force constant again and μ is a reduced mass shown by:

$$\mu = \frac{m_1 \cdot m_2}{m_1 + m_2} \tag{2.3}$$

where m_1, m_2 are the masses of atom that are in a diatomic molecule involved in a vibration.

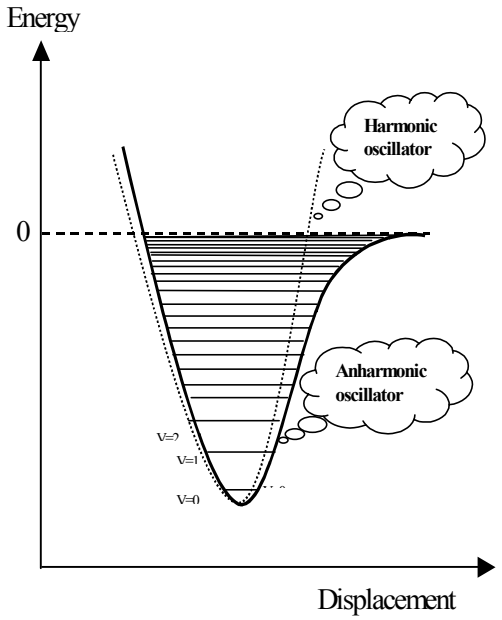

Figure 2.2: Energy diagram of vibrational modes.

If the diatomic molecule shows an ideal harmonic oscillator the potential energy is,

$$E = \frac{1}{2} k y^2 \qquad 2.4$$

The parabola in Figure 2.2 shows that the potential energy is a maximum when the spring is stretched or compressed to its maximum aptitude A and decreases to zero at the equilibrium position. However, the equations above do not completely describe the behavior of the system; since the quantized nature of molecular vibration energies does not appear. Rearranging and developing these equations give the potential energy as:

$$E = \left(v + \frac{1}{2} \right) \frac{h}{2\pi} \sqrt{\frac{k}{\mu}} \qquad 2.5$$

where *h* is Planck constant and v is the vibrational quantum number, which can take only positive integer values including zero.

By rearrangement of equation 2.2 and 2.5:

$$E = \left(v + \frac{1}{2} \right) h\nu \qquad 2.6$$

is obtained. Thus only discrete energy levels would be allowed. The separation between levels would be:

$$\Delta E = h\nu \qquad 2.7$$

According to the quantum mechanical selection rules, transitions between more than one vibrational state are forbidden for such an ideal harmonic oscillator. Therefore only the fundamental vibrations occur and there would be no near infared spectrum.

However, molecules are not ideal oscillators and vibrations in real molecules. They show anharmonic oscillator. At higher vibrational states, departures from harmonic behaviour occur. When a molecule contains a high level of potential energy, it has a tendency to dissociate and can no longer return to its equilibrium position by a restoring force. (Figure 2.2). There are two types of anharmonicity that affect the vibrations, namely mechanical anharmonicity, and electrical anharmonicity. [4] Mechnical anhrmonicity is observed, when the potential energy is as a function of displacement has terms third or higher order.Electrical anharmonicity is observed if the dipole moment of the vibration is nonlinearly related to the displacement of the vibration.

The energy values of the anharmonic oscillator are provided by the following equation:

$$E = h\nu\left(v + \frac{1}{2} \right) - xh\nu\left(v + \frac{1}{2} \right)^2 + yh\nu\left(v + \frac{1}{2} \right)^3 + ... \qquad 2.8$$

where *x* and *y* are the anharmonicity constants. The more important consequence of anharmonicity is that it allowes transitions between more than one vibrational state and leads the absorption of NIR radiation by molecules.

Combination bands in NIR spectrum consists of strecthing and bending combinations. Strecthing vibrations occur at higher energy levels and they are either symmetric or asymmetric. Bending vibrations occur at high wavelengths and they are in-

plane or out-of-plane. In-plane bending consists of scissoring and rocking, out-of-plane bending consists of twisting and wagging. From the higher wavelengths to the lowest wavelengths strecthing, in-plane bending (scissoring), out-of-plane bending (wagging); twisting and rocking occur

For a given molecule many overtones and combination bands occur in NIR region.Therefore NIR spectra are very complex because of the overlapping bands. In qualitative analysis it is less useful than the mid-IR region. However in quantative analysis it is more conventional by the help of multivariate calibration methods.

2.3 INSTRUMENTATION OF NEAR-INFRARED SPECTROSCOPY

A wide variety of NIR instruments have been designed and great improvements have been made in the last 50 years. There has been a reduction in instrument noise, and along with noise reduction improvements have been made in accuracy and sensitivity. Simple filter spectrometers to high speed Fourier transform spectrometers have been manufactured for many applications. According to their properties, NIR instruments can be categorized into two classes, dispersive instruments, and Fourier-Transform instruments. [6]

2.3.1 Dispersive Instruments:

UV–visible instruments are generally designed to cover the wavelength range from 780 to 2500 nm and they are called UV-visible-NIR spectrophotometers; since UV-visible and NIR infrared spectrometers show similarities in the optical elements. Most NIR instruments have detector both reflectance and transmittance measurements together. Therefore it allows the analysis of both solid and liquid samples easily. Dispersive instruments depend on the monochromator type. (Figure 2.3) Holographic gratings have replaced the old mechanically grooved and replicated gratings, since the holographic gratings have fewer detect, anomalies and greater throughput across the entire NIR spectrum. They are produced by photoetching process, and manufactured easily because they are economical. [6]

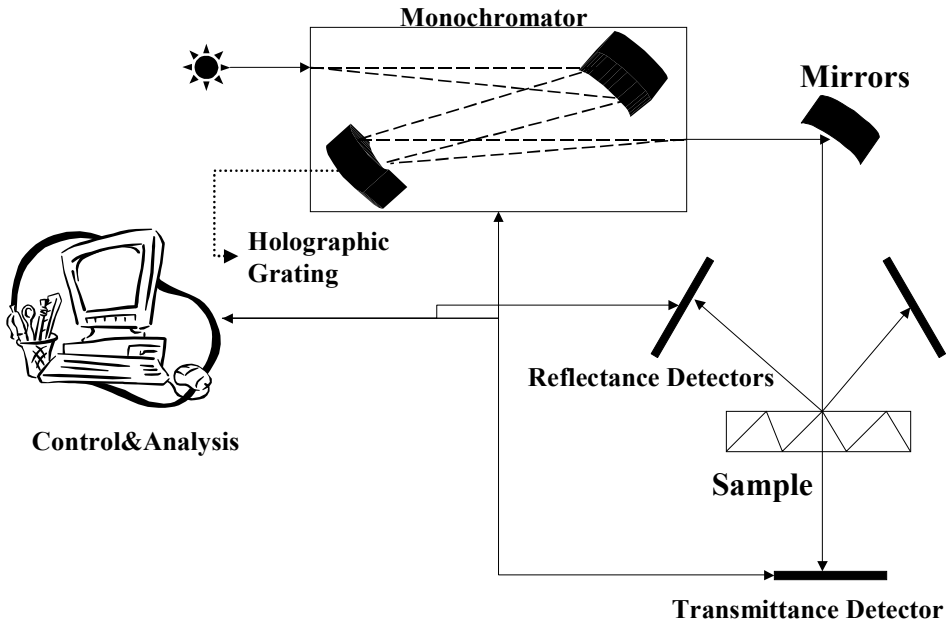

Figure 2.3: Schematic representation of a NIR spectrometer.

Tungsten-halogen lamps with quartz envelopes are usually used as source for the NIR instruments. These lamps provide high-energy output over the 360–3000 nm region. In addition they can be used for a long time; because the halide inside the lamp envelope performs a bathing action that keeps the quartz window cleaner than it is possible with ordinarily evacuated tungsten lamps. More expensive xenon discharge lamps can also be used as a source. In recent years lasers and light-emitting diodes have been started to use. [7]

Sample handling is relatively easier in the NIR region than in the mid-IR range. Sample cells are made from quartz or fused silica. Their pathlengths vary between 1mm to 10 cm.

Detectors are designed for two aims. First one is for reflectance measurements and the second one is for transmittance measurements. For this purposes silicon sensors, lead sulfide (PbS), antimony sulfide (Sb_2S_3), lead selenide (PbSe), and bismuth sulfide (Bi_2S_3) have been used. In some cases gallium arsenide (GaAs) detectors are used, but small surface areas of such detectors do not lend themselves to integration of diffusely reflected light. [6]

2.3.2 Fourier Transforms Instruments

Fourier Transform near-infrared (FT-NIR) instruments have been commercially available for several years. Michelson interferometer, Fabry–Perot based interferometer or common type of interferometer-based instruments are the most widely used types in the instrument design. [6] The Michelson interferometer-based instruments are usually called multiplex spectrometers in which all wavelengths of the radiation from source are observed simultaneously. Therefore this type instruments has a very high speed and signal-to-noise ratio (S/N). The Michelson interferometer contains a beam splitter and two plane mirrors, one fixed and one movable mirror. After a beam of polychromatic radiation is passed through a sample cell, it is split into two equal power beams, one of them directed to the fixed mirror and the other is to the movable mirror. After reflection from the plane mirrors, the beams are recombined and sent to the detector. And the variations

in the intensities of the combined beams can be measured as a function of pathlength differences.

The Fourier transform is applied for signal processing in multiplex instruments to convert time domain spectra into the frequency domain. Figure 2.4 shows an artificial interferogram and its wavelength domain spectrum. Therefore they are called Fourier transform instrument. This property gives some advantages to this instrument: The first advantage is, the large light flux reaching to the detector since there are no entrance slits. Using interferometer leads wavelength accuracy and precision that are providing high signal to noise ratio (S/N). And the last one is, all the wavelengths of the radiation from the source are recorded simultaneously. Thus an entire spectrum can be recorded in a second. This extremely high speed can be also used for signal averaging, which also provides a better S/N ratio.

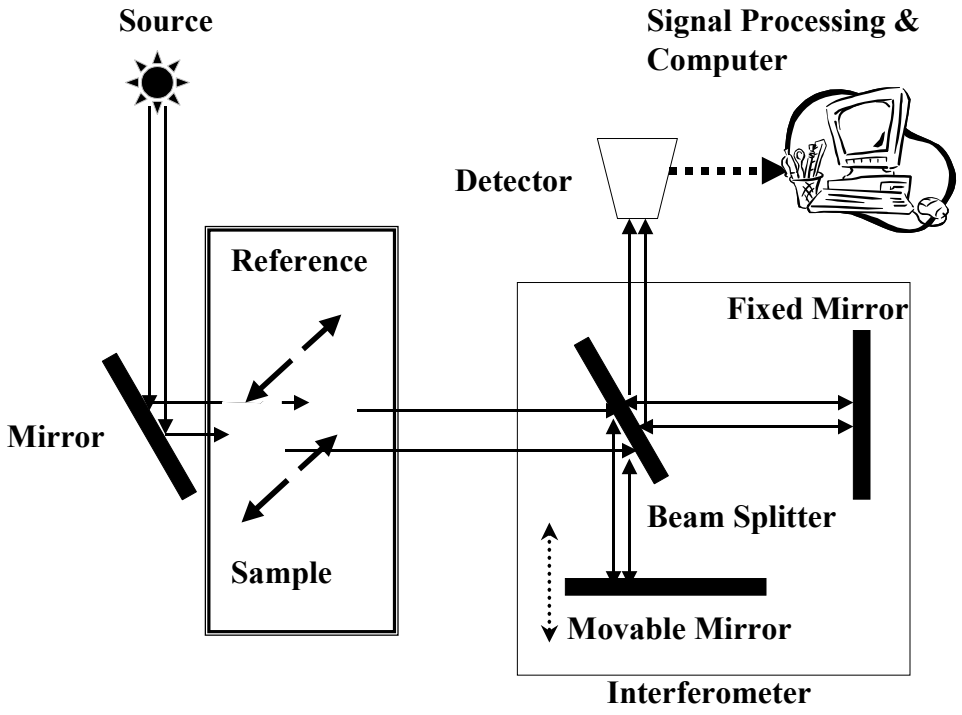

Figure 2.4: Simplified block diagram of a Michelson Interferometer-based Fourier Transform Near–Infrared Spectrometer

2.4 ADVANTAGES OF NEAR-INFRARED SPECTROSCOPY

Near–Infrared spectroscopy has numerous advantages over the traditional infrared analysis methods.[7]

All organic compounds absorb in NIR region, but not strongly. It gives good sample penetration for the NIR light and makes Beer's Law useful for NIR region. Longer pathlenghts can be used for the liquid analysis. Wide slits monochromators, efficient detectors, bright sources and high throughput optics provide high S/N ratio that gives effectiveness for the NIR spectrum. The reproducibility of intensity measurements is on the order of micro absorbance units; since the intensity of NIR bands is less than the mid-IR bands. It is very rapid analysis, since it does not need necessarily sample preparation. Because of the glass optics, grating spectrometers are relatively cheap among the other IR instruments. The sensitivity of NIR instruments is increasing and the dedection limit is decreasing, as little as 0.001 to 1% when ideal conditions exist. One can determine as many components in the samples as there are spectrally independent wavelengths. This number is commonly around ten.

CHAPTER 3
MULTIVARIATE CALIBRATION METHODS IN SPECTROSCOPY

3.1 INTRODUCTION

Spectroscopy is a branch of analytical chemistry that gives both quantitative and qualitative information about a sample of interest. Since spectroscopic analysis is based on indirect measurements, it requires a calibration process. The word “Calibrate” in daily life means to determine the inner diameter or capacity of a gun or some other cylinder. However in quantitative analysis the word “CALIBRATE” is used for empirical data and prior knowledge for determining how to predict unknown quantitative information from available measurements with some mathematical transfer function [8]. In this context, calibration is one of the key steps associated with the many biological, industrial and environmental materials.

In the past, data acquisition and analysis were often time-consuming and tedious activities in analytical laboratories. Advances in instrumentation and computing have allowed analytical chemists to collect huge amounts of data on a wide variety of problems of interest and to construct a calibration model for instrument. However more data do not necessarily mean more information. Only when the data are interpreted and put to use, they become valuable to the chemist and to society in general, then the data become information. For this reason, the data analysis or methods helped to broaden to use of analytical techniques for difficult problems.

In the simplest conditions, a model such as

$$y = ax + b \qquad 3.1$$

has been used to express the relationship between a single measurement (y) from an instrument and the level (x) of the analyte of interest. Instrumental measurements are obtained from species in which the amount of the analyte has been determined by some independent and inherently accurate results assay. [9, 10] Instrumental measurements and results from the independent assays are used to construct a model that relates the analyte

level to the instrumental measurements. Then, this model is used to predict the analyte levels of future samples based on the instrumental measurements.

In general calibration models are divided into two types:

1. Univariate Calibration
2. Multivariate Calibration

Multivariate analysis is a collection of powerful mathematical and statistical tools that can be applied to chemical analysis when more than one measurement is acquired for each sample. To understand the evolution of multivariate calibration methods; it is useful to review the univariate calibration methods and its limitations.

3.2 UNIVARIATE CALIBRATION METHODS

In general univariate calibration methods involve the use of a single measurement from an instrument. It also requires the signal from instrument only depends on the analyte of interest without interference. [9]

The calibration in these methods is emphasized into two steps: (Figure 3.1)

1. Calibration
2. Prediction

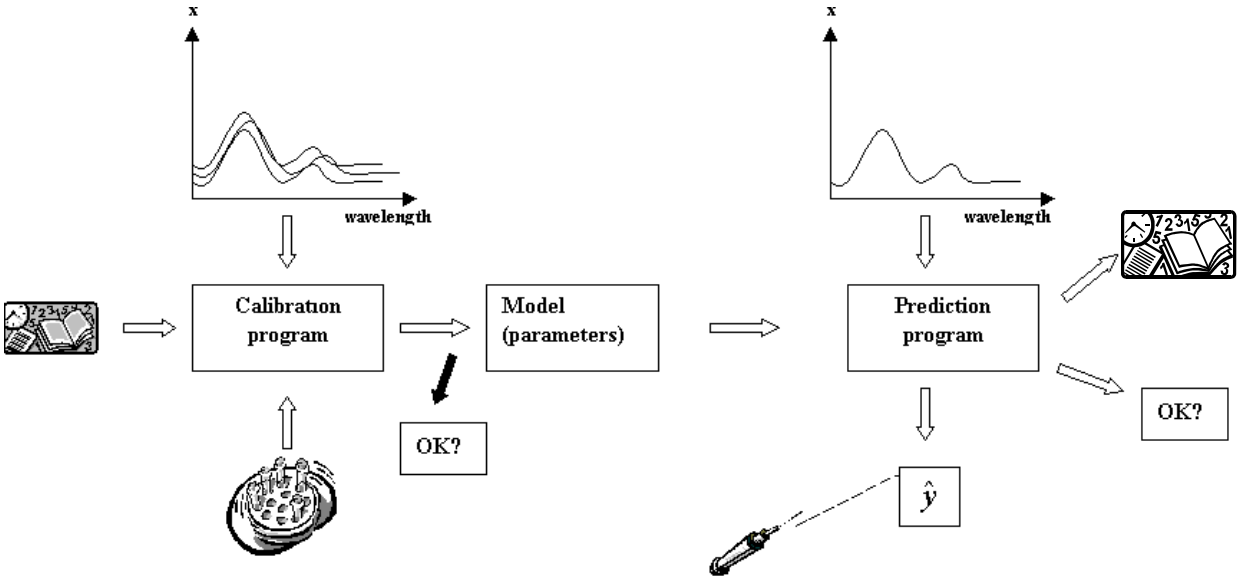

Figure 3.1: A schematic diagram of the calibration and prediction process.

In the first step (calibration), indirect instrumental measurements are obtained from the species whose concentrations have been determined by another independent method. The set of instrumental measurements and results from the analyte referred the calibration or training set. These are used to construct a model that relates the amount of analyte to the instrumental measurement. In general, this step is the most time-consuming and expensive part of the overall calibration procedure because of the preparing of reference samples.

Next the model developed in the calibration step combined with the measurement of a new species to predict the analyte level or amount. The prediction step is illustrated in Figure 3.2.

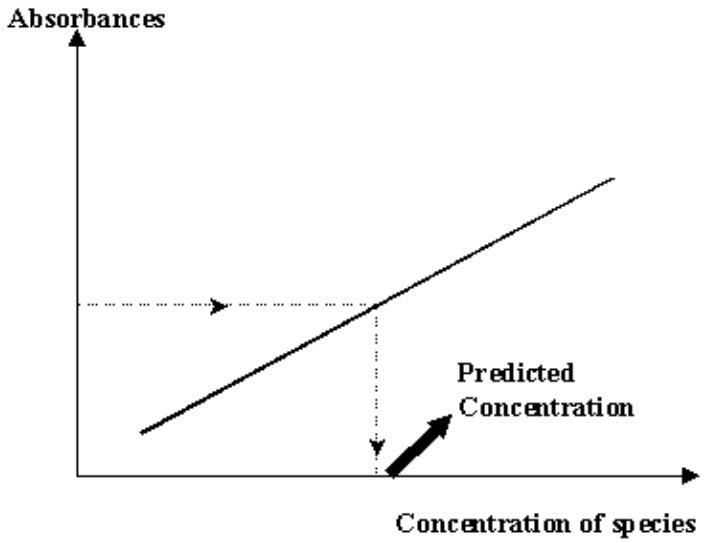

Figure 3.2: A calibration graph generated by simple linear least squares for a set of sample.

For a univariate calibration model, the response of the instrument, *s*, at a fixed frequency is related to the analyte concentration**,** ***c***, by the calibration function defined as:

$$s = f(c) + e_s \qquad 3.2$$

where e_s is the error associated with the instrument response. In spectroscopy, the relationship described by *f(c)* is assumed to be linear according to the Beer's law.

The two most common method of modeling are the classical and inverse methods when the assumption for the linear relationship between instrument response and analyte concentration holds true. [9]

3.2.1 Classical Univariate Calibration

In the classical univariate calibration method, the expressed statistical model can be shown as:

$$a_i = b_0 + b_1 c_i + e_i \qquad 3.3$$

where a_i and c_i are the instrument response and analyte concentration, respectively for the ith sample of m calibration standards. The measurement error associated with a_i is represented by e_i. In ideal case there is no error associated with the instrumental response. Therefore the solution of the Equation 3.3 simply gives a straight line. That is b_0 and b_1 are the intercept and slop of the line. However; there is no such ideal case in real world samples or applications and there is always some sort of error associated with instrumental response or in the concentration values. Therefore, the traditional quantitative spectroscopic analysis methods begin with plotting the instrument responses against the analyte concentration for a set of m calibration standards. Then, b_0 and b_1, which produces a straight line that is best fit to the plotted data, are estimated using methods of least squares. The method of least squares minimizes the sum of the squares (SS) of the residuals for all the points in the classical univariate calibration method. The SS of the data set can be defined by rearranging Equation 3.3 for m calibration standards as:

$$SS = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} \left(a_i - b_0 - b_1 \times c_i\right)^2 \qquad 3.4$$

To minimize SS, partial derivatives of SS needs to be taken with respect to each of the two parameters that are being estimated and resulting expressions are set to zero. [11, 12]

Thus:

$$\frac{\partial SS}{\partial b_o} = 2\sum_{i=1}^{m}\left(a_i - b_0 - b_1 \times c_i\right)(-1) = 0 \qquad 3.5$$

and

$$\frac{\partial \boldsymbol{SS}}{\partial \boldsymbol{b}_1} = \boldsymbol{2}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\left(\boldsymbol{a}_i - \boldsymbol{b}_0 - \boldsymbol{b}_1 \times \boldsymbol{c}_i\right)\cdot\left(-\boldsymbol{c}_i\right) = \boldsymbol{0} \qquad 3.6$$

After dropping 2 and –1 from Equations 3.5 and 3.6, the solutions of b_0 and b_1 can be obtained by solving normal equations:

$$\sum_{i=1}^{m}\left(a_i - b_o - b_1 \times c_i\right) = 0 \qquad 3.7$$

and

$$\sum_{i=1}^{m}\left(c_i a_i - c_i b_0 - b_1 c_i^2\right) = 0 \qquad 3.8$$

These equations then become;

$$mb_0 + b_1\sum_{i=1}^{m} c_i = \sum_{i=1}^{m} a_i \qquad 3.9$$

and

$$b_0\sum_{i=1}^{m} c_i + b_1\sum_{i=1}^{m} c_i^2 = \sum_{i=1}^{m} c_i a_i \qquad 3.10$$

The solution of these equation produces the least squares estimated values of b_0 and b_1:

$$\hat{\boldsymbol{b}}_1 = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{c}_i \boldsymbol{a}_i - \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{c}_i\right)\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{a}_i\right) \Big/ \boldsymbol{m}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{c}_i^2 - \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{c}_i\right)^2 \Big/ \boldsymbol{m}} \qquad 3.11$$

and

$$\hat{b}_o = \bar{a} - \hat{b}_1 \bar{c} \qquad 3.12$$

where $\bar{a}$ and $\bar{c}$ are the mean values of instrumental responses and analyte concentrations for *m* calibration samples. Now the estimated calibration equation can be written as:

$$\hat{a} = \hat{b}_0 + \hat{b}_1 \bar{c} \qquad 3.13$$

and then concentration of an unknown sample can be calculated by;

$$c_u = \frac{a_u - \hat{b}_o}{\hat{b}_1} \qquad 3.14$$

where c_u is unknown analyte concentration and a_u is instrument response for that sample. The correlation coefficient (R^2) is a numerical measure that express the strength of the linear relationship between c and a and can be defined as:

$$R^2 = \frac{\sum_{i=1}^{m} (\hat{a}_i - \bar{a})}{\sum_{i=1}^{m} (a_i - \bar{a})} \qquad 3.15$$

This equation produces a unit free number. The values for R^2 range from 0 to1 and it should as close as 1, possible.

3.2.1.1 Matrix Algebra Applied to Simple Linear Regression

Univariate calibration method can also be described in matrix notation. Then the equation becomes as:

$$a = C\beta + e_a \qquad 3.16$$

where ***a*** is the *mxl* vector of instrument responses, ***C*** is the *mx2* matrix of analyte concentrations, β is the *2xl* vector of regression parameters (b_0 *and* b_1) and e_a is the *mxl* matrix of the errors associated with *a* or residuals that are not fit by the model. Note that the first column of the ***C*** matrix is a vector of ones, which is necessary to estimate b_0 when the multiplication is performed.

The two normal Equations 3.5 and 3.6 given in previous section can be represented in matrix form as:

$$(C' \cdot C) \cdot \beta = C' \cdot a \qquad 3.17$$

Then least squares solution to Equation 3.17 during calibration is:

$$\hat{\beta} = (C' \cdot C)^{-1} \cdot C' \cdot a \qquad 3.18$$

where $\hat{\beta}$ is the *2x1* vector of least square estimate parameters b_0 and b_1 with the sum of squared residuals not fit by the model being minimized. Once the parameters are estimated, then the concentration of an unknown sample can be calculated by the equation 3.14.

3.2.2 Inverse Univariate Calibration

The second univariate calibration method is called as the inverse method that assumes inverse Beer's law. The modeling can be implied as:

$$c_i = p_0 + p_1 \cdot a_i + e_i \qquad 3.19$$

where e_i is the error associated with the reference value c_i. In the calibration step, the model parameters (p_0 *and* p_1) are estimated by the method of least squares described in classical univariate calibration. The following equations represent the estimated model parameters that are the slope and the intercept of the calibration line.

$$\hat{p}_i = \frac{\sum_{i=1}^{m} a_i c_i - \left(\sum_{i=1}^{m} a_i\right)\left(\sum_{i=1}^{m} c_i\right) \Big/ m}{\sum_{i=1}^{m} a_i^2 - \left(\sum_{i=1}^{m} a_i\right)^2 \Big/ m} \qquad 3.20$$

and

$$\hat{p}_0 = \bar{c} - \hat{p}_1 \cdot a \qquad 3.21$$

where $\bar{a}$ and $\bar{c}$ are the mean values of instrumental response and analyte concentrations for m calibration samples. Now the estimated calibration equation can be written as:

$$\hat{c} = \hat{p}_0 + \hat{p}_1 \cdot a \qquad 3.22$$

In the prediction step, concentration of unknown sample can be calculated by:

$$c_u = \hat{p}_o + \hat{p}_1 \cdot a_u \qquad 3.23$$

where c_u is unknown analyte concentration and a_u is instrument response for that sample.

Univariate calibration methods offer simplicity for certain types of applications where selective measurements can be found or when the analyte contains no interferences. However, their applications are limited due to the shortcomings of the single measurement based analysis. First of all, interference free systems are rarely encountered in real applications and concentrations of the interfering species are usually unknown. To make matters worse, the amounts of the interfering species in samples are not always the same. Also a large amount of noise associated with the instrument response at selected wavelength results in poor calibration. Another problem with the use of univariate calibration methods is lack of constant baseline for every measurement.

Although the predictions obtained by classical and inverse method will be different for a given example, in many cases these differences are insignificant. The choice of particular univariate calibration method depends on whether the reference values of calibration standards or the instrument responses are more precise.

3.3 MULTIVARIATE CALIBRATION METHODS

There is a need for improved quantitative information in science and technology. According to this purpose, multivariate calibration is a general selectivity and reliability enhancement tool. [8] It is applicable to determination of major constituents as well as minor components and other qualities, and for a very wide range of instrument types. It includes transformations of measurements into informative results. Calibration establishes this transformation.

The multivariate calibration means determining how to use many measured variables simultaneously for quantifying some target variable. For example, the measured variables could be chromatographic or spectroscopic measurements, and the target variable could be analyte concentration. [8]

The reasons for using multivariate calibration methods besides the traditional methods are given below:

If the constituent interacts with the other constituents in the samples or the solvents, the spectrum of the analyte in the complex samples may be different from the spectrum of the analyte in pure form. So, calibrating the analyte in isolated, purified model systems may be used a little; it will have to be done empirically on realistic samples from the actual process. But there are some problems: The analyte may not be stable and/or homogenous, and the measurements may be contaminated by the interferences. This means that there is a need of multivariate calibration methods.[8]

3.3.1 Interferences and Measurement Errors

In the real world it is very difficult to find the ideal measurements of the selectivity enhancement tool for the analyte that are interested. There will be some interferences or errors in the sampling method and analysis of sample.

a) In the Figure 3.3 there is a sample analyzed in a high-precision instrument. This measurement produces a selective measurement that is linearly related to the concentration of the analyte. There is no error and any interference. [8]

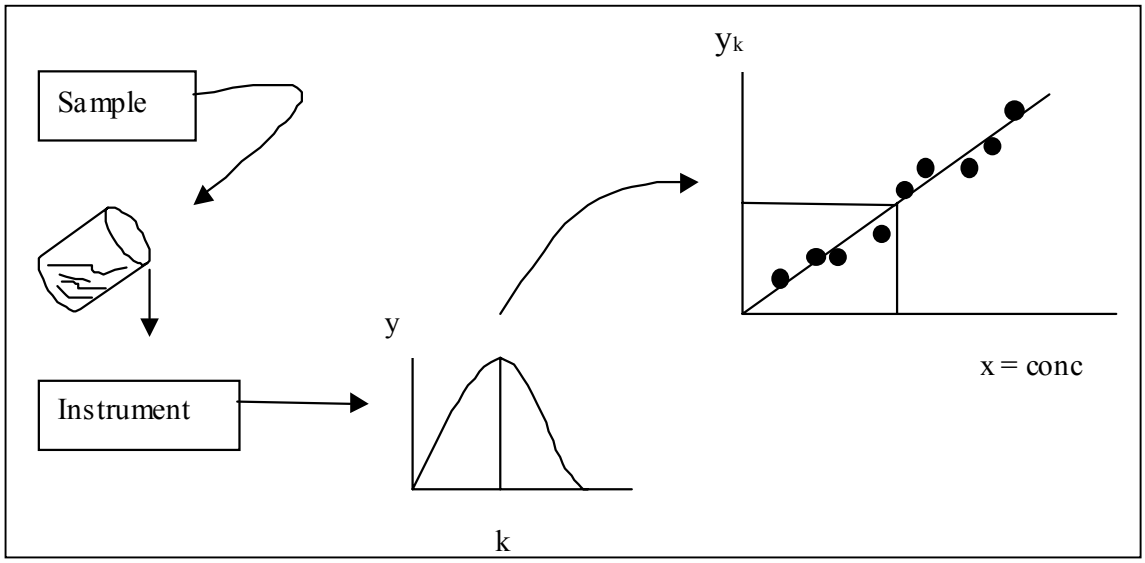

Figure 3.3. An ideal measurement.

b) In the Figure 3.4, the instrumental measurement is not selective for the analyte, and the instrument response is non-linear. There might be interferences in the sample. The nonlinearity originates from the change in the concentration of constituents and the change in the levels of the interferents. [8]

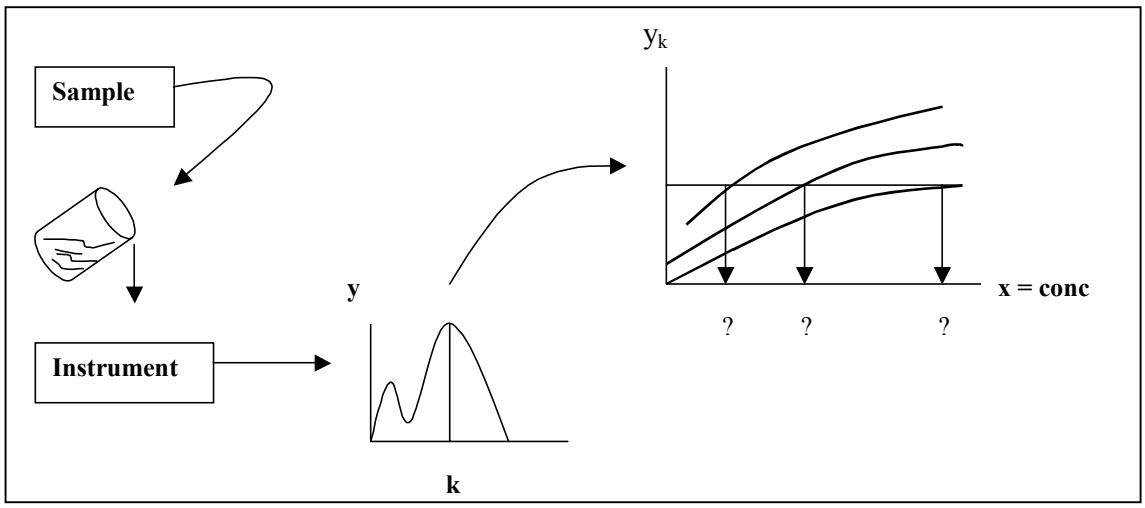

Figure 3.4. Nonlinear results for the measurements.

c) Traditionally, interferences had to be removed physically to ensure selectivity: Cleaning, standardizing and diluting each sample prior to single-channel measurement. The calibration is limited to the "linear range" and only narrow range of the instrument scales could be used. (Figure 3.5) Hopefully the data from new samples contain no unexpected trouble. [8]

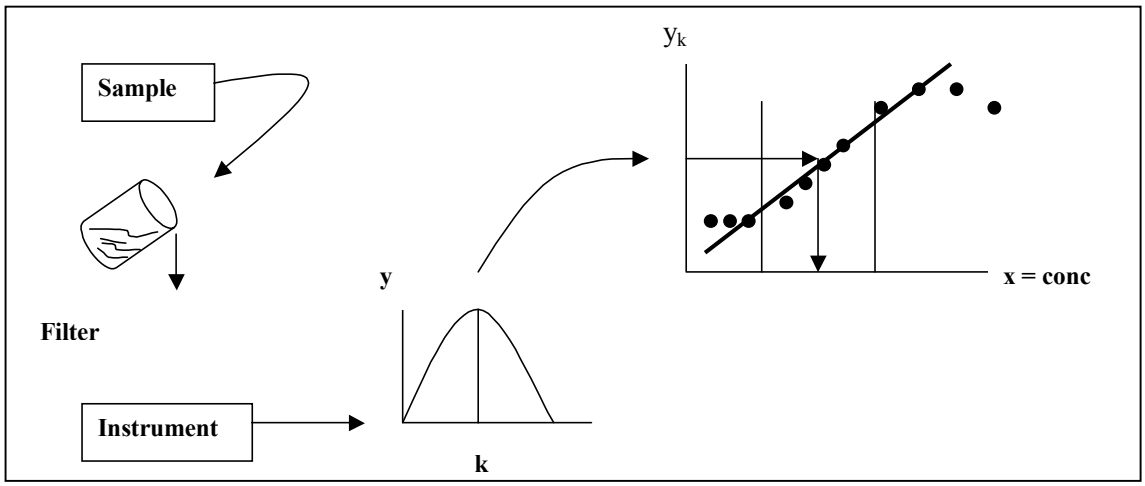

Figure 3.5. Traditional selectivity enhancement.

d) With multivariate calibration methods, interferences and individual non-linearities represent fewer problems. The cleaning, standardizing and diluting are more or less replaced by mathematical modeling of multi channel measurements. This process removes interference effects, extends the linear range of the calibration and allows automatic outlier detection if new samples contain unaccepted trouble. (Figure 3.6) [8]

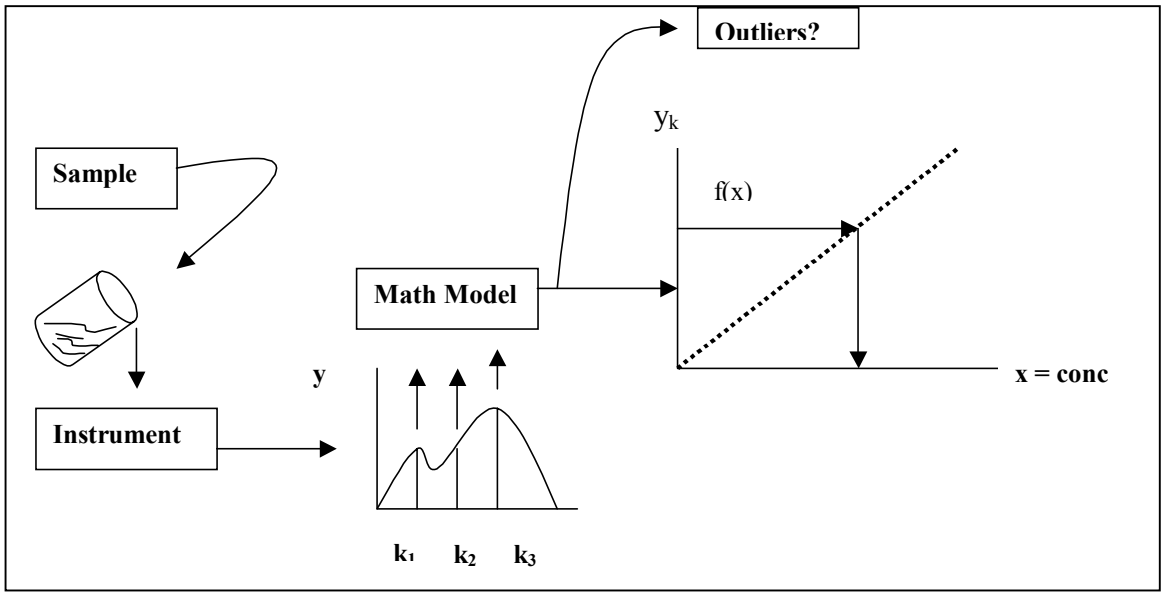

Figure 3.6.A selectivity enhancements by multivariate calibration method.

3.3.1.1 Chemical Interferences:

"Chemical interferences" is used to describe the systematic errors, which are come from the other components in the mixture or by chemically induced variations in the sample. [8] In real world, most of the samples contain more than one constituent; especially biological samples. Many constituents may affect the nature of the sample and the measurements of the sample. The main aim of the measurement is finding the relationship between the response and concentration according to the Beer law, since all the analytical chemists try to find the simple measurement way to obtain the sufficiently selective result for the analyte.

Figure 3.7.a shows the absorbance spectra of an analyte under the room conditions at various concentrations. The height of peaks depends on the concentration of the analyte. The calibration curve is obtained by plotting the absorbance value versus the concentration value. (Figure 3.7.b) For this kind of well behaved data, univariate and multivariate calibration methods give away the same prediction ability and high precise results.

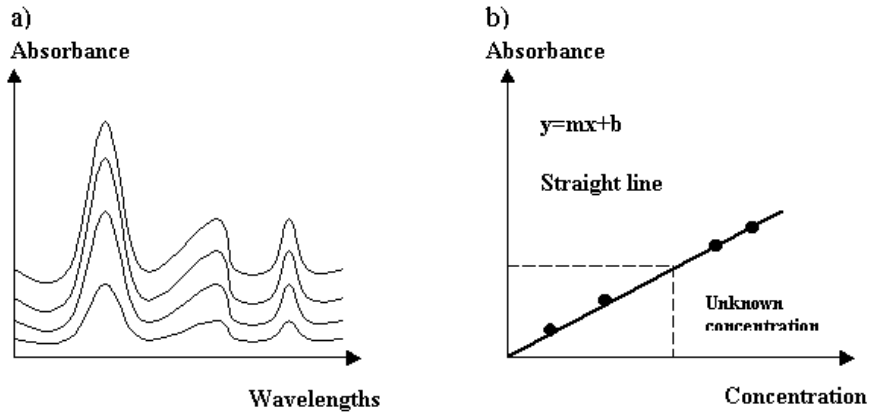

Figure 3.7. No interference problems. a) A spectra of an analyte at various concentrations. b) Univariate calibration curve.

If the interferents are in the sample mixture, multivariate calibration methods give outlier warnings, which is not the case of univariate calibration methods. Figure 3.8.a shows overlapping spectrum and in this spectrum, the absorbances of the analyte will not be proportional to the concentration of the sample. Therefore the calibration curve will not be a straight line and the prediction ability will get worse. (Figure 3.8.b) Traditionally, interferent should be removed or an interference–free wavelength must be found for plotting the calibration curve. If there is not enough time, to get the desired selectivity in the results, interferents should be removed mathematically, which is the case of the multivariate calibration methods.

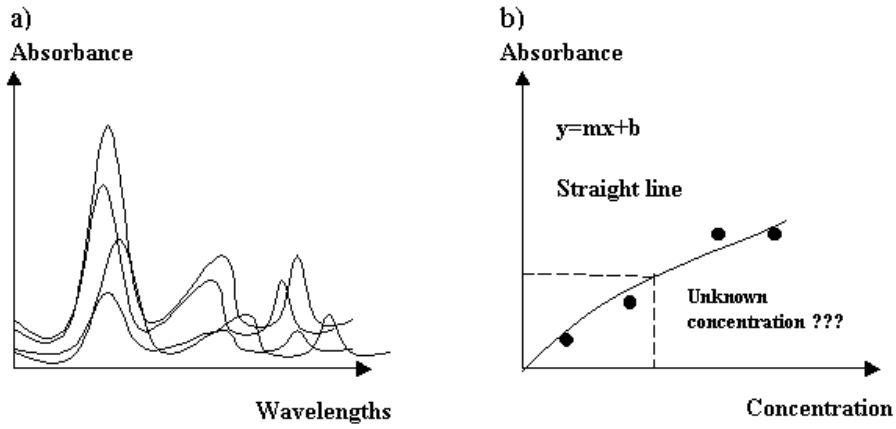

Figure 3.8.Chemical interferences from other constituents. a) Spectra of a multi component mixture that have overlaps; b) Univariate calibration curve.

Also some variations in the sample spectrum will change the precision of results. For example, the sample preparation condition is very important, since it may change the spectra of the analytes that we are interested in. Again, if there is not enough time for extra sample preparation, it can be changed mathematically using multivariate calibration methods.

3.3.1.2 Physical Interferences:

"Physical Interferences" means systematic errors in quantitative determination of a chemical constituent caused by physical effects rather than the effect of other chemical constituents with similar instrument response. [8]

The physical phenomena in the samples can affect the measured signal strongly. To get the desired selectivity, such effects can be compensated mathematically. Figure 3.9.a illustrates one common type of physical interference due to the samples, namely that of light scattering in spectroscopy. The spectra include the baseline, analyte peak and the solvent peaks. It is strongly affected by light scattering. Figure 3.9.b shows the calibration line of the analyte that obtained from a univariate calibration method. To correct it, the baseline peak should be first subtracted from each spectrum in a way that is a primitive

multivariate calibration method. By using multivariate calibration methods there is no extra prior sample preparation.

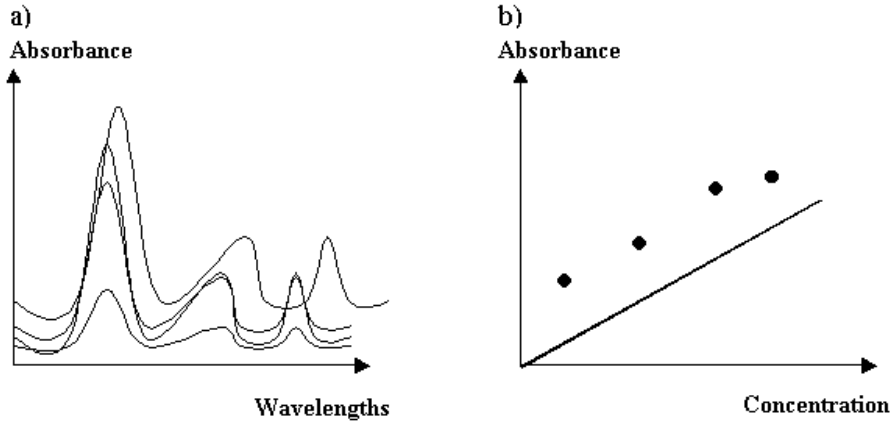

Figure 3.9: Physical interference from the sample: a) spectra of samples with baseline and solvent peak; b) calibration curve obtained from a univariate calibration method.

Another example of physical interference originating from the sample is the temperature effect on the analysis. Today it is well known in that the water absorption peaks in the near-infrared range is affected by temperature.

Chemical or physical interferences are not the only interferences in the measurement procedure. It should be concerned that the systematic errors are due to the way of the measurement made. Also problems of human mistakes must be taken into the consideration, especially in routine analysis, where each sample cannot be given too much attention. Univariate calibration methods cannot correct for such effects, multivariate calibration methods, it is possible to build a 'bridge' across such discontinuities. With them, many types of the sample abnormalities or instrumental problems can automatically be detected as outliers. These methods are:

- Classical Least Squares (CLS)
- Inverse Least Squares (ILS)
- Principle Component Analysis (PCA)
- Principal Component Regression and Partial Least Squares (PCR &PLS)
- Genetic Regression (GR)
- Genetic Classical Least Squares (GCLS)
- Genetic Inverse Least Squares (GILS)
- Genetic Partial Least Squares (GPLS)

3.2 CLASSICAL LEAST SQUARES

This method assumes that the modeling is based on Beer's Law. With this law, the absorbance at each frequency is proportional to the analyte concentrations. [9,10,13,14] It defines a relationship between four different variables:

- the spectral response (A_λ)
- the constituent absorptivity constant (ε_λ)
- the pathlength of light (b)
- the constituent concentration (c)

The goal of spectroscopic quantitative method is solving for the absorptivity constants. However, if the pathlength of the samples is kept constant then Beer's law can be written as:

$$A_\lambda = K_\lambda C \qquad 3.24$$

In this notation K is a combination of the absorptivity coefficient and pathlength.

If we have a single sample this equation can be solved easily. Only the absorbance of the sample is measured, then with known concentration, the value of K_λ is calculated. For predicting the concentration of an unknown sample calculating is as simple as measuring absorbance at the same wavelength. Finally arranging the Equation 3.24, the unknown concentration can be calculated as:

$$C = \frac{A_\lambda}{K_\lambda} \qquad 3.25$$

Due to the limitations of noise, instrument error, sample handling error, and many other possibilities; it is the best way to measure the absorbances of a series of samples and calculate the slope of the best line through all the data points. Just like the univariate calibration.

If we have a sample with two constituents then the problem will become more complex. In mathematical solution the number of mathematical equations should equal to the number of the unknown variables. In this case it is necessary to have two equations.

$$\mathrm{A}_{\lambda_1} = \mathrm{K}_{\mathrm{a},\lambda_1} \times \mathrm{C}_{\mathrm{a}} \qquad 3.26$$

$$A_{\lambda_2} = K_{b,\lambda_2} \times C_b \qquad 3.27$$

where the $A_{\lambda 1}$, $A_{\lambda 2}$ are the absorbances of constituents at two different wavelengths; C_a, C_b are the concentrations of the constituents in the sample; and $K_{a,\lambda 1}$, $K_{b,\lambda 2}$ are the absorptivity constants for the two constituents at the indicated wavelengths. Two different solutions can be applied for this process. The first one is based on the assumption that does not include any interference. By it, the solutions of equations become simple and independent from each other. Since bands are well resolved. (Figure 3.10)

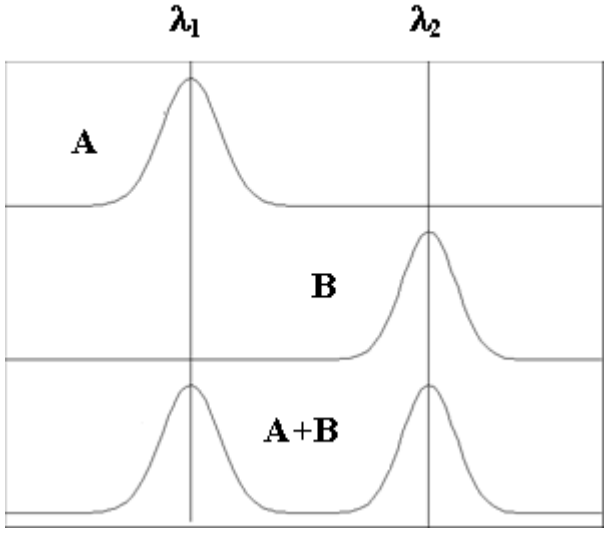

Figure 3.10. Hypothetical spectra of two different pure constituents A and B and a mixture of the two.

The second one is based upon another assumption, which includes interferences. That means the constituents in the sample have similar properties. (Figure 3.11)

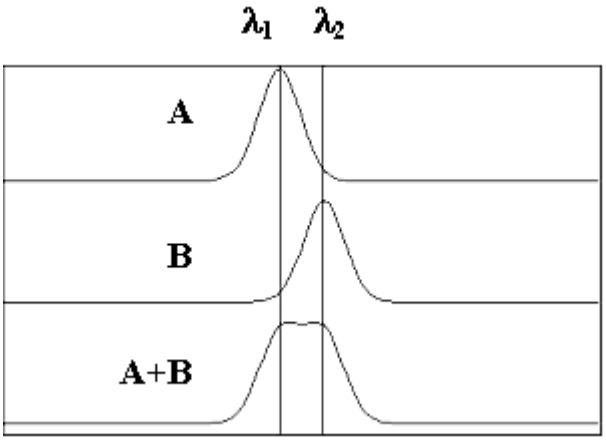

Figure 3.11. Hypothetical spectra of two alternative pure constituents, A and B, and a mixture of the two.

In this case the bands of the constituents spectra overlap and the equations must be solved simultaneously for both A and B. This solution can be done only taking the advantage of Beer's law; the absorbances of constituents at the same wavelength are additive. Thus the equations of two constituents for a single spectrum;

$$A_{\lambda 1} = K_{a\lambda_1} \times C_a + K_{b,\lambda_1} \times C_b \qquad 3.28$$

$$A_{\lambda 2} = K_{b\lambda_2} \times C_b + K_{a,\lambda_2} \times C_a \qquad 3.29$$

In the assumption there is no error in the measurements and in the predictive ability of the equations, these equations are used to predict unknowns. In real world it is impossible to find samples, there is always an error. Thus the error should be in problem solution.

$$A_{\lambda 1} = K_{a\lambda_1} \times C_a + K_{b,\lambda_1} \times C_b + E_{\lambda_1} \qquad 3.30$$

$$A_{\lambda 1} = K_{a,\lambda_2} \times C_a + K_{b,\lambda_2} \times C_b + E_{\lambda_2} \qquad 3.31$$

$E_{\lambda 1}$ and $E_{\lambda 2}$ are the spectral residual errors that are not fit by the model. In these conditions when two constituents samples were studied, errors were zero to obtain best line for the calibration model. However, with all the calibration models, Classical Least

Squares requires a lot of training samples with multiple constituents to build an accurate calibration model. With it, it is possible to build model simultaneously for all samples.

More than two constituents or wavelengths are getting the problem harder. An efficient way to solve it is matrix mathematics. It has also many calculations, but the importance of this way is its suitability for computers.

In matrix terms the Equations 3.30 and 3.31 can be formulated as:

$$\begin{vmatrix} A_{\lambda 1} \\ A_{\lambda 2} \end{vmatrix} = \begin{vmatrix} K_{a,\lambda 1} K_{b,\lambda 2} \\ K_{a,\lambda 2} K_{b,\lambda 2} \end{vmatrix} \begin{vmatrix} C_a \\ C_b \end{vmatrix} + \begin{vmatrix} E_{\lambda 1} \\ E_{\lambda 2} \end{vmatrix} \qquad 3.32$$

or more simply

$$A = KC + E \qquad 3.33$$

If this model extended to more complex mixture, it is should be as:

$$\begin{vmatrix} a_{1,1} \ldots\ldots a_{n,1} \\ a_{1,m} \ldots\ldots a_{m,n} \end{vmatrix} = \begin{vmatrix} k_{1,1} \ldots\ldots k_{1,n} \\ k_{1,l} \ldots\ldots k_{l,n} \end{vmatrix} \begin{vmatrix} c_{1,1} \ldots\ldots c_{1,l} \\ c_{m,1} \ldots\ldots c_{m,l} \end{vmatrix} + \begin{vmatrix} e_{1,1} \ldots\ldots e_{n,1} \\ e_{1,m} \ldots\ldots e_{m,n} \end{vmatrix} \qquad 3.34$$

where *A* is the *mxn* matrix of spectral absorbances (calibration spectra), *K* is the *lxn* matrix of absorptivity-pathlength constants, *c* is the *mxn* matrix of the constituent concentrations, and *E* is the *mxn* matrix of spectral errors or residuals not fit by the model. The subscripts indicate the dimensionality of matrix; m is the number of samples, n s the number of the data points (wavelength) used for calibration and *l* is the number of constituents in the sample mixtures.

Using matrix algebra, it is simple to solve the Equation 3.32 with computers. The first step to build model and find the best fit least squares line for the data is finding the value of the K. K represents the matrix of pure component spectra at unit concentration and unit pathlength. If the K matrix is solved; it can be used to predict concentrations of unknown samples. It can be showed simply as:

$$K = A \times C^{-1} \qquad 3.35$$

However, to calculate the inverse of a matrix requires that the matrix must be square. It means the calibration set has the same number of samples as constituents, but

practical it is not true. For best representation of the true calibration equation more samples should be used. Therefore an alternative solution is required.

$$K = A \times C^T (C \times C^T)^{-1} \qquad 3.36$$

This solution is a "pseudo-inverse" of *C* matrix. C^T is the transpose of the *C* matrix and "-1" refers the inverse of eqution.

This method is known as K matrix or Classical Least Squares (CLS). Also it can be considered as factor analysis method since the spectral absorbances matrix is represented as the product of two smaller matrices C and K.

Classical Least Square is a full-spectrum method therefore it can provide significant improvements in precision, allow simultaneous fitting of spectral base lines and make available for examination and interpretation least squares estimated pure-component spectra and full-spectrum residuals. One interesting side of this property is that if the entire spectrum is used for calibration, the rows of K matrix are spectra of the absorptivities for each of the constituents. These will like similar to the pure component spectra.

In spite of these advantages, this technique has one major disadvantage. It requires knowing the complete composition (every constituents) of the calibration mixtures and predicted (unknowns) must be mixtures of the exactly the same constituents. If the concentration of any constituents accepted as an outlier the predicted absorbance will be incorrect. Therefore, CLS can be applied to only the samples in which constituents' concentration well known. [15] In addition, there must be no interference in the sample. If there is, the K matrix will not be reflecting the absorptivities of the constituents and the predictions of unknowns. The CLS can be applied to the samples that contain only minimal or no inter constituents. Therefore it performs efficiently for gas–phase's samples.

3.3 INVERSE LEAST SQUARES

In real world the constituents of the samples interfere with each other and the exact composition of sample cannot be known. So the CLS cannot be applicable. To eliminate the disadvantage of Classical Least Squares, the Inverse Least Squares (ILS) assumes that the responses of the samples depend on the concentration. It accepts that the Beer's law can be taken inversely. Therefore the equation rearranged by taking the advantage of algebra. There are two solutions: The one is:

$$C = \frac{A_\lambda}{\varepsilon_\lambda b} \tag{3.37}$$

And the other is combining the absorptivity coefficient and the pathlength in a single constant.

$$C = PA_\lambda + E \tag{3.38}$$

In this equation C is the *mxl* matrix of constituent concentrations, A is the *mxn* matrix of spectral absorbances, P is the *nxl* matrix of the unknown calibration coefficients relating the l component concentrations to the spectral intensities, and E is the *mxl* vector of random concentration error or residuals that are not fit by the model. Since the model error is presumed to be error in the component concentrations, this method minimises the squared errors in concentration during calibrations. Consider the following equations:

$$C_a = A_{\lambda_1} \times P_{a,\lambda_1} + A_{\lambda_2} \times P_{a,\lambda_2} + E_a \tag{3.39}$$

$$C_b = A_{\lambda_1} \times P_{b,\lambda_1} + A_{\lambda_2} \times P_{b,\lambda_2} + E_b \tag{3.40}$$

by these equations, if the constituents of sample does not known exactly, the matrix of coefficients can still be calculated correctly.

This model known, as Inverse Least Squares (ILS) or Multiple Linear Regression (MLR) or P matrix and it seem as a best approach for the quantitative analysis. There is no need to recognise every components in the sample. Again like CLS, p matrix of absorbances is not square the "pseudo-inverse" must be used.

$$P = C \times A^T (A \times A^T)^{-1} \tag{3.41}$$

Notice that the ILS is different from the CLS. The inverse representation of Beer's law has given a significant advantage. "The analysis based on this model is invariant with respect to the number of chemical constituents." With this assumption the model also can be reduced for one component at a time.

$$c = Ap + e \tag{3.42}$$

where c is the mx1 vector of concentrations of the constituents, p is the nx1 vector of calibration coefficients, and e is the mx1 vector of concentration results not fit by the model. That means, if one constituent in the sample is known, the model can be applied to the all samples of interested. During the calibration step, the least squares estimate of p is:

$$\hat{p} = \left(A^T A\right)^{-1} A^T \times c \tag{3.43}$$

where $\hat{p}$ is the estimated calibration coefficients. Once $\hat{p}$ is calculated then the concentration of the analyte of interest can be predicted with the equation below:

$$\hat{c} = a^T \times \hat{p} \tag{3.44}$$

where $\hat{c}$ is the scalar estimated concentration and a is the spectra of the unknown sample.

This model seems to be best modelling technique in all words. Since it can build models accurately when only the one component is known in the sample. The only requirement is selecting wavelengths that correspond to the absorbances of the desired constituents.

Unfortunately the ILS calibration has some drawbacks. Due to the dimensionality of the matrix equations, the number of selected wavelengths cannot exceed the number of training samples that are occurred the calibration set. But in theory it should be possible to measure many more training samples to allow for additional wavelengths. However the absorbances in a spectrum tend to all increase or decrease together the concentrations of the constituents in the sample mixture and that is caused a new problem, which is called *collinearty*. This effect causes the mathematical solution to become less stable for the each constituent. Another problem with adding more wavelengths to the model is an effect known as *over fitting*. Generally it improves the prediction ability of the model. However when too much information in the spectrum is used to calibrate, the model starts

to include spectral noise, which is unique to the training set and the prediction accuracy for unknown samples.

In the ILS the wavelength selection is very important. For this reason many sophisticated algorithms are used to choose the "best" set of the wavelengths that represent the each constituent in the sample mixture. In spite of these disadvantages, ILS allows the calibration of very complex mixtures since only knowledge of constituents of interest is required and it is relatively fast.

3.4 THE EIGENVECTOR QUANTITATION METHODS

In real samples, there are many different variables that make a spectrum. These are:

- the constituents of sample
- inter-constituent interactions
- instrument variations (i.e., detector noise)
- changing environmental conditions that affect the baseline and absorbance
- differences in sampling handling

Even with all these complex changes occurring, there should be finite number of these variations when the spectral data are occurring. Hopefully the largest variations in the calibration set would be the concentrations of the constituents of the mixtures in the spectrum. If it was possible to calculate a set of "variation spectra" that represented the changes in the absorbances at all the wavelengths in the spectra, then this data could be used instead of the raw spectral data for building the calibration model. The "variation spectra" could be used to reconstruct the spectrum of sample by multiplying each one by a different constant scaling factor and adding the results together until the new spectrum closely matches the unknown spectrum. Each spectrum in the calibration set would have a different set of scaling constants for each variation since all the concentrations of the constituents are different. Therefore, the fraction of each spectrum that must be added to construct the unknown data should be related to the concentration of the constituents. The "variation spectra" are often called eigenvectors or spectral loadings or loading vectors or

principle components or factors. The scaling constants used to reconstruct the spectra are generally known as scores.

The eigenvectors must relate to the concentrations of the constituents that make up the samples, since they came from the original calibration data.

The calculated scores are unique to each separate principle component and training spectrum, and can be used in place of absorbances. Since the representation of the mixture spectrum is reduced from many wavelengths to a few scores as shown in Figure 3.12. This method is combining both the CLS and ILS methods together in the same calculation. Since it is better than the classical models in the meaning of accuracy and robustness.

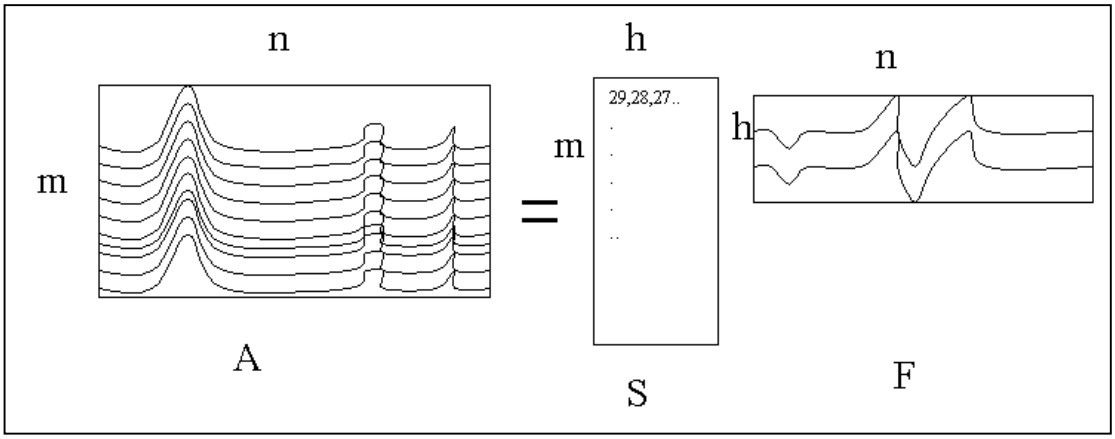

Figure 3.12. PCA breaks the spectral data into most common spectral variations (factors, eigenvectors, loadings) and the corresponding scaling coefficients (scores)

The trick in using these models comes from the calculation of the eigenvectors. These models are based on the concentration predictions and changes in the data, not the absolute absorbance measurements that are used in all Classical models.

In order to calculate the PCA model, the spectral data must change in some way. To accomplish this it is the best way to vary the concentration of the constituent. Since there can be problem with collinearty. For example, if the concentrations of the two constituents present always in same ratio, the model will detect only one constituent not two**.** Also not only the concentration of the constituents if the absorbance peak of the A increase or decrease when constituent B also increases or decreases, only one variation

will be detected and this is the changes in the mixture of A and B. Therefore, it is very important to have randomly concentration ratios in the mixtures.

3.5 PRINCIPLE COMPONENT ANALYSIS AND PRINCIPLE COMPONENT REGRESSION

One way to calculate all possible variation in the spectra is Principle Component analysis (PCA). It requires a group of training spectra that represents the composition of the sample that is interested in. The main property of this set is it should contain the range of expected for the unknown samples.

PCA is effectively process elimination. By this property, it is possible to create a set of eigenvectors (principle components) that are presentation of the changes in the absorbances. When the training data set has been fully processed by the PCA algorithm, it is reduced to two main matrices: the eigenvectors and the scores. The matrix representation shown as:

$$A = S \cdot F + E_A \qquad 3.45$$

where A is mxn matrix of spectral absorbances, S is mxh matrix of score values for all of the spectra, and F is an hxm matrix of eigenvectors. The E_A mxn matrix is the error or spectral residuals that are not fit by the model. The dimensions of the matrices are representative of the data they hold; m is the number of samples (spectra), n is the number of the data points (wavelength) used for calibration, and f is the number of eigenvectors.

The model equation should look familiar, and in fact it is very similar to the CLS model for the spectra. However, the spectral data is not constructed from the concentrations and absorptivity coefficient spectra. And there is no limit to the number of the wavelengths that can be used; so all the data up to the entire spectrum can be included in the model. The concentrations matrix C has not played a role in the calculations at all. Therefore, PCA cannot be used alone as a model for predicting constituent concentrations.

The eigenvectors are represented the spectral variations that are common to all the calibration data. The F matrix from PCA performs a similar task to the K matrix in the CLS model; it stores the constituent spectral data. However this does not mean the rows

of F matrix are the spectra of the pure constituents because they are not. On the other hand, the scores in the S matrix are unique to each calibration spectra and as a spectrum it is represented by a collection of absorbances at a series of wavelengths. Just like classical models it is possible to regress *C* against the scores *S*. In this case, the regression technique from the ILS model is the best choice. Since, there is no priori knowledge of the complete sample composition and some robustness in the original calibration mixtures. The model equation then will become:

$$C = B \cdot S + E_C \qquad 3.46$$

where *C* is the *mxn* matrix of constituent concentrations, *B* is an *mxh* matrix of the regression coefficients, and *S* matrix is the scores from the PCA model. *m* is the number of the constituents that are used in the calibration set, n is the number of the samples (spectra), and *h* is the number of eigenvectors. As with ILS, the B coefficients matrix can be solved by the regression:

$$B = \left(S \cdot S^T\right)^{-1} CS^T \qquad 3.47$$

Thus the name of this type of this regression called is Principle Component Regression (PCR), since it combines both PCA and ILS to solve the calibration equation for the model. By rearranging the model, a single unified equation will come up to represents PCR model.

$$S = AF^T \qquad 3.48$$

It is not necessary to use pseudo-inverse of the F to solve this equation. Since the F matrix is a special type of matrix that is called *orthonormal matrix*. When this matrix is multiplied by it's own transpose, the *identity matrix* is the result. Multiplying any matrix by the identity matrix is the same as multiplying a single number by one; the result is always the number again.

Finally by combining the concentration equation with scores equation, the final PCR model equation emerges:

$$C = BAF^T + E_C \qquad 3.49$$

The PCR model is not completely free of problems. Since PCR is a two-step process; PCA eigenvectors and scores are calculated and then the scores regressed against

the constituent concentrations using a regression method similar to ILS. N the first step, PCA calculates the factors and/or scores independtly of any knowledge of the constituents' concentrations. Hopefully the variation spectra are the results of the constituents' concentrations but there is no guarantee.

3.6 PARTIAL LEAST SQUARES

Partial Least Squares (PLS) is another method that also calculates the variations in the spectra. It is soft modeling techniques in which the data are decomposed into new variables that are linear combinations of the original data. This new variable is named as principal components or factors and therefore, PLS is often called factor methods. The way in which the new variables are created can be visualized for a two dimensional system. If the instrument responses for a set of *m* samples at two wavelengths (n=2) are plotted against each other, a new axis is formed in the direction that represents maximum variability of the data. This new axis is called first principle component or first eigenvector. If all the samples fall on this new axis, then all of the variations can be described using only one eigenvector. [26] Otherwise a second eigenvector can be found that is perpendicular or orthogonal to the first eigenvector. The second one describes the maximum amount of residuals, not fit by the first one, in the data set and so on. If more than two wavelengths are included in instrument response matrix, the plotting space becomes multidimensional and several eigenvector can be found, each one successfully accounting for the maximum possible amount of remaining variability and each orthogonal to others. In general, the number of principle component or factor that can be generated is less than or equal to the number of sample. [16]

PLS is full-spectrum method so it retains the full spectrum advantages of CLS. However, al of the component concentrations need to be known because, both the PLS can perform the analysis one component at a time while avoiding the ILS wavelength selection problems. PLS and PCR differ in the way the matrix of the spectra decomposed into two smaller matrices. In the PCR, decomposition is performed independently of analyte concentration whereas in the PLS, the concentration information is used to extract

factors. Therefore, the PLS method is expected to provide better calibration models and prediction. [17] The model for either the PLS is described as:

$$\boldsymbol{A} = \boldsymbol{TB} + \boldsymbol{E_A} \qquad 3.50$$

where A is the same before, B is a hxn matrix of basis vectors or loading spectra. T is an mxh matrix of intensities or scores in the new coordinate system defined by the h loading vectors. E_A is now the mxn matrix if spectral residuals not fit by the model. The difference between CLS and these factor methods is that the loading vectors in B are not pure component spectra but they are linear combinations of the original calibration spectra. Also the intensities in the new coordinate system are no logger constrained to the concentrations as were in CLS, but modeling can be done to relate the scores in T to the component concentrations. The number of basis vectors, h, to represent original calibration spectra is determined by an algorithm during the calibration step.

The spectral intensities in the new coordinate system can be related to the concentrations of the analyte with an ILS model given by:

$$\boldsymbol{c} = \boldsymbol{Tv} + \boldsymbol{e_c} \qquad 3.51$$

where c is the *mxl* vector of component concentrations, v is the *hxl* vector of coefficients which relate spectral intensities to the component concentration and e_c is the *mxl* vector of errors in reference values of the component that is being analyzed. The least-squares solution for v is similar to the Equation 3.43 in ILS, however, since the columns of the T matrix are orthogonal, inversion of the diagonal ($T^T T$) matrix is trivial. The estimate of v vector is given as:

$$\hat{\boldsymbol{v}}_h = \left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{T}^T\boldsymbol{c} \qquad 3.52$$

where $\hat{v}_h$ is the least-squares estimate of v. The T and B matrices are calculated in a stepwise manner (one vector at a time) until the desired model has been obtained. As mentioned earlier, PLS and PCR differ in the way they generate T and B matrices. In the PCR model, NIPALS (nonlinear iterative partial least squares) algorithm developed by Wold [20] is used. The NIPALS algorithm extracts the full spectrum loading vectors without using concentration information in the decomposition of spectral matrix A. Therefore; the prediction of component concentrations is expected to be poorer that the

results obtained by PLS which applies a modified version of NIPALS algorithm. [16] This modified version of the algorithm uses concentration information in the process of obtaining loading vectors thereby resulting in a generator predictive ability.

There are two PLS methods that are available today in the analysis of complex chemical mixtures. These are called PLS1 and PLS2 methods. In the PLS1 method, the analysis performed one component at a time and other component concentrations not included in the model building step. This is the most commonly used form the PLS method and it is reported that the predictions obtained with PLS1 are better that those obtained PLS2. It is suggested that PLS2 algorithm should be used for qualitative application.

Before applying the factor based methods to the data, it is common practice to do some sort of data pretreatment such as mean centering and scaling. [21] The mean centering is usually applied to both calibration spectra and corresponding analyte concentrations in which the average concentrations for the component of interest are subtracted from each spectrum and from given component concentrations, respectively. After the data pretreatment, a CLS calibration model is selected for the analysis of one component at a time. Then the PLS1 algorithm starts with the calculation of the estimated first weighed loading vector, $\hat{w}_h$, by setting h to 1. This is done with the method of least squares and is given by:

$$\hat{w}_h = A^T c\left(c^T c\right)^{-1} \qquad 3.53$$

where $\hat{w}_h$ is an *nxl* vector representing the first order approximation of the pure component spectra for the component that is being analyzed. This weighted loading vector is then used to form the score vector $\hat{t}_h$, with an ILS prediction model. The method of least squares is used to regress A on $\hat{w}_h$ which produces the first estimated $\hat{t}_h$ vector as given:

$$\hat{t}_h = A\hat{w}_h \qquad 3.54$$

With a linear least-squares regression, this score vector can be related to the component concentrations. The scalar regression coefficient, $\hat{v}_h$, is estimated by:

$$\hat{v}_h = \hat{t}_h{}^T c\left(\hat{t}_h^T \hat{t}_h\right)^{-1} \quad 3.55$$

The least-square estimated regression coefficient is later used to obtain concentration residuals. In order to eliminate collinearty problems, the PLS loading vector, $\hat{b}_h$, is now calculated with a new model for A. Once again the method of least squares is used to find estimated b vector by:

$$\hat{b}_h = \hat{t}_h{}^T A\left(\hat{t}_h^T \hat{t}_h\right)^{-1} \quad 3.56$$

where $\hat{b}_h$ is an *nx1* vector. It is now possible to calculate the first PLS approximation to the calibration spectra by multiplying the score vector ($\hat{t}_h$) with transpose of PLS loading vector ($\hat{b}_h^T$). The first residual matrix is calculated by subtracting the PLS approximation matrix from A matrix. The residuals in concentration vector calculated in a similar manner where scalar regression coefficient ($\hat{v}_h$) is multiplied with score vector and this product is subtracted from original concentration vector. The following equations provide residuals in both *A* and *c*.

$$E_A = A - \hat{t}_h \hat{b}_h^T \quad 3.57$$

and

$$e_c = c - \hat{v}_h \hat{t}_h \quad 3.58$$

This is the end of the first iteration in the calibration step. This is the process is repeated for a desired number of loading vectors by incrementing *h*, substituting E_A for A and e_c for concentration in the first CLS calibration model at the beginning of the algorithm.

The prediction step of PLS1 algorithm involves the calculation of final calibration coefficients, b_f, which have the dimension of an original spectrum. Once the b_f is calculated, it is possible to calculate the concentration of a new sample using the average concentration of the analyte and its spectra. The following equations show the prediction step in PLS1.

$$b_f = \hat{W}\left(\hat{B}\hat{W}^T\right)^{-1}\hat{v} \quad 3.59$$

where $\hat{W}$ and $\hat{B}$ contains individual $\hat{w}_h$ and $\hat{b}_h$ vectors, respectively and $\hat{v}$ is formed from individual regression coefficients ($\hat{v}_h$) The final prediction equation is then given as:

$$\hat{c} = a^T bf + c_o \qquad 3.60$$

where $\hat{c}$ is the predicted unknown sample c a is the spectrum of that sample and c_0 is the average concentration of calibration samples.

The process of determining the optimal number of PLS factors may vary from algorithm. The cross-validation approach is one of the methods for this. [22] For *m* calibration spectra, the PLS1 algorithm is performed on m-1 spectra and the left out spectrum is used to validate the model. This process is repeated until each spectrum is left out once in the calibration set. The predicted concentration for each left out sample is then compared with their original values and the prediction error sum of the squares (PRESS) is calculated for each added factor. The PRESS is a measure of how well a particular model fits the calibration data and given by:

$$PRESS = \sum_{i=1}^{m} \left(\hat{c} - c_i\right)^2 \qquad 3.61$$

where c_i is the reference (known) concentration of the *ith* sample and concentration is the predicted concentration of the *ith* sample for m calibration standard.

It is not the minimum PRESS value, however, that is used for the selection of optimal number of PLS factors since this may lead to over fitting resulting in a poorer prediction. Therefore a comparison needs to be done between two models that contain h and h+1 factors. Here, the better model is the one with smaller number of factors where the difference between the two PRESS values is determined by the F test to be significant. Figure 3.13 is an example of the determination of the optimal PLS factors. In this example the optimum number of factor is 6.

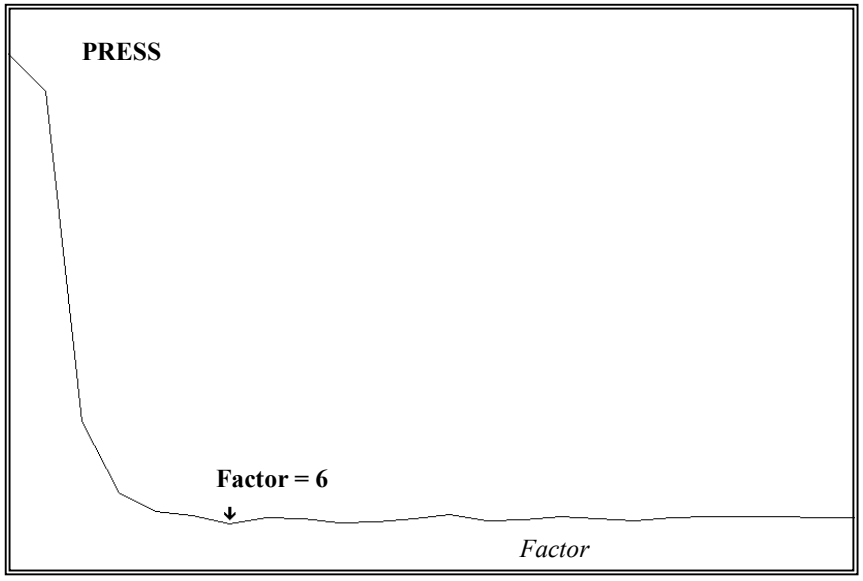

Figure 3.13: The graphic of PRESS to Factor Number

Although factor based calibration methods (PLS and PCR) eliminate most of the shortcomings of the hard modeling multivariate calibration methods such as CLS and ILS, they are much more complicated in terms of the mathematics involved in the decomposition of spectral matrix and obtaining basis vectors.

3.7 GENETIC REGRESSION

In principle, the performances of each calibration method are differentially affected by each underlying factors. Briefly CLS is a multivariate least-square procedure based on Beer's law. The CLS model accounts for errors in the spectral measurements. CLS can accommodate spectral intensities at all frequencies for all calibration samples. All overlapping spectral components should be known for optimal performance of CLS. By being a full-spectrum method, CLS has the ability to achieve improved precision since there is signal-averaging effect when many or all the spectral intensities are included in the analysis. ILS is a least-squares method that uses the inverse of Beer's law as its model. The ILS model accounts for errors in the reference concentrations. ILS is a frequency-limited method and, therefore, is not capable of the precision improvements of CLS from

signal averaging of multiple intensities. However, ILS can often be a useful method even if only one component is known for the calibration samples.

PLS and PCR are both factor-based methods that are capable of being full-spectrum methods. Like ILS, PLS and PCR can be employed when only one component is known in the calibration samples. Both PLS and PCR methods factor the spectral data calibration matrix into the product of two smaller matrices. This amounts to a data compression step where the intensities at all frequencies used in a new full-spectrum coordinate system. This new coordinate is composed of loading vectors that can be used to represent the original spectral data. The intensities in the new full-spectrum coordinate system are then used in a model where concentration is presumed to be a linear function of these intensities. Thus PLS and PCR are methods that concerned with modeling both spectra and concentrations during calibration. PCR performs the factoring of the spectral data matrix without using information about the concentrations. Therefore, there is no guarantee that the full-spectrum basis vectors that are associated with PCR are relevant for concentration prediction. On the other hand PLS performs the spectral factoring trying to account for the spectral variation while assuring that the new basis vectors relate to the calibration concentrations. Thus, the PLS sacrifices some fit of the spectral data relative to PCR in order to achieve better correlations to concentrations during predictions. [23]

In recent years, a new approach to the calibration has been reported where a genetic algorithm is used to optimize several linear regression models. [44-48] This new method is called Genetic Regression (GR) and is considered to be a hybrid calibration method since it uses full-spectrum information and obtains a single score for each constituents in the calibration samples. GR performs the calibration process with this score using simple linear regression. The idea behind genetic regression is simple to understand and apply but as powerful as any other multivariate calibration methods in the analysis of samples with multiple constituents.

3.7.1 Genetic Algorithms

Genetic Algorithms (GA) are global search and optimization methods based on the principles of natural evolution and selection as developed by Darwin. [49] For thousands of years, human beings have tried to answer the ultimate questions "why, how, and when did life exactly start on the earth" and produced several hypothesis. According to the Darwin's theory of evolution "struggle for life and survival of the fittest ", individuals better fitted to the environment they live in more likely survive and breed, thus passing their genetic information to their offspring. Individuals who are not fit and unable to adapt will eventually be eliminated from the population. This process progresses slowly over a long period of time (or may never end) through generations and the species will evolve into better and fit forms.

In the last couple of decades, scientists have been trying to take advantages of the natural evolutions as an improvement concept in the process of solving large-scale optimization problems. In the 1960's biologists have begun to perform the simulation of genetic systems experiments with computer. The pioneering work in genetic algorithms was done by Holland who developed a GA in his research on adaptive systems in the early 1960's and is considered the father of the field. [30] Over the years, GA's have received increasing attention and have been applied to large number of global optimization problems in many areas of applied science. [30-39] Lucaiasius and Kateman pioneered the first applications of genetic algorithms to calibration problems in analytical chemistry in the late 1980's [40,41] Since then, there have been several applications of GA's to wavelength selection [22,23,42-48] and calibration transfer problems in spectroscopy. [24-26]

Computationally the implementation of a typical GA is quite simple and consists of five basic steps including initialization of gene population, evolution of the population, and selection of the parent genes for breeding and mating, crossover and mutation, and replacing the parents with their offspring. These steps have taken their names from the biological foundation of the algorithm.

A gene is a potential solution to a given problem. The exact form of a gene may vary from application to application and depends upon the problem being investigated. In

this study, the term gene is used to describe a collection of wavelength pairs combined with simple mathematical operator (+, ×, –, and /) and each gene produces a score, which relates the instrument response to the constituent concentration. The term population is used to describe the collection of individual genes in the current generation. In order to evaluate each gene's success in the prediction of analyte concentration, a fitness function such as the inverse of standard error of calibration (SEC) and standard error of prediction (SEP), which are the derivatives of standard deviation, is used

3.7.2 Genetic Regression

Genetic regression is an implementation of a GA for selecting wavelengths and mathematical operators to build calibration models. GR is a hybrid calibration method, which optimizes simple linear regression models through an evolving selection of wavelengths and simple mathematical operators. (+, ×, –, and /). The advantage of GR is that it uses a simple underlying model that is easy to understand and explain while applying the optimization power of a GA. GR follows the same basics initialize/breed/mutate/evaluate algorithm as other GA's but differs in the way it encodes genes. Most GA's use a bit field representation for encoding the gene to simplify computer manipulation. The GR algorithm used here has a simple structure consisting of a wavelength pair and a mathematical operator. The implementation of GR consists of five basic steps as in most GA's shown in Figure 3.14

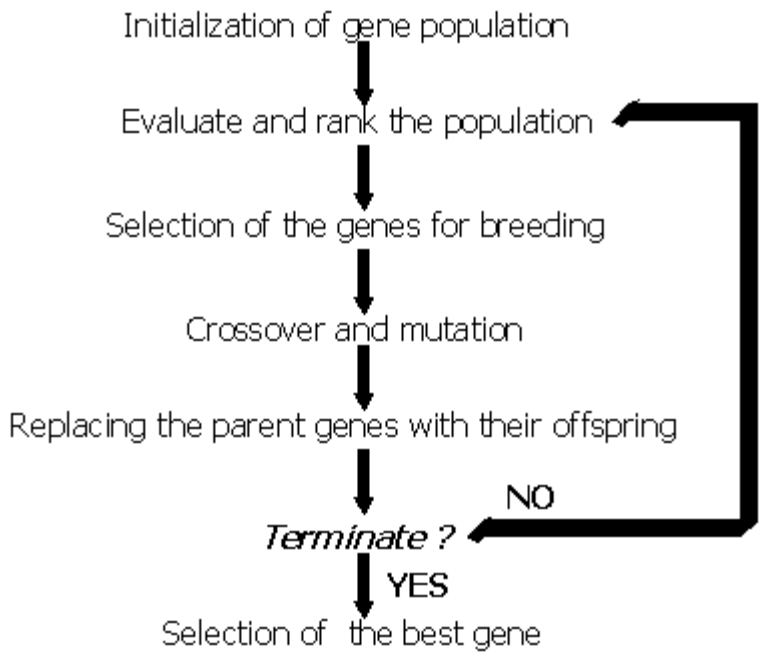

Figure 3. 14. Flow chart of the Genetic Regression (GR) program.

3.7.2.1 Initialization

The initialization step randomly creates the first generation of genes with a fixed population size. Although random initialization helps to minimize bias and maximize the number of possible recombination, GR is designed to select initial genes in somewhat biased random fashion in order to start with genes better suited to the problem than those that would be randomly selected. Biasing typically done with a function, which is orthogonal to the fitness function used to evaluate and rank genes later. This becomes very important when the search space (spectra in case) contains in large regions, which do not contain any useful information. However this biasing must not be too strict in order to avoid being trapped in a local maximum and to ensure a large variance in the initial genes. The size of the gene pool is a user-defined parameter in the GR. Though it is possible to optimize the population size, no extensive study was done for this purpose other than testing various values and observing how they affect the quality of the results. It is important to note that the larger the population size, the no longer the computation time.

An improvement in the diversity with a large number of genes results in lower computation speed.

In the initial gene pool, a gene consists base pairs between 2 and 50. A minimum of base pairs is a requirement for GR to allow mating and upper limit was set 50 to speed up the initialization. A base pair contains two randomly selected wavelengths and a randomly selected mathematical operator to combine these wavelengths. Each base pair is then added to give a score as shown:

$$S = (A_{5248} \times A_{6854}) + (A_{7954} - A_{8241}) + (A_{8417} - A_{8764}) \qquad 3.62$$

where S is so-called genetic score of the gene, A is the absorbance measured at the indicated wavelength. Figure 3.12 shows the schematic representation of a final best gene, for the prediction of a constituent in a mixture of a sample.

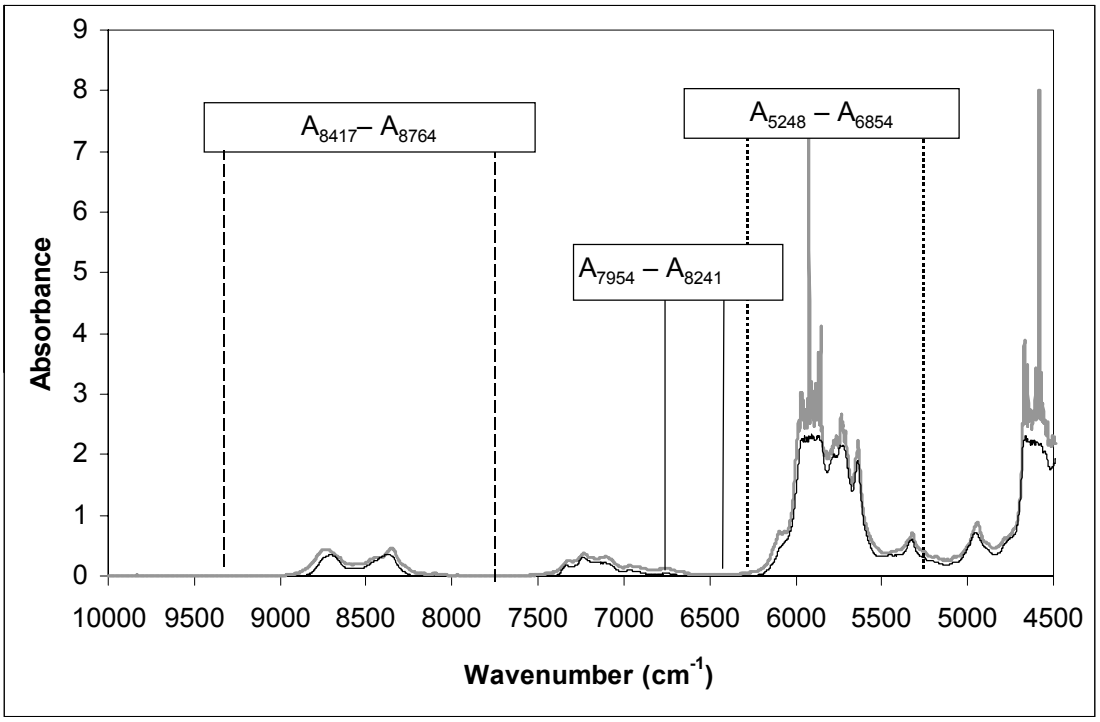

Figure 3.12. The schematic representation of a final best gene, for the prediction of a constituent in a mixture of a sample.

3.7.2.2 Evaluate and rank the population

This step involves the evaluation of the genes using fitness function, which is the inverse of standard error of calibration (SEC), followed by ranking process from the gene whose fitness is the highest to the one that has the lowest fitness. Here, a score is against known concentrations using simple least squares in the model-building step. The SEC is a derivative of the standard error (SE), which is calculated:

$$SE = \sqrt{\frac{\sum_{i=1}^{m} \left(\hat{c}_i - c_i^2\right)^2}{df}} \qquad 3.63$$

where c_i and $\hat{c}_i$ are the known and predicted analyte concentrations for m samples respectively, and df is the degrees of freedom in the calculations given by:

$$df = m - k \qquad 3.64$$

where k is the number of parameters extracted from the data set. For a calibration data set where a linear model assumed, there are only two parameters to be extracted including slop of the line and the intercept. In this case the degrees of freedom would be equal to m–2. If the df term is replaced by $m - 2$ in the Equation 3.63, it becomes standard error of calibration (SEC) as given:

$$SEC = \sqrt{\frac{\sum_{i=1}^{m} \left(\hat{c}_i - c_i\right)^2}{m-2}} \qquad 3.65$$

The success of each gene is measured by its fitness value, which is obtained by taking the inverse of SEC (Fitness = 1/SEC)

3.7.2.3 Selection of the genes for breeding

This step is the basic principle of natural evolution is put to work for GR as in all GA's. It involves the selection of the parent genes from the current population for breeding according to their fitness value. The goal is to give higher chance to those genes with fitness so that only the best performing members of the population will survive in the long run and will be able to pass their information to the net generations. Here, it is expected that the genes

Better suited for the problem will generate even better offspring. The genes with the low fitness values will be given lower chance to breed and hence most of them will be unable to survive. There are number of selection methods that can be used for parent selection[49] Top down selection is one of the simplest methods for parent selection. After genes are ranked in the current gene pool, they are allowed to mate in a way that the first gene mates with the second gene, third one with the forth one and so on. All the members of the current gene are given a chance to breed.

Roulette wheel selection method, which is used in GR, is the one where the chance of selecting gene is directly proportional to its fitness. In this method, each slot in the roulette wheel represents a gene. The gene with the highest fitness has the biggest slot and the gene with the lowest fitness has the smallest slot. Therefore, when the wheel is spun, there is a higher chance of being selected for a gene with high fitness than for a gene with a low fitness. There will also the genes, which are selected multiple times and some of the genes will not be selected at all and will be thrown out from the gene pool. After all the parent genes are selected, they are allowed to mate top-down, whereby the first gene (G_1) mates with the second gene (G_2). G_3 with G_4 and so on until all the genes mates. Since no ranking is done for the roulette wheel selected genes, the genes with low fitness have a chance to mate with better performing genes, thus resulting in an increased possibility of recombination.

3.7.2.4 Crossover and Mutation

The genetic algorithm does most of its work in the breeding/mating step. The step involves breaking the genes at random points and cross-coupling them as illustrated in the following example:

Parents

$$S_1 = (A_{347} \times A_{251}) + \#(A_{352} + A_{415})$$
$$S_2 = (A_{265} - A_{235}) + \#(A_{536} / A_{322}) + (A_{243} + A_{342})$$

Offspring

$$S_3 = (A_{347} \times A_{251}) + (A_{536} / A_{322}) + (A_{243} + A_{342})$$
$$S_4 = (A_{265} - A_{235}) + (A_{352} + A_{415})$$

The points where the genes are cut for mating are indicated by#

Here the first part of S_1 is combined with the second part of the S_2 to give the S_3, likewise the second part of the S_2 to give S_4. This process is called single point cross over and it is the one used in GR. There are also another types of cross over methods such as two points cross over and uniform cross over, each having their advantages and disadvantages. In the uniform case, each gene is broken at every possible point and many possible combinations are possible in the mating step, thus resulting in more exploitation. However, it is more likely to destroy good genes. Single point cross over will not provide different offspring if both parent genes are identical, which may happen in the roulette wheel selection, and broken at the same point. To avoid this problem, two points cross over, where each gene is broken in two points and recombined, can be used. Single point cross over generally does not disturb a good gene but it provides as many recombinations as other types of cross over schemes. Also mating can increase or decrease the number of base pairs in the offspring.

Mutation, which introduces random deviations into the population, was also introduced into the GR during the mating step at a rate of 1% as is typical in GA's. Replacing one of the base pairs in an existing gene with a randomly generated new base pair usually does this. Mutation allows the GR to explore the search space and incorporate new material into the genetic population. It helps keep the search moving and can eject GR from a local minimum on the response surface. However, it is important not to set mutation rate too high since it may keep the GA from being able to exploit the existing population.

3.7.2.5 Replacing the parent genes by their offspring

After crossover, the parent genes are replaced by their offspring and the offspring are evaluated. The ranking process based on their fitness values follows the evolution step. Then the selection for breeding/mating starts all over again. This is repeated until a predefined number of iterations are reached.

At the end, the gene with the lowest SEC (highest fitness) is selected for model building, which is done by simple least squares. This model is used to predict the concentrations of component being analyzed in the validation (test) sets. The success of the model in the prediction of the validation sets are evaluated using standard error of prediction (SEP) which is calculated as:

$$SEP = \sqrt{\frac{\sum_{i=1}^{m} (\hat{c}_i - c_i)^2}{m}} \qquad 3.66$$

where m is now denotes the number of validation samples.

3.7.2.6 Termination

The termination of the algorithm can be done in many ways. The easiest way is to set predefined iteration number for the number of breeding/mating cycle. However no extensive statistical test has been done to optimize it, though it can also be optimized.

Because the random processes are heavily involved in the GR as in all the GA's, the program has been set to run any number times for each component in a given multi-component mixture. The best run, i.e. the one generating the lowest SEC for the calibration set and at the same time produced SEP's for validation sets that are in the same range with SEC was subsequently selected for evaluation and further analysis.

GR has some major advantages over classical univariate and multivariate calibration methods. It is a hybrid calibration method in which it uses full spectra information and reduced it to a singe score to build simple calibration models. First of all, it is as simple as univariate calibration in terms of the mathematics involved in the model building and prediction steps, but at the same time it has the advantages of the multivariate calibration methods since it uses the full spectrum to extract genetic scores. It automatically corrects baseline fluctuations with the use of simple mathematical operators while forming the base pairs. Also no data pretreatment is necessary before calibration, which saves the extra time in the data processing and it can use the data that are collected at different wavelength intervals.

Another big advantage of the GR is that it can be used as a multi-instrument calibration method or so-called calibration transfer without any further change on the collected on multiple instruments into a singe calibration set as if they were all collected on a single instrument and uses this model in the prediction of analyte concentrations whose spectra were collected on several different instruments.

3.9 GENETIC INVERSE LEAST SQUARES (GILS)

The major drawback of the CLS is that all of the interfering species must be known and their concentrations included in the model. This need can be eliminated using the inverse least squares (ILS) method, which uses the inverse of Beer's Law. In the ILS method, concentrations of an analyte are modeled as a function of absorbance measurements as mentioned previously. Because modern spectroscopic instruments are very stable and provide excellent signal-to-noise (S/N) ratios, it is believed that the majority of errors lie in the reference values of the calibration sample, not in the measurement of their spectra.

The major disadvantage of ILS can be seen in Equation 3.41 where the matrix, which must be inverted, has dimensions equal to the number of wavelengths in the spectrum and this number cannot exceed the number of calibration samples. This is a big restriction since the number of wavelengths in a spectrum will generally be more than the number of calibration samples and the selection of wavelengths that provide the best fit for the model is not a trivial process. Several wavelength selection strategies, such as stepwise wavelength selection and all possible combination searches, are available to build an ILS model that fits the data best. Here we used the same genetic algorithm by GCLS described later to build genetic inverse least squares (GILS) models with one difference. This difference is in the way the mating and single point crossover operations are carried out. Because the number of wavelengths is restricted in response matrix **A** in the ILS, the size of the largest gene is restricted to one less than the number of calibration samples in the concentration vector. However, if the single point crossover is set to take place in any point of a gene, then the mating step could produce new genes that have a larger number of wavelengths than the number of calibration samples even though all the genes in the initial gene pool were set to have smaller number of wavelengths than the size of the concentration vector. In order to avoid this problem, the crossover operation is only performed in the middle of each gene in GILS so that the new generations will never have larger sizes than the number of calibration samples. The rest of the algorithm is the same as the one used in GCLS. The genetic algorithm of GCLS can be described as:

In the initialization step, an even number of genes are formed from full a spectral data matrix and each gene are used to form a CLS model. These models are then evaluated and ranked using the fitness function described in GR. The roulette wheel method is then used to select the gene population for breeding. After the selection procedure is completed, the selected genes are allowed to mate top-down without ranking whereby the first gene mates with second gene and third one with fourth one and so on as described in above with one difference. Since the genes used in GCLS are only vector of wavelengths and contains no base pairs as described in GR, for each gene a random number is generated between 1 and the length of the gene and the single point crossover process is performed using this number. After crossover, the parent genes are replaced by their offspring and the offspring are evaluated. The ranking process is based on their fitness values and follows the evaluation step. Then the selection for breeding/mating starts all over again. This is repeated until a predefined number of iterations are reached. During the each iteration the best gene with the lowest SEC is stored in order to compare it with the best gene of the next generation. If the next generation produces a better gene then it is replaced with the older one; otherwise the old one kept for further iterations. At the end, the gene with the lowest SEC is selected for model building. This model is used to predict the concentrations of component being analyzed in the validation (test) sets as described in GR. Another difference between GR and GCLS is that there is no mutation step in GCLS.

CHAPTER 4
EXPERIMENTAL SECTION

4.1 ESTERIFICATION REACTIONS

The aim of this project is to investigate the possibility of analyzing the complex mixtures of carboxylic acids, alcohols, esters and water by using near infrared spectroscopy and multivariate calibration methods (GR, GILS, PCR, and PLS) and also to investigate the possibility of developing calibration models for the solutions of these compounds. Due to this reason, the performance of these calibration methods in developing these models were investigated and they were compared with each other. Later on, it will be searched if it is possible to use the obtained models to measure these compounds in their real processes (for ex; in industrial process)

As seen above these constituents are the component of the esterification reactions. Esterification reactions are the reaction between the carboxylic acids and the alcohols. They are based on condensation reaction. Figure 4.1 illustrates the esterification reaction: [50]

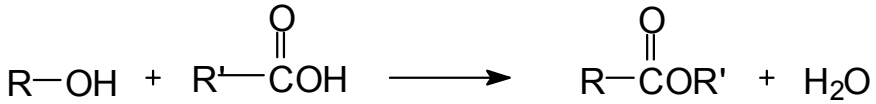

Figure 4.1. General reaction of esterification

These reactions are extremely slow without catalysts. Because of this reason these reaction are realized in the presence of homogeneous and/or heterogeneous catalysts. Mineral acids such as sulphuric acid, and organic acids such as p-toluene-sulphonic acid are the examples of homogenous catalysts and cation-exchange resins such as Amberlyst-15 or Dowex 50W are the examples of the heterogeneous catalysts. However homogeneous catalysts have not been used as much as in the chemical industry. Their separation from the product is very difficult and it needs expensive process constructions; since these are highly corrosive. Nowadays in many chemical processes heterogeneous catalysts are being used or planned to be used because of the following advantages: a) they eliminate the corrosive environment, b) the catalyst from the reaction mixture can be

removed by decantation or filtration, c) the purity of the products is higher since the side reactions can be eliminated or are less significant. [51,52]

The reaction mixtures contain organic materials, which are analyzed by titration, gas chromatography (GC) or liquid chromatography (LC). However it is not possible to measure the concentration values of all compounds by one method. For instance, water cannot be determined using GC. At least titration and chromatographic methods must be used to determine all of them. These methods are time-consuming and not reliable for the researchers. Recent advances in the instrumentation and multivariate calibration methods have increased the use of Near-Infrared Spectroscopy (NIR). And it has also been used for on-line or in-line monitoring when it is equipped with fiber-optic probes. And also in NIR collecting the data takes only seconds.

Many research workers have studied esterification of carboxylic acid with alcohol in the presence of homogeneous and/or heterogeneous catalysts. For the kinetic reaction many esterification reactions were investigated. [51-55] Problems seen in developing calibration models for complex systems are the chemical correlation and problems originating from different interferences in the obtained spectroscopic data. Because of this reason, it was investigated to remove the possible correlations and interferences to develop the most appropriate calibration models that can be used in real samples.

4.2 INSTRUMENTATION

In this project the spectra were collected with a FTS–3000 NIR spectrometer. (Bio–Rad, Excalibur, Cambridge, MA) This spectrometer was equipped with Tungsten – Halogen lamp as a source, Calcium Fluoride (CaF_2) as a beam splitter, and Lead Selenide (PbSe) as a detector. The samples were contained in Infrasil quartz cell with a pathlength of 2 mm and the data collections were done between the 4500–10000 cm^{-1}. Resolution was optimized to the 16 cm^{-1} and 64 scans were done. Duplicate measurements were done for each sample and background was taken as an empty infrasil quartz cell.

4.3 DATA ANALYSIS

The spectra were transferred to a separate PC after collection on the instrument and MS Excel (Microsoft Office 200, Microsoft Corporation) was used to prepare text files that are required for the methods used in this study. The new genetic algorithms based multivariate calibration methods (GR, and GILS) were written in MATLAB programming language using Matlab 5.3 (MathWorks Inc, Natick, MA) and PLS and PCR methods were taken commercially from Grams/32.

4.4 DESIGNS OF THE DATA SETS

Four sets that have 40 samples were prepared with carboxylic acids, alcohols, esters and water. The range of concentrations each constituent in the sets were in the range of 0 – 86 % for acid, 0 – 80 % for alcohol, 0 – 14 % for the water, and 0 – 66.4 % for esters by volume. These sets were the preparation of the methyl acetate, ethyl acetate, propyl acetate and buthyl acetate, respectively. In all process, the same values were taken. Table 4.1 and Table 4.2 show the concentration of he each constituent for the each set.

Table 4.1: Concentration profiles for calibration set. All concentrations are given in grams. (SN: Sample number, Acac: Acetic acid, Alc: Alcohol, Est: Ester, and W: Water.)

SN	W	Alc	Est	Acac
1	1.75	9.53	1.84	8.89
2	1	9.30	11.55	2.37
4	3	4.54	7.35	7.99
5	0.5	3.26	15.23	5.14
8	0.63	14.65	7.88	0.88
9	1.75	10.70	5.78	4.94
10	1.25	4.65	3.94	11.85
11	2.25	6.51	4.20	9.28
13	1.5	13.25	4.46	3.95
14	0.25	8.14	6.30	7.90
15	2.5	4.65	4.20	10.67
17	3.25	8.14	6.56	5.33
20	0.5	8.37	5.25	8.30
21	1.75	7.21	2.36	10.47
24	3.4	8.32	2.26	8.30
25	0.3	3.44	20.48	1.19
28	1	10.00	4.46	7.11
29	2.9	15.35	4.04	1.38
31	1	4.42	4.90	11.52
33	0.62	0.47	16.67	6.32
34	1.7	5.95	16.62	0.85
36	2.63	0.70	11.68	8.30
37	2.26	7.85	10.55	3.36
39	2	6.51	9.98	5.14
40	0.87	5.70	6.43	9.39

Table 4.2: Concentration profiles for prediction set. All concentrations are given in grams. (SN: Sample number, Acac: Acetic acid, Alc: Alcohol, Est: Ester, and W: Water.)

SN	W	Alc	Est	Acac
3	2	9.53	3.15	7.70
6	2.5	4.65	8.40	7.51
7	0.75	6.98	4.46	9.88
12	1.12	6.98	6.17	8.30
16	1	5.12	6.56	9.68
18	2.38	6.16	11.81	3.75
19	1.63	9.53	11.15	1.98
22	1.37	8.72	7.49	5.62
23	0.8	7.21	3.15	10.63
26	1.25	10.00	3.68	7.51
27	1.05	8.56	3.41	9.09
30	1.3	2.33	15.80	4.86
32	0.25	6.98	11.81	4.74
35	1.33	5.70	2.72	11.81
38	2.73	10.55	5.39	4.58

CHAPTER 5

RESULTS AND DISCUSSION

All samples were analyzed using NIR spectrometer and the data were collected for the prediction. Each set corresponding to the esterification reaction was divided into two sets: One was for calibration and the other was for prediction. Calibration set contained 50 spectra of 25 samples and prediction set contained 30 spectra of 15 samples. These samples which were in calibration or prediction set were chosen randomly. We only paid attention on calibration set because the samples have maximum and minimum concentration value, were in this set. Concentrations of all constituents were predicted using PCR, PLS, GR, and GILS. The unit of concentration was taken as grams.

In the NIR spectral region the absorbance bands are often broad and overlapping. The NIR spectral changes that result from the varying concentration of the compounds in the esterification reaction mixture are difficult to interpret visually. The figures 5.2 to 5.5 show the spectra of pure components and their mixtures for each esterification reactions.

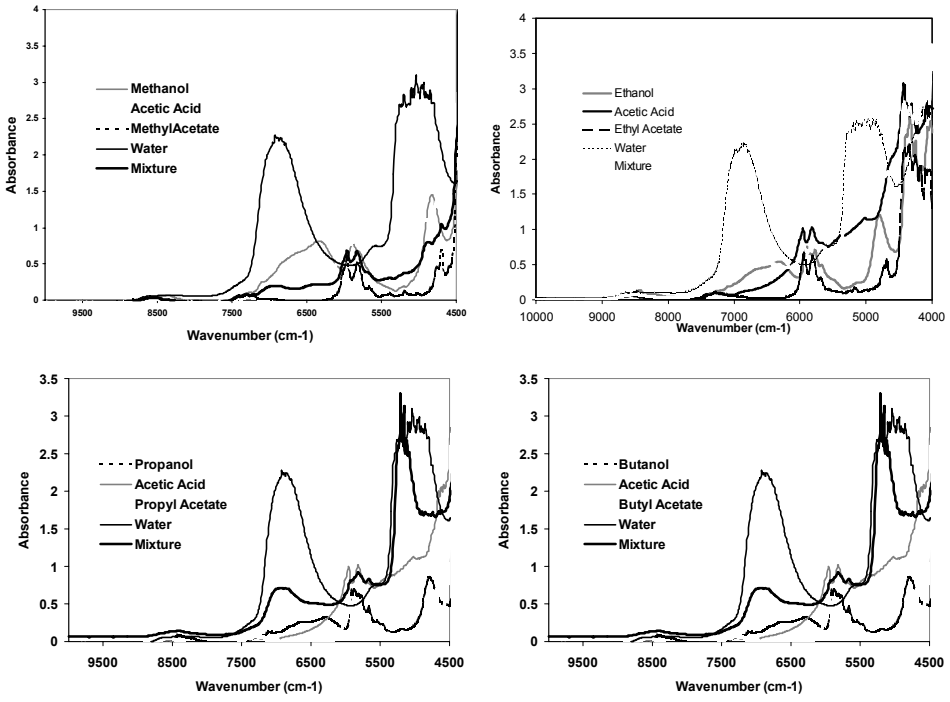

Figure 5.1. NIR absorbance spectra of each esterification process

From these spectra, it is evident that each constituent exhibits very similar spectral characteristics, which makes it necessary to use a multivariate calibration method to resolve the mixtures of these compounds. As seen from spectra, generally, the alcohols and acids' bands overlap esters' bands around 5900 cm^{-1}. Only the water band at 5200 cm^{-1} can be interpreted visually. There is no spectral disturbances in the spectra therefore preprocessing of the data is not required. Due to this reason, the performance of these four calibration methods in developing these models were investigated and they were compared with each other.

The PLS and PCR regression calibration models for each reactions were first calculated using cross–validation. And all data was mean–centered not scaled. Cross validation attempts to emulate predicting "unknown" samples by using the training set data itself. There are two main advantages of cross–validation methods. The first is estimation of the performance of the model. Since the predicted samples are not same as the samples used build the model. The second benefit of cross–validation is better outlier detection since each sample is left out of the models during the cross–validation process. On the other hand, it is a very time–consuming process. Mean centering translates the collection of data to the origin of multivariate space where analysis will be performed. It also removes the need for an intercept from the regression model. Since fewer terms in the regression model may need to be estimated and estimated analyte concentrations may be more precise following mean centering of the data. More of the information content of a data set can usually be described with a simpler model if the data is mean centered. The major effect of mean–centering is removing the broad sloping background from the data collection. Also PLS–1 algorithm (one component at a time) was used here. Using by these methods the standard error of calibration (SEC) was found between 0.0994 to 0.2497 grams and standard error of prediction (SEP) was found between the 0.08115 to 0.3157 grams

Genetic Regression (GR) is a hybrid calibration between univariate and multivariate calibration techniques in which it optimizes simple linear regression models through an evolving selection of wavelengths and simple mathematical operators (+, -, *, /). It is also used a full-spectrum of each sample. In this regression, data are not mean centered but cross-validated. It is a hybrid calibration method that uses the full spectral

information and reduces it to a single score upon which simple calibration models are built. First of all, it is as simple as univariate calibration in terms of the mathematics involved in the model building and prediction steps, but at the same time it has the advantages of the multivariate calibration methods since it uses the full spectrum to extract genetic scores. It automatically corrects baseline fluctuations with the use of simple mathematical operators while forming the base pairs. Note that no data pretreatment is necessary before calibration, which saves the extra time in the data processing. Using by these methods the standard error of calibration (SEC) was found between 0.1049 to 0.6249 grams and standard error of prediction (SEP) was found between the 0.0980 to 0.3157 grams

Genetic Inverse Least Squares (GILS) is new method. As it was explained above ILS uses the inverse Beer's Law and it can build the model if only one constituent is known in the samples. The only requirement is selecting the wavelengths that correspond to the absorbances of the desired constituents. However, the number of selected wavelengths can not be exceed the number of training samples that are occurred in the calibration set; due to the matrix dimensionality. If the training sample numbers are increased; additional wavelengths can be selected. However it causes collinearity and overfitting which affects the precision of the model. GILS is a modified version of the original ILS method in which a small set of wavelengths is selected from a full spectral data using a genetic algorithm. The algorithm used to select the optimum number of wavelengths in GILS is quite similar to the GR algorithm, but differs in the way it encodes the gene. In GILS, the term 'gene' describes a vector whose elements are randomly selected wavelengths. And the mating and single point crossover operations are carried out in the algorithm. Because the number of wavelengths is restricted in response matrix in the ILS, the size of the largest gene is restricted to one less than the number of calibration samples in the concentration vector. However, if the single point crossover is set to take place in any point of a gene, then the mating step could produce new genes that have a larger number of wavelengths than the number of calibration samples even though all the genes in the initial gene pool were set to have smaller number of wavelengths than the size of the concentration vector. In order to avoid this problem, the crossover operation is only performed in the middle of each gene in GILS so that the new

generations will never have larger sizes than the number of calibration samples. Using by these methods the standard error of calibration (SEC) was found between 0.0569 to 0.2105 grams and standard error of prediction (SEP) was found between the 0.09181 to 0.3214 grams

Overall data results are shown in Tables 5.1 to 5.4 for each esterification reactions. These results are obtained after several regressions were done. Then using predicting values, the actual values of each component were plotted against these predicted values and we seen the calibration plots. These plots are shown in Figures 5.2 to 5.17.

Table 5.1 The SEC. SEP, and R^2 results for all the components and all the methods for methyl acetate process.

Name of Method	Components	SEC	SEP	R^2 (SEC)	Factor Number
PLS	Acetic Acid	0.1923	0.2026	0.9969	6
	Methanol	0.1933	0.1081	0.9979	7
	Methyl Acetate	0.2342	0.2153	0.9974	6
	Water	0.0999	0.0811	0.9898	4
PCR	Acetic Acid	0.1181	0.068	0.997	11
	Methanol	0.1871	0.1975	0.9974	11
	Methyl Acetate	0.2425	0.2228	0.9977	11
	Water	0.1937	0.0986	0.9866	11
GILS	Acetic Acid	0.162	0.2139	0.9978	
	Methanol	0.1075	0.1991	0.9992	
	Methyl Acetate	0.2435	0.2775	0.9972	
	Water	0.0712	0.0918	0.9945	
GR	Acetic Acid	0.2617	0.2172	0.9942	
	Methanol	0.6249	0.3661	0.9734	
	Methyl Acetate	0.3637	0.2568	0.9948	
	Water	0.1225	0.098	0.984	

Table 5.2. The SEC. SEP, and R^2 results for all the components and all the methods for ethyl acetate process.

Name of Method	Components	SEC	SEP	R^2 (SEC)	Factor Number
PLS	Acetic Acid	0.1739	0.2194	0.9975	4
	Ethanol	0.2349	0.1813	0.9961	6
	Ethyl Acetate	0.2497	0.3157	0.9976	16
	Water	0.124	0.139	0.9837	7
PCR	Acetic Acid	0.1885	0.2151	0.9971	10
	Ethanol	0.2766	0.1675	0.9942	10
	Ethyl Acetate	0.4505	0.2555	0.9922	10
	Water	0.1352	0.1496	0.9807	10
GILS	Acetic Acid	0.1064	0.1716	0.999	
	Ethanol	0.1125	0.2292	0.9991	
	Ethyl Acetate	0.2105	0.2292	0.9983	
	Water	0.1053	0.1477	0.988	
GR	Acetic Acid	0.2891	0.2634	0.9929	
	Ethanol	0.4255	0.2744	0.9875	
	Ethyl Acetate	0.6648	0.6332	0.983	
	Water	0.1224	0.098	0.984	

Table 5.3. The SEC. SEP, and R^2 results for all the components and all the methods for propyl acetate process.

Name of Method	Components	SEC	SEP	R^2 (SEC)	Factor Number
PLS	Acetic Acid	0.0732	0.1518	0.9995	12
	Propanol	0.084	0.1338	0.9995	12
	Propyl Acetate	0.1654	0.1456	0.9989	11
	Water	0.0938	0.2826	0.9905	4
PCR	Acetic Acid	0.089	0.1484	0.9993	19
	Propanol	0.0847	0.1369	0.9995	19
	Propyl Acetate	0.1762	0.1662	0.9988	19
	Water	0.1181	0.2967	0.9855	19
GILS	Acetic Acid	0.0847	0.1644	0.9994	
	Propanol	0.0913	0.1735	0.9994	
	Propyl Acetate	0.1768	0.1735	0.9994	
	Water	0.0569	0.2812	0.9965	
GR	Acetic Acid	0.2589	0.2163	0.9943	
	Propanol	0.5438	0.5028	0.9797	
	Propyl Acetate	0.4406	0.4111	0.9924	
	Water	0.1049	0.2791	0.9882	

Table 5.4. The SEC. SEP, and R^2 results for all the components and all the methods for buthyl acetate process.

Name of Method	Components	SEC	SEP	R^2 (SEC)	Factor Number
PLS	Acetic Acid	0.1957	0.2542	0.9967	6
	Propanol	0.1579	0.1895	0.9987	7
	Propyl Acetate	0.1821	0.1656	0.9987	7
	Water	0.2109	0.0924	0.9526	4
PCR	Acetic Acid	0.1542	0.1699	0.9978	10
	Propanol	0.1542	0.1699	0.9984	10
	Propyl Acetate	0.2228	0.2051	0.9981	10
	Water	0.1702	0.1024	0.9748	10
GILS	Acetic Acid	0.1608	0.3214	0.9978	
	Propanol	0.1615	0.221	0.9988	
	Propyl Acetate	0.1615	0.2561	0.999	
	Water	0.0656	0.0736	0.9953	
GR	Acetic Acid	0.1782	0.2247	0.9973	
	Propanol	0.5516	0.5308	0.9792	
	Propyl Acetate	0.3864	0.3354	0.9942	
	Water	0.1409	0.0696	0.9785	

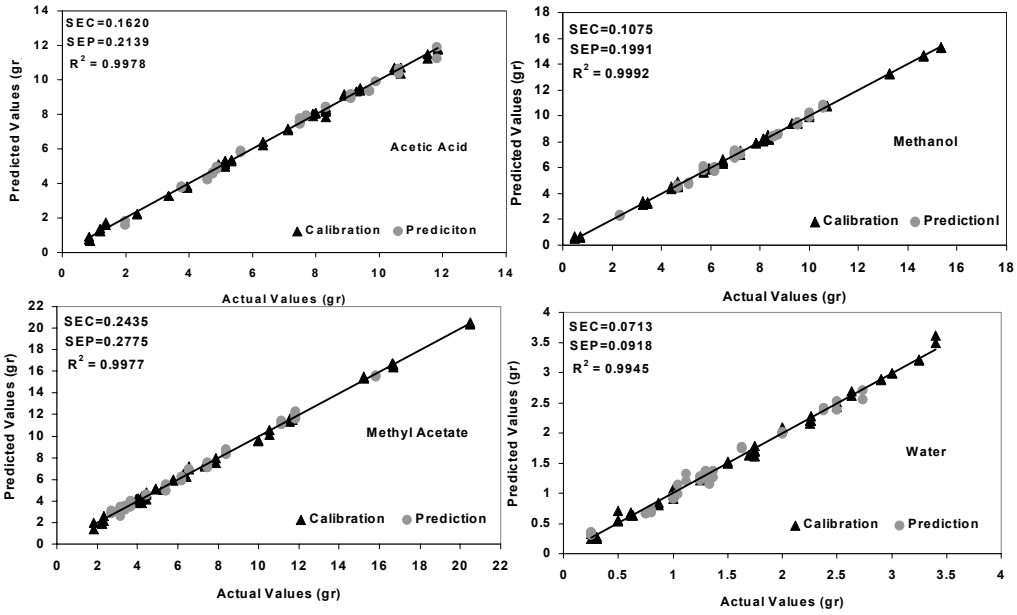

Figure 5.2. Calibration plots obtained with GILS for methyl acetate process

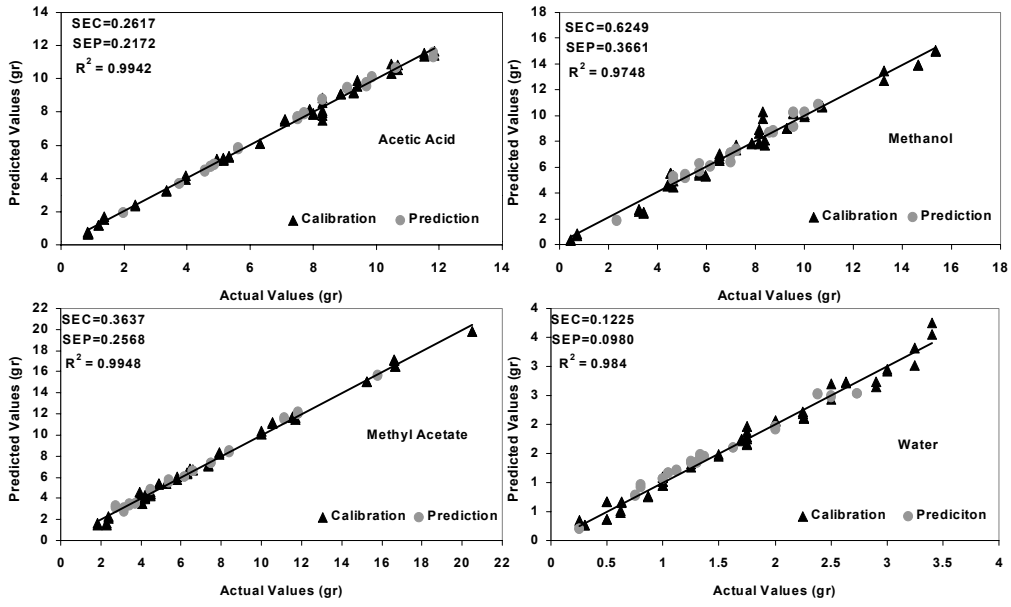

Figure 5.3. Calibration plots obtained with GR for methyl acetate process

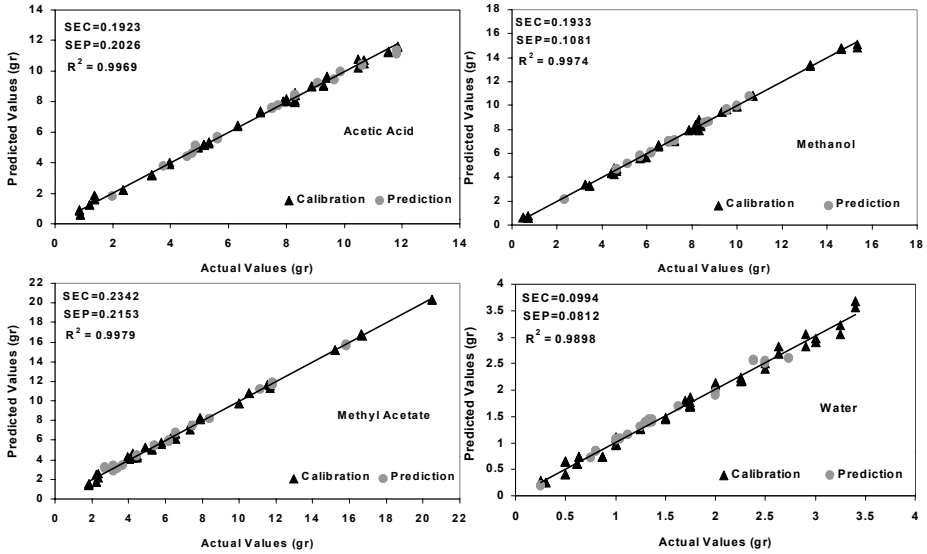

Figure 5.4. Calibration plots obtained with PLS for methyl acetate process

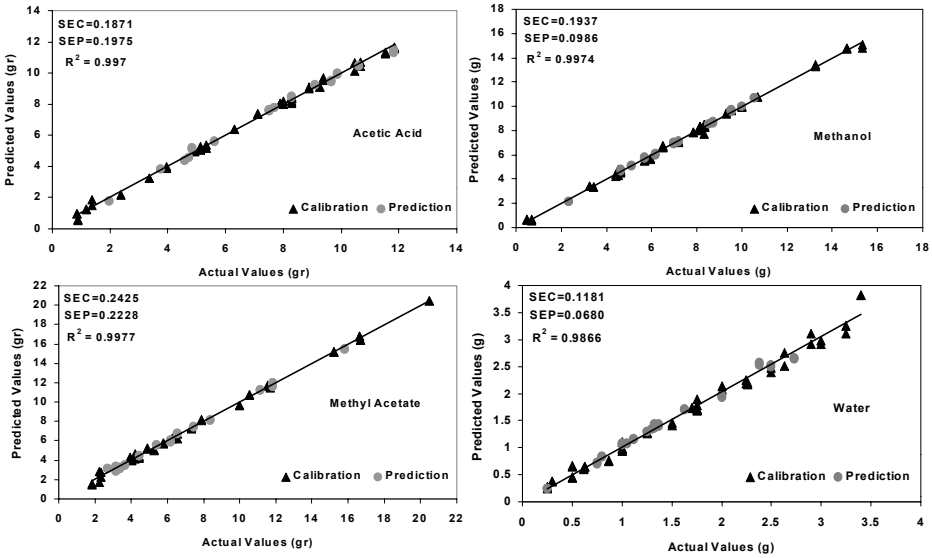

Figure 5.5. Calibration plots obtained with PCR for methyl acetate process.

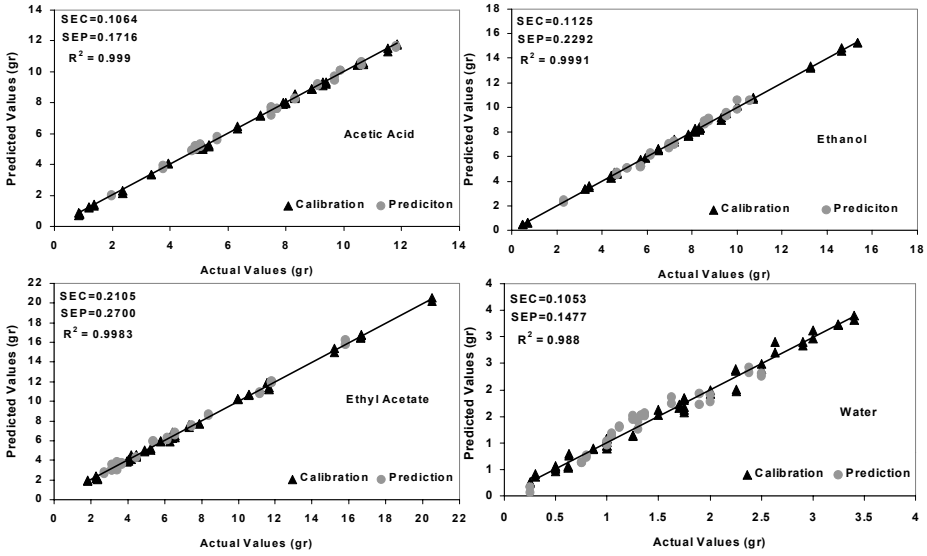

Figure 5.6. Calibration plots obtained with GILS for ethyl acetate process.

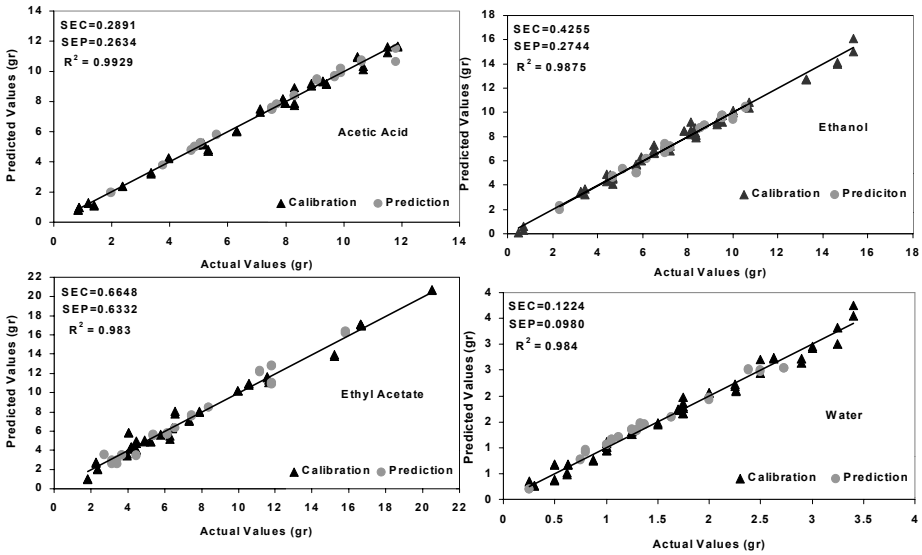

Figure 5.7. Calibration plots obtained with GR for ethyl acetate process.

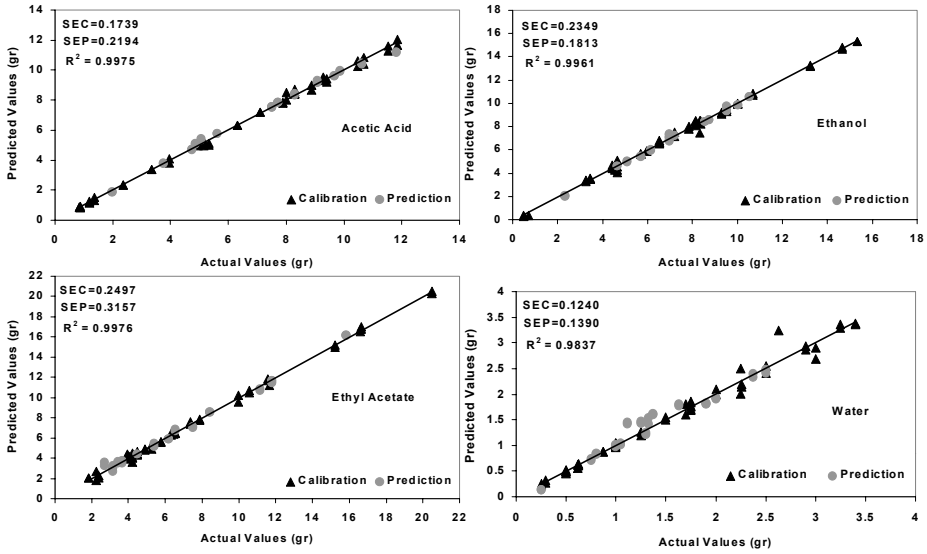

Figure 5.8. Calibration plots obtained with PLS for ethyl acetate process.

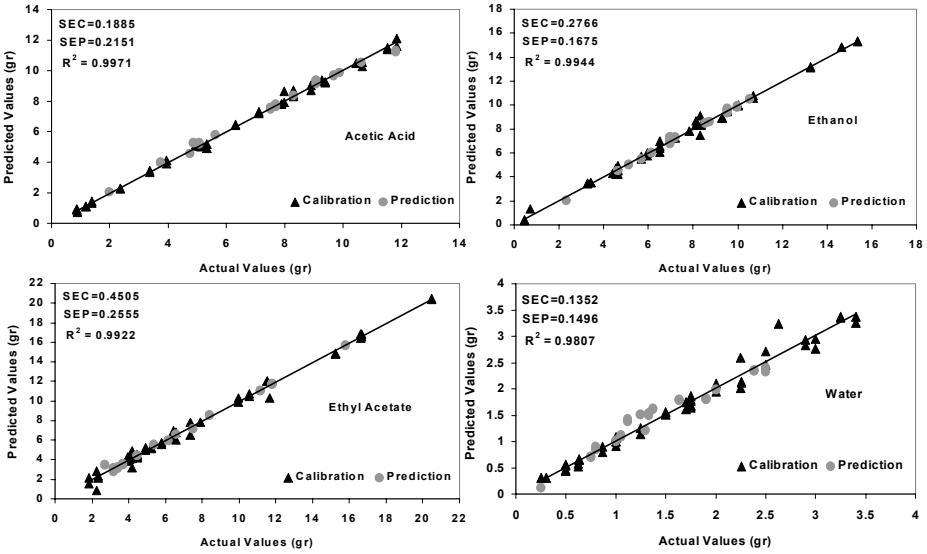

Figure 5.9. Calibration plots obtained with PCR for ethyl acetate process.

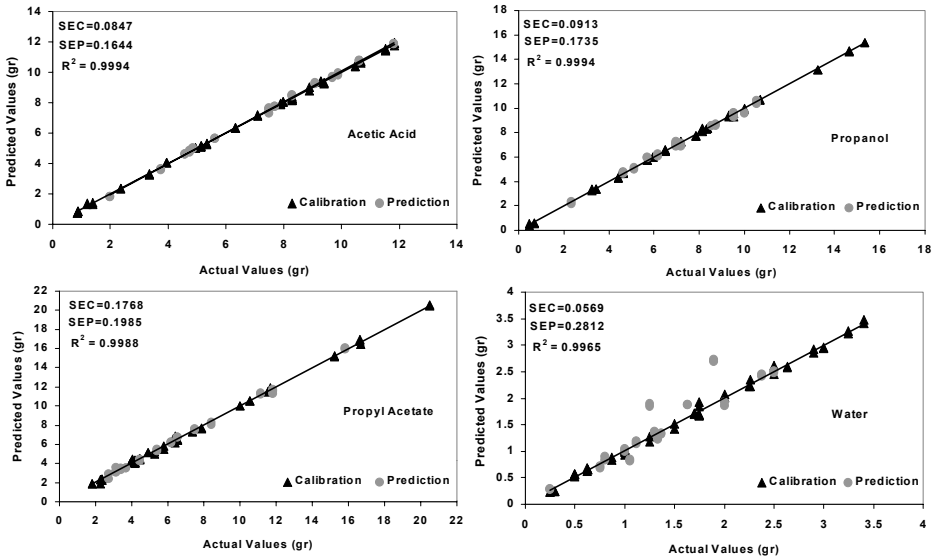

Figure 5.10. Calibration plots obtained with GILS for propyl acetate process.

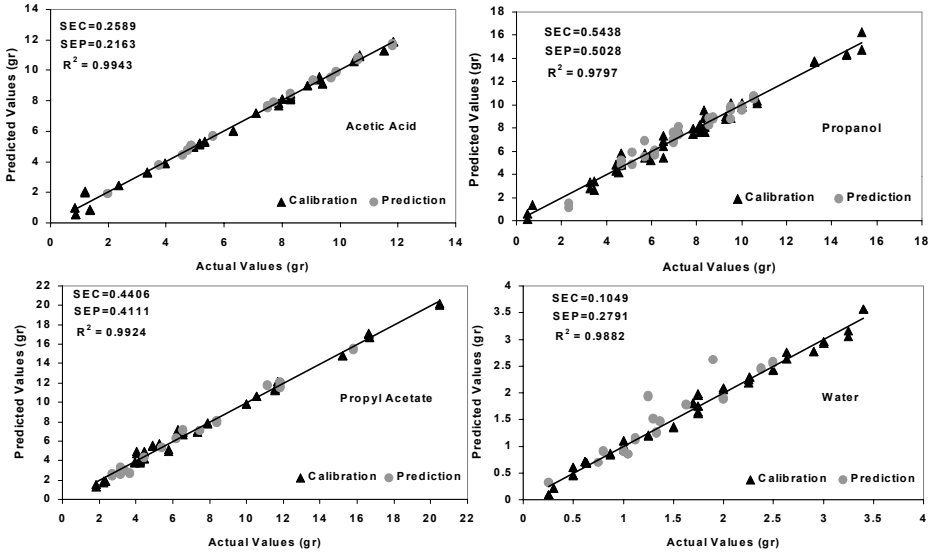

Figure 5.11. Calibration plots obtained with GR for propyl acetate process.

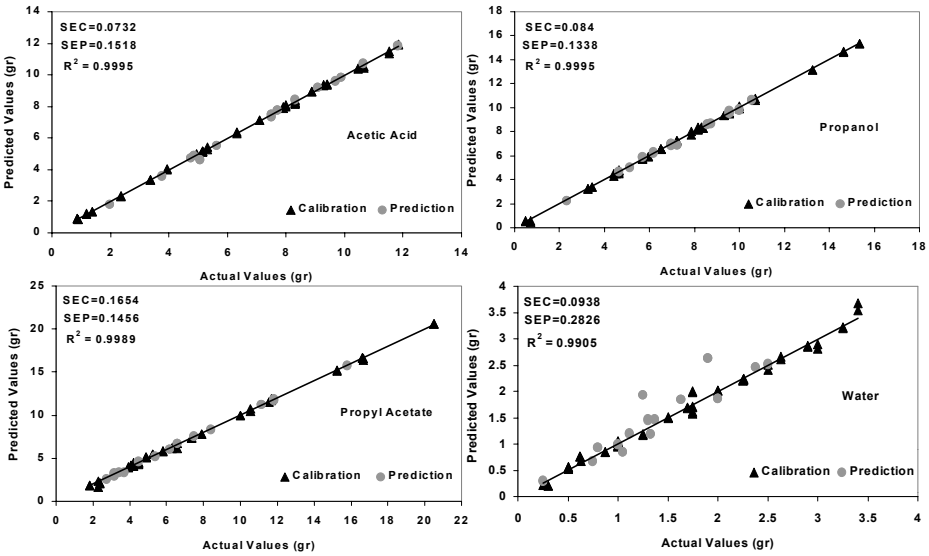

Figure 5.12. Calibration plots obtained with PLS for propyl acetate process.

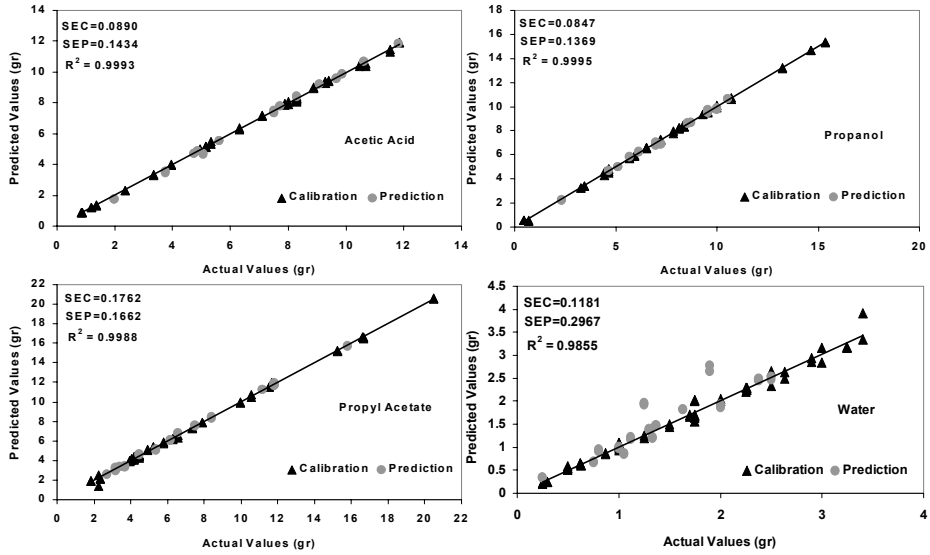

Figure 5.13. Calibration plots obtained with PCR for propyl acetate process.

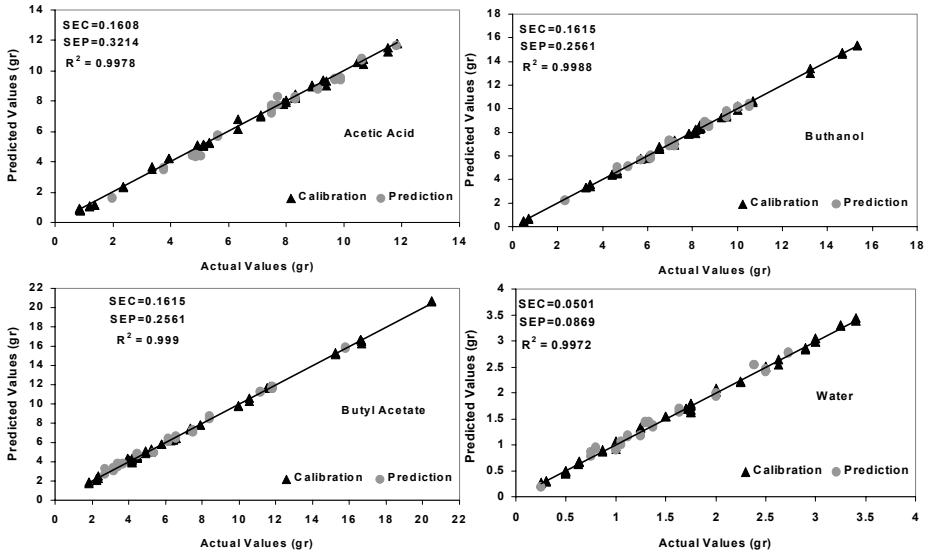

Figure 5.14. Calibration plots obtained with GILS for butyl acetate process.

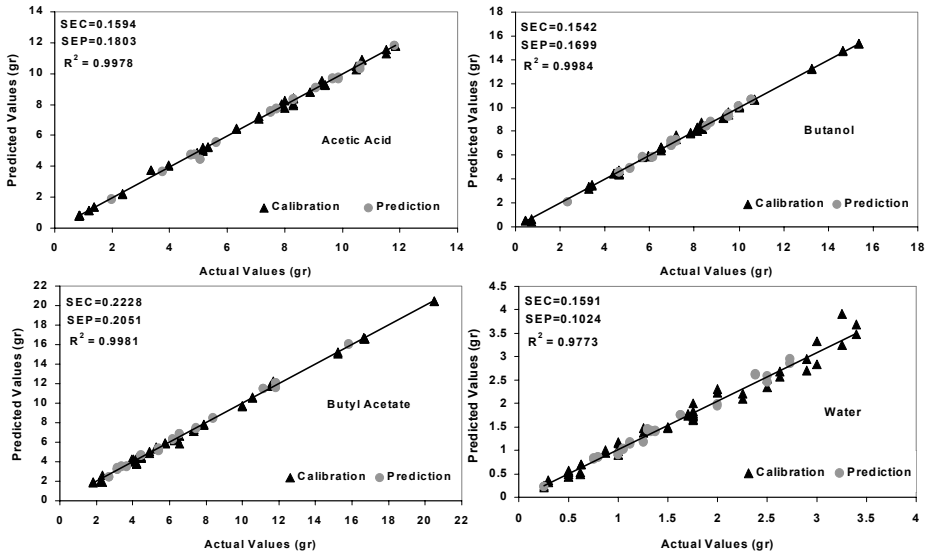

Figure 5.15. Calibration plots obtained with PCR for butyl acetate process.

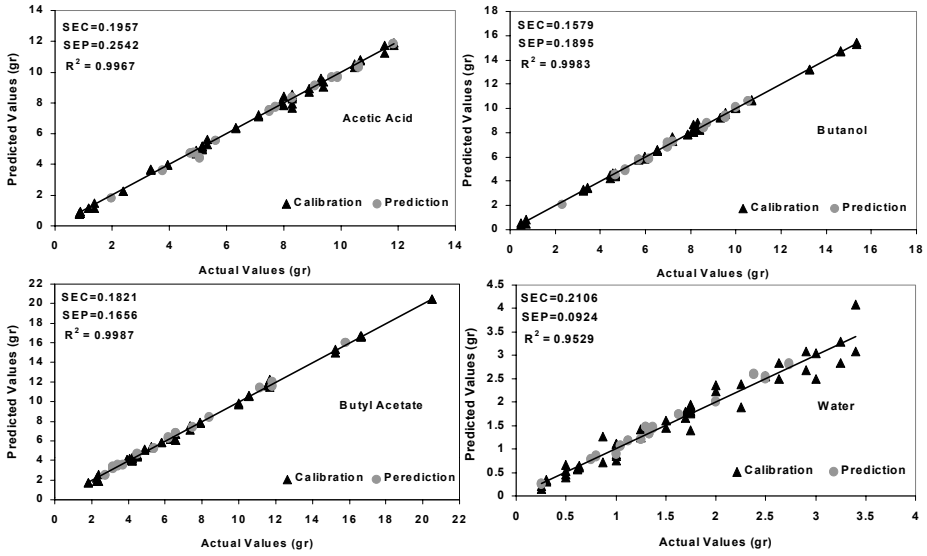

Figure 5.16. Calibration plots obtained with PLS for butyl acetate process.

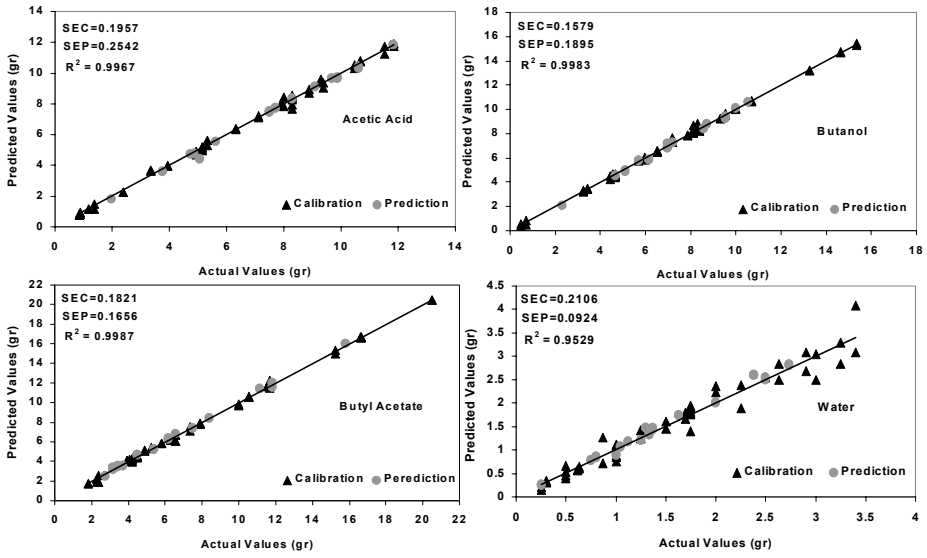

Figure 5.17. Calibration plots obtained with GR for butyl acetate process.

Among these plots and tables, it is seen that PLS and GILS are best regression methods; since they have minimum SEC value and best regression coefficient. And also when they tested they were shown to have the minimum SEP value. As we said above, the aim of this study is to find the best regression method for esterification process. To accomplish this aim the condition of homogeneity of two regression methods can be checked with the F-statistical test. The theoretical value for one-tailed test and calculated F values for each esterification process are shown in Table 5.5

Table 5.5. The results of F-test for each esterification reactions.

Methyl Acetate		
Constituent	**F**	**F Critical one-tail**
Water	0.97	0.62
Metanol	1.00	1.61
Methyl Acetate	1.00	0.62
Acetic Acid	1.00	1.61

Ethyl Acetate		
Constituent	**F**	**F Critical one-tail**
Water	0.97	0.62
Etanol	1.03	1.61
Ethyl Acetate	0.99	0.62
Acetic Acid	0.98	0.62

Propyl Acetate		
Constituent	**F**	**F Critical one-tail**
Water	1.00	0.62
Propanol	1.00	0.62
Propyl Acetate	0.99	0.62
Acetic Acid	1.00	1.61

Butyl Acetate		
Constituent	**F**	**F Critical one-tail**
Water	0.96	0.62
Butanol	0.99	0.62
Butyl Acetate	0.99	0.62
Acetic Acid	1.00	0.62

According these results, we concluded that these methods could be used monitoring the esterificaton process.

Chapter 6
CONCLUSION

The ability to build calibration models using collected NIR absorbance spectra has been successfully demonstrated with GILS, GR, PCR, and PLS. Several calibration methods were built. For all esterification process SEC and SEP values were calculated. The lowest SEC and SEP are selected for the best calibration model. Also actual vs. predicted values plots` regression coefficients values should be equal or closed to 1. The best results are obtained from GILS and PLS compared to the other methods. For GILS, it may be explained with the fact that ILS can be powerful multivariate calibration method when accompanied with proper wavelength selection methods. And it is also indicated that there is no interference in our sample.

REFERENCES

1. Skoog, D. A.; Holler, F. J.; Nieman, T. A.; *Principles of Instrumental Analysis*; Saunders College Publishing; 5th. Edition; 1998, 380
2. Workman, J. J.; "*Interpretive Spectroscopy for Near-Infrared*" *Applied Spectroscopy Reviews*; 31 (3); 1996; 251
3. Weyer, L. G.; "*Near Infrared Spectroscopy of Organic Substances*" *Applied Spectroscopy Reviews*; 21 (1&2); 1985; 1
4. Miller, C. E.; "*Near Infrared Spectroscopy of Synthetic Polymers*" *Applied Spectroscopy*; 26 (4); 1991; 277
5. Ingle, Jr., J. D.; "*Spectrochemical Analysis*"; Prentice-Hall Inc. NJ; 1988;404
6. Mcclure, W. F.; "*The Giant Is Running Strong*" Anal. Chem.; 66 (1); 1994; 43A
7. Drennen, J. K.; Kraemer, E. G.; Lodder, R. A.; "*Advances and Perspectives in Near Infrared Spectrophotometry*"; *Critical Reviews in Anal. Chem*.; 22 (6); 1991; 443
8. Martens, H.; Naes, T.; *Multivariate Calibration*; John Wiley &Sons Publication NY; 2001
9. Beebe, K.R.; Kowalski, B. R.; *Anal Chem*; Vol 59; No 17; 1987; p. 1007A
10. Thomas, E. V.; *Anal. Chem.*; Vol. 66; No. 15; 1994; p. 795A
11. Ryan, T. P.; *Modern regression Methods*; John Wiley & Sons Publication NY; 1997
12. Draper, N.R.; Smith, H; *Applied Regression Analysis*; John Wiley & Sons Publication NY; 1980
13. Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S.; "*Matrix representation and criteria for selecting analytical wavelengths for multicomponent spectroscopic analysis"; Anal. Chem*. Vol. 54; 1982; p. 1472
14. Haaland, D. M.; Thomas, E. V.; "*Partial least squares for regression analysis*"; *Anal. Chem*.; Vol 60; 1988 p. 1193-
15. Haaland, D. M.; Han, L.; Niemczyk, T. M.; "*Use of CLS to understand PLS IR calibration for trace detection of organic molecules*"; *Applied spectroscopy*; Vol. 53; No. 4; 1999; p. 390

16. Haaland, D. M.; Melgaard, D. K.; "*New augmented classical least squares methods for improved quantitative spectral analysis*"; *Vibrational Spectroscopy*; Vol 29; 2002; p. 171
17. Mark, H.; *Anal. Chem*.; Vol 58; 1986; p.2816
18. Martin, K. A.; *Applied spectroscopy*; 27; 1992; p. 325.
19. Lindberg, W.; Persson, J.; "*Partial Least Squares Method for Spectroflurometric Analysis of Humic Acid and Ligninsulfonate*"; *Anal. Chem*.; 55; 1983; p. 643
20. Wold, H.; *Multivariate Analysis*; John Wiley & Sons Publications NY; 1966; p. 391
21. Geladi, P.; Kowalski, B. R.; "*Partial Least Squares Regression*"; *Anal. Chim. Acta;* 185; 1986; p. 1.
22. Malinowski, E. R.; "*Theory of error in Factor Analysis*"; *Anal. Chem*.; Vol 4; 49; 1977; p. 606
23. Thomas, E. V.; Haaland, D. M.; "*Comparison of Multivariate Calibration Methods for Quantitative Analysis*"; *Anal. Chem*.; Vol 60; 1990; 1091-1099
24. Lawerence, D.; *Handbook of Genetic Algorithms*; Van Nonstrand Reinhold NY; 1991
25. Hartnett, M. K.; Bos, M.; Van Der Linden, W. E.; Diamond, D.; *Anal. Chim. Acta*; 316; 1995; 347
26. Hartnett, M. K.; Diamond, D.; *Anal. Chem*.; Vol 69; 1997; 1909
27. Shaffer; R. E; Small, G. W.; Arnold, M. A.; *Anal. Chem*.; Vol. 68; 1996; 2663
28. Weinke, D.; Lucaius, C. B.; Katemann, G.; *Anal. Chim. Acta*; 265; 1992; 211
29. Leardi, R.; Boggia, R.; Terrile, M.; *Journal of Chemometrics*; 6; 1992; 267
30. Gilbert, R. J.; Goodacre, R.; Woward, A. N.; Kell, D.B.; *Anal. Chem*.; Vol 69; 1997; p. 4381
31. Fontain, E.; *Anal. Chim. Acta*; 265; 1992; p. 227
32. Cong, P.; Li, T.; *Anal. Chim. Acta*; 293; 1994; p191
33. Wienke, D.; Lucasius, C. B.; Ehrlich, M.; Kateman, G.; *Anal. Chim. Acta*; 271; 1993; p. 253
34. Hibbert, D. B.; *Chemom. Intell. Lab. Syst*.; 19; 1993; p. 277
35. Lucasius, C. B.; Kateman, G.; *Trends Anal. Chem*.; 10; 1991; p 254

36. Lucasius, C. B.; Kateman, G.; "*Understanding and using GA*"; *Chemom. Intell. Lab. Syst*.; 19; 1993; p. 1
37. Bangolore, A. S.; Shaffer, R. E.; small, W.; Arnold, M. A.,; *Anal. Chem*.; Vol 68; 1996; p. 4200
38. Broadhurst, D.; Goodacre, R.; Jones, A.; Rowland, J. J.; Kell, D. B.; *Anal. Chim. Acta*; 348; 1997;p. 71
39. Acros, M. J.; Ortiz, M. C; Villahoz, B.; Sarabia, L. A.; *Anal. Chim. Acta*; 339; 1997; p. 63
40. Rimbaud, D. J.; Massart, D. L., Leardi, R.; De Noord, O. E.; *Anal. Chem*.; 67; 1995; 4295
41. Horchner, U.; Kalivas, J. H.; *Anal.Chim. Acta*; 315; 1995; p.1
42. Lucasius, C. B.; Beckers, M. L.; Kateman, G.; *Anal. Chim. Acta*; 286; 1994; p. 135
43. Li, T.; Lucasius, C. B.; Kateman, G.; *Anal. Chim. Acta*; 268; 1992; p. 123
44. Paradkar, R. P.; Williams, R. R.; *Appl. Spectrosc*; 52; 1996; p.753
45. Paradkar, R. P.; Williams, R. R.; *Appl. Spectrosc*; 51; 1997; p. 92
46. Ozdemir, D.; Mosley, R. M.; Williams, R. R.; *Appl. Spectrosc*; 52; 1998; p. 599
47. Ozdemir, D.; Mosley, R. M.; Williams, R. R.; *Appl. Spectrosc*; 52; 1998; p. 1203
48. Ozdemir, D.; Williams, R. R.; *Appl. Spectrosc*; 53; 1999; p. 210
49. Wang, Y.; Veltkamp, D. J.; Kowalski, B. R.; *Anal. Chem*.; 63; 1991; p. 2750
50. Solomons, T. W. G.; Fryhle, C. B.; *Organic Chemistry* John Wiley&Sons, Inc.; 7th Edition; 2000; 828
51. Helminen J.; Leppämäki M.; Paatero E.; Minkkinen P.; "*Monitoring the kinetics of the ion-exchange resin catalyzed esterification of acetic acid with etanol using near-infrared spectroscopy with partial least squares model*"; *Chemo. And Intell. Lab*. Syst.; 44, 1998; pages 341-352
52. Altiokka, M. R.; Citak, A.; "*Kinetics Study of Esterification of Acetic Acid with isobutanol in the Presence of Amberlyst catalyst"*; *Applied Catalysis;* 6205; 2002; p 1-8
53. Ramalinga, K.; Vijalaykashmi, P, Kaimal. T. N.B.; *Tetrahedron Letters*; 43; 2002; p 9879-882

54. Ooi, T; Sugimoto, H; Doda, K; Maruoka, K; "*Esterification of carboxylic acids catalyzed by in situ generated tetraalkylammonium fluorides*" *Tetrahedron Letters*; 42; 2001;p 945-9248
55. Lilja,j.; Murzin, D. Yu; Salmi, T; Aumo, J; Mäki-arvela, P; Sundell, M.; *"Esterification of different acids over heterogeneous and homogeneous catalysts and correlation with Taft equation*"; *Journ. Of Mole. Catly. A: Chemical*; 3498; 2002; 1-9