Technical University of Crete, Greece
School of Electronic and Computer Engineering

Spectral deconvolution and concentration mapping in complex biochemical stains

Fani Abatzi

Thesis Committee:
Professor Costas Balas (Supervisor)
Professor Michalis Zervakis
Associate Professor Pagona-Noni Maravelaki

Chania, July 2014

Acknowledgements

First of all, I would like to thank George Epitropou for his invaluable help and cooperation, his time, advice and support throughout the research and implementation of this diploma thesis, without which I wouldn't have learned so much and I wouldn't have reached this point. In addition, I would like to thank professor Costas Balas
for the opportunity he gave me to deal with such an interesting topic, as well as both professors Michalis Zervakis and Noni-Pagona Maravelaki for their participation in the committee of my thesis and their comments and advice. I also feel very grateful for the help in providing the chemical solutions and the spectrophotometer needed for the spectra acquisition by the staff of the Analytical and Environmental Chemistry Laboratory of Technical University of Crete.

I would also like to thank my friends and especially Nikos Kofinas who has patiently been next to me through the writing of this thesis and helped me so much to cope with it, as well as my best friends Vaso Manikaki, Christos Rossos and my sister Olympia for the psychological support and uplift all of these months and all the members and fellow students of Optoelectronics lab for their help before the presentation. Last but not least, I would like to thank my parents for their financial support and their understanding throughout all the years of my studies without which my years at university would be so much more difficult.

Abstract

Spectral Imaging (SI) combines spectroscopy and imaging, enabling the acquisition of a stack of images at narrow spectral bands comprising the so-called spectral cube. A complete spectrum can be calculated for every image pixel from the multidimensional spatio-spectral space of the cube. This study aims at identifying the concentration of solvents in mixtures of multiple biochemical stains with overlapping spectral signatures. More specifically, a series of experiments has been undertaken via experimental design arrangements (full factorial, face-centered & half factorial) employing both spectrum acquisition by spectrophotometer and spectral imaging acquisition. Furthermore, an extensive number of algorithmic methods, based on Beer Lambert'As law generalization, including Classical Least Squares (CLS), Inverse Least Squares (ILS) with forward or backward selection, Principal Components Regression (PCR) and Partial Least Squares (PLS) has been implemented and applied to both simulated and experimental data. It was found that PLS can predict the concentrations in mixtures of two and three solvents with high accuracy on the datasets of spectral images. The combination of SI with concentration prediction algorithms can provide a valuable tool for quantitative assessment of the uptake of biological stains in histochemistry applications.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Chapter 1

Introduction

1.1 Thesis Contribution

This thesis studies the decomposition of optical spectra from biomedical stain mixtures for chemometrics and especially quantitative analysis. Its main contribution resides in the quantification,in other words, the estimation of the individual dissolved substances' concentrations and, thus, it provides a solution using four methods based on least squares regression (CLS, ILS, PCR, PLS). It also contributes a comparison among these four methods, highlighting their advantages and disadvantages in the concentration estimation process and which of them reciprocates better for our thesis' purposes. In order to achieve this, some error metrics, such as root mean squared errors and standard error of regression have been employed.

Furthermore, another contribution of this thesis is that an experimental design has been constructed and followed for the preparation, acquisition and processing of mixtures of two, three, four and five components. More specifically, this design can be used in datasets acquired with a spectrometer or in simulated data, as well as in hyperspectral images acquired with a microscope and its greatest advantage is that it built in order to incorporate the most information with the fewest inputs. Thus, it would be important to follow it for future experiments in Optoelectronics lab.

The system that has been created has great potential in process analysis, in vivo diagnostics, security-related detection systems and many other areas, especially biomedical engineering. This lies in the use of biomedical stains that aid the designation of biomarkers of diagnostic importance, especially in the field of immunohistochemistry.

1.2 Thesis Outline

Chapter 2 provides general information about the theoretical background of the problem, including spectroscopy and spectral imaging, but it also analyses more specific issues connected to our work, such as the Beer-Lambert Law, Chemometrics field, previous related work and applications.

Chapter 3 is concerned with the problem specifications, such as the hardware devices used for the experimental process and their features. Furthermore, in this chapter we describe the biomedical dyes used in the experimental stage,the experimental design followed, the experimental data acquired from the above devices (pure solutions or mixtures) and the simulated data used. The chapter is also concernced with the inevitable and critical to the processing stage errors that occur during data acquisition.

In Chapter 4 we discuss all the methods applied in this thesis. First, we make a brief reference to the calibration process and its significance, beginning with univariate and continuing to multiple and multivariate calibration. Subsequently, we analyze the four calibration/regression methods that result in the mixture concentration prediction. In the last part of the chapter we mention the concentration and overall regression errors calculated to validate the procedure.

Chapter 5 is the most critical one, because it presents the results obtained after the methods and their algorithms are applied on the various experimental (spectral, or image) data or the simulated data. A plethora of tables and figures are provided to cover the topic.

Finally, in Chapter 6 we summarize the conclusions we were guided towards and the possible future research directions on the problem.

Chapter 2

Theoretical Background

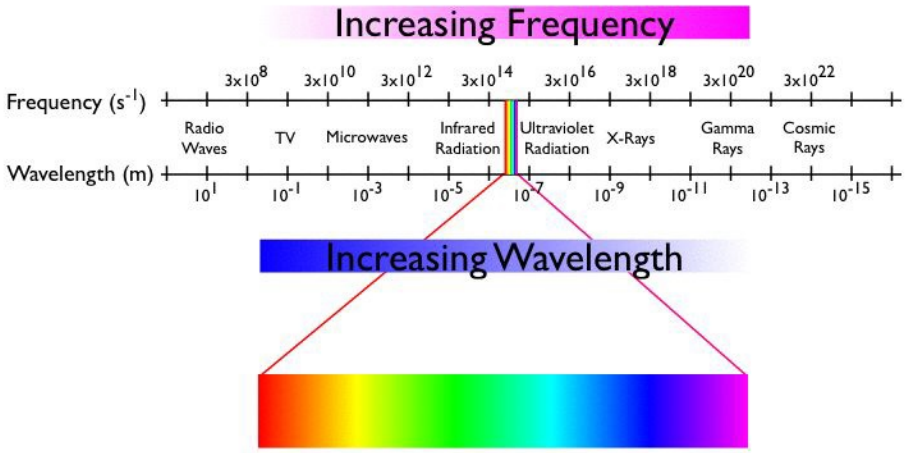
2.1 Electromagnetic Radiation

Electromagnetic radiation is a transverse wave, consisting of an oscillating electric field E and an oscillating magnetic field M. The two fields are perpendicular to each other as well as to the propagation direction of the wave. The wave is described by a wavelength λ, which corresponds to the physical length of a full oscillation and a frequency v, which corresponds to the number of oscillations per second. Electromagnetic radiation exhibits properties of both waves and particles. This wave-particle duality was initially understood by Albert Einsten who expressed it as a continuous flow of particles or wave energy packages, also known as photons [1].
One photon carries energy equal to:

$$E_{photon} = hv = \frac{hc}{\lambda} \tag{2.1}$$

where h is Planck's constant (6.6261×10^{-34} Js), c is the speed of light, v is the frequency and λ is the wavelength of the radiation.

For reasons of convenience, the electromagnetic spectrum is divided into a number of spectral bands, each one of which interacts with matter in a different way. The visible spectrum constitutes a small part of the total radiation spectrum [2]. Most of the radiation that surrounds us cannot be seen, but can be detected by dedicated sensing instruments. The electromagnetic spectrum ranges from very short wavelengths (including gamma and x-rays) to very long wavelengths (including microwaves and broadcast radio waves) as shown in figure 2.1.



Spectral Band	**Wavelength**
γ rays	¡ 0.03 nm
X-rays	0.03-10 nm
Ultraviolet	10-400 nm
Visible light	400-800 nm
Near Infra-red	0.8-2.5 μm
Mid Infra-red	2.5-15 μm
Far Infra-red	15-200 μm
Microwaves	0.2-7 mm
Radio waves	100-10000 m

Figure 2.1: Spectral bands of Electromagnetic Radiation

2.2 Spectroscopy/Spectrometry

Spectroscopy is the study of the interaction between electromagnetic radiation and matter as a function of wavelength (λ). In fact, historically, spectroscopy referred to the use of visible light dispersed according to its wavelength, e.g. by a prism. Later, the concept was expanded greatly to comprise any measurement of a quantity as function of either wavelength or frequency. Thus, it can also refer to interactions with particle radiation or to a response to an alternating field or varying frequency (v). A further extension of the

scope of the definition added energy (E) as a variable, once the very close relationship $E = hv$ for photons was realized.

Spectrometry is the spectroscopic technique used to assess the concentration or amount of a given species. In those cases, the instrument that performs such measurements is a spectrometer or spectrograph.

Spectroscopy/spectrometry is often used in physical and analytical chemistry for the identification of substances through the spectrum emitted from or absorbed by them. Spectroscopy/spectrometry is also heavily used in astronomy and remote sensing. Most large telescopes have spectrometers, which are used either to measure the chemical composition and physical properties of astronomical objects or to measure their velocities from the Doppler shift of their spectral lines [3].

2.2.1 Analytical Spectroscopic Methods

The type of spectroscopy depends on the physical quantity measured. Normally, the quantity that is measured is intensity, either of energy absorbed or produced. Most spectroscopic methods are differentiated as either atomic or molecular based on whether or not they apply to atoms or molecules. Along with that distinction, analytical methods of spectroscopy/spectrometry can be classified on the nature of their interaction as follows [4] :

Classic: Related to mass. These are either *Gravimetric* (Weighing scales) or *Volumetric* (calibrated glassware).

Instrumental: Related to energy. These are either *Optical* (electromagnetic radiation-sample interaction - absorption, emission, scattering), *Electrochemical* (end point in volumetric measurements with electrical devices) or *Specialized* (chromatographic, immunochemical).

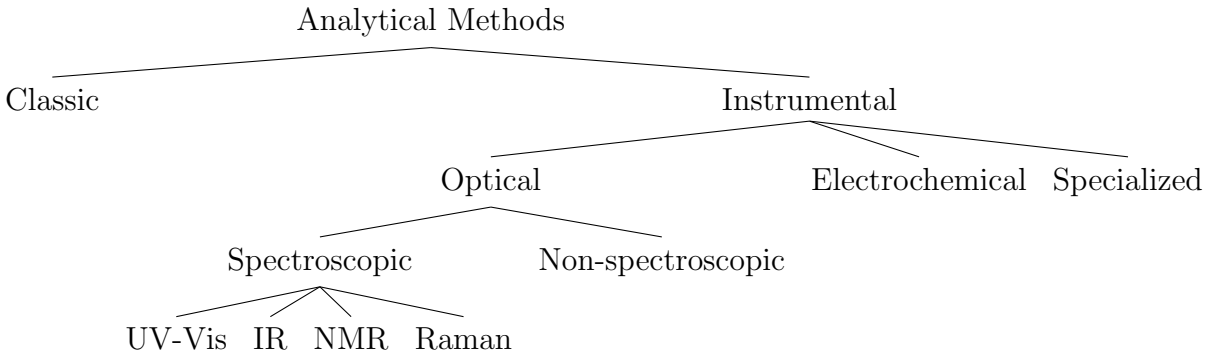
Optical methods are separated into Spectroscopic Techniques and Non-Spectroscopic Techniques. *Spectroscopic* techniques are based on the ability of various substances to emit or interact with radiation of typical frequencies and on spectral measurements (wavelength, power-intensity of radiation). *Non-spectroscopic* techniques are based on the interaction of electromagnetic radiation and matter, which entails

change in the direction or physical properties of radiation. It does not involve spectral measurements.

There are four categories of Instrumental Optical Spectroscopic Techniques:

1. **UV-Visible (UV-Vis)**. The absorption of visible or ultraviolet radiation causes electron transitions of outer orbitals.
2. **Infra Red (IR)**. The absorption of infra red radiation causes vibrational, deformational and rotational stimulations to the bonds of molecules.
3. **Nuclear Magnetic Resonance (NMR)**. Changes in nuclei's energy.
4. **Raman**. Scattering radiation.

The following diagram summarizes the above distinction between the various methods of spectroscopy:



In this diploma thesis we have dealt with instrumental, optical, visible spectroscopic methods, and especially with absorption spectroscopy. Absorption spectroscopy is the study of light absorbed by molecules. In it, white light is caused to pass through a sample and then through a device (such as a prism) that breaks the light up into a spectrum. When such light is passed through a sample, under the right conditions, the electrons of the sample will absorb those wavelengths of light that can change them to other levels. Thus, the light coming out of the prism will be missing those wavelengths corresponding to the allowed energy levels of the electrons in the sample. We will see a spectrum with black lines where the absorbed light would have been if it had not been removed by the sample [5].

2.2.2 Absorption Spectra

An absorption spectrum is the representation of absorption, permeability, or intensity of the radiation as a function of wavelength [4]. In this diploma thesis, in order to estimate the concentrations of the spectral mixtures, we dealt with absorption spectroscopy and we conducted experiments measuring absorption spectra at different wavelengths in the visible spectrum area when the device allowed us, or measuring transmission spectra and converting them into absorption spectra.

2.3 Spectral Imaging

Spectral imaging (SI) is a branch of spectroscopy which combines spectroscopy with imaging. Hyperspectral imaging is part of spectral imaging and the difference with multi-spectral imaging lies in the number of spectral bands under investigation. Spectral imaging can be considered as the equivalent of color photography, but each pixel needs to acquire many bands of light intensity data from the spectrum, instead of just the three bands of the RGB color model displayed in figure 2.2.

2.3.1 Hyperspectral Cube

The Spectral Imaging (SI) systems acquire a 3-dimensional dataset of spectral and spatial information, known as spectral or hyperspectral cube. The hyperspectral cube comprises a set of images of a size of $x \times y$ pixels acquired at N different wavelengths. The 3-D spectral and spatial character of the data (intensity I at 3 dimensions x,y and λ) and a more analytical representation of the hyperspectral cube are illustrated in figure 2.3.

2.3.2 Spectral Signatures

Spectral signatures are a collection of pixels representing categories of materials with different structural or chemical composition [6]. More specifically, for any given material, the amount of solar radiation that reflects, absorbs, or transmits varies with wavelength. This important property of matter makes it possible to uniquely identify different physical or chemical substances or classes and separate them by their spectral signatures, as shown in figure 2.4. For example, at some wavelengths, sand reflects more energy

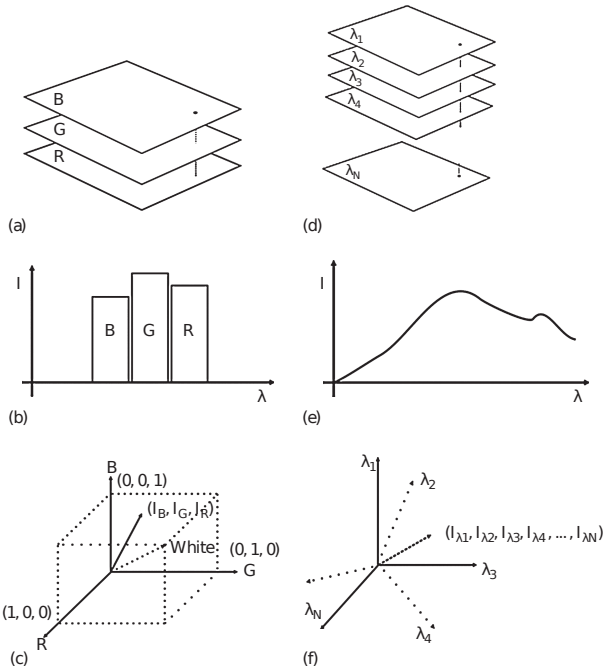


Figure 2.2: (Image data capturing and representation in color (a-c) and spectral (d-f) cameras.

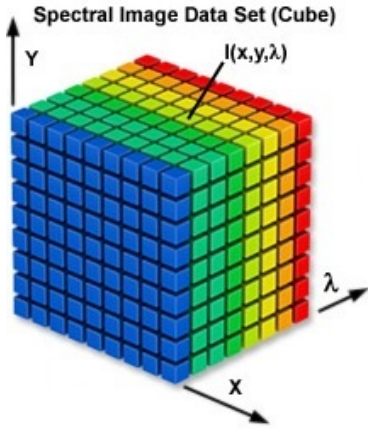


Figure 2.3: (Hyper)spectral Cube representing the spectral (λ) and spatial (x,y) character of the data

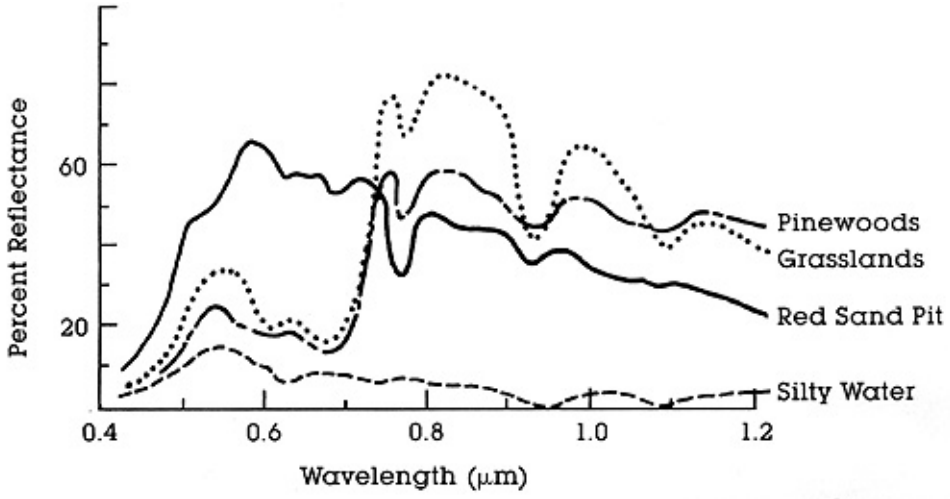


Figure 2.4: Spectral signatures of different physical substances.

than green vegetation but at other wavelengths it absorbs more (reflects less) than does the vegetation. In principle, we can recognize various kinds of surface materials and distinguish them from each other by these differences in reflectance [7].

2.4 Chemometrics and quantitative analysis

Quantitative analysis as used in chemistry, chemical engineering and physics is the determination of the absolute or relative abundance (often expressed as a concentration) of one, several or all particular substance(s) present in a sample.

Chemometrics is the use of statistical and mathematical techniques in order to extract information from chemical systems using multivariate statistics and applied mathematics to address problems in chemistry, biochemistry, medicine, biology and chemical engineering. The field is generally recognized to have emerged in the 1970s as computers became increasingly exploited for scientific investigation. The term chemometrics was introduced by Svante Wold in a grant application 1971 and the International Chemometrics Society was formed shortly thereafter by Svante Wold and Bruce Kowalski, two pioneers in the field. Wold was a professor of organic chemistry at Umea University, Sweden, and Kowalski was a professor of analytical chemistry at University of Washington, Seattle [8] [9] [10] [11]. Chemometrics is a relatively new branch of science and often

involves using linear algebra methods to make qualitative or quantitative measurements of chemical data.

In this diploma thesis, we have dealt with quantitative analysis and the field of chemometrics in order to estimate the absolute abundances (concentrations) of chemical solutions. The concentrations of such solutions are measured in Molarity (units: $\frac{mol}{L}$ or M), which represents the number of moles of a dissolved substance per litre of solution [10].

Advances in computational hardware and the development and application of chemometrics were necessary to transform the field of spectroscopy into a useful analytical tool. This same combination of hardware and software tools must be further exploited to analyze and extract the information contained in hyperspectral images.

2.5 Beer-Lambert Law

One of the keys to quantitative analysis in any scientific field is the assumption that the amounts(concentrations) of the constituents of interest in the samples are somehow related to the data from a measurement technique used to analyze them. The ultimate goal is to create a calibration equation (or series of equations) which, when applied to data of "unknown" samples measured in the same manner, will accurately predict the quantities of the constituents of interest. In order to calculate these equations, a set of "standard" samples are made which reflect the composition of the "unknowns" as closely as possible. The standards are then measured by an instrument. Together, this collection of known data (the composition of each standard) and the measured data from the instrument form what is known as a **training set** or **calibration set**. The calibration equations that describe the relationships between these two sets of information are calculated from this data. The exact equation or set of equations that make up the calibration is also known as a **model**. Thus, this process is often called, "solving the calibration model." One advantage of using spectroscopy as a measurement technique is that the **Beer-Lambert Law** (also known as Beer's Law) defines a simple linear relationship between the spectrum and the composition of a sample. This law, which should be familiar to all spectroscopists, forms the basis of nearly all other chemometrics methods for spectroscopic data. Simply stated, the law claims that when a sample is placed in the beam of a spectrometer, there is a direct and linear relationship between the amount (concentration) of its constituent(s) and the amount of energy it absorbs.

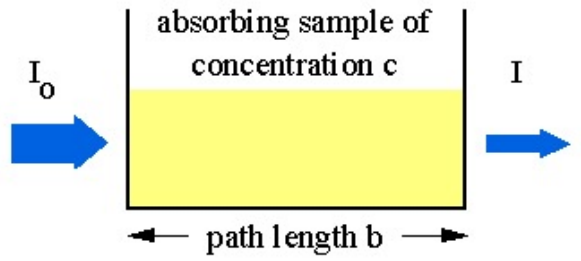


Figure 2.5: A beam of incident light of intensity I_0 passes through absorbing sample of concentration c and of a pathlength d and exits with an intensity of I.

In mathematical terms [10]:

$$\mathbf{A} = log(\frac{I_0}{I}) = -log(T) = log(1/T) = \varepsilon b\mathbf{C} \tag{2.2}$$

where
A= Absorbance of a sample as a function of wavelength. It's a rational number.
I_0= Intensity of the incident light.
I = Intensity of the transmitted light (see figure 2.5).
T = Transmittance ($\frac{I}{I_0}$), usually expressed as a percentage %T.
C = the concentration of the sample's solution measured in $\frac{mol}{L}$ or M (molarity).
b = the pathlength that the light beam has travelled inside the sample(in cm).
ε = the molar absorptivity of the solution, which is a constant number also proportional to the respective absorbance wavelengths.
When referring to mixtures instead of single substance solutions Absorbance, molar absorptivity and concentration are matrices of the following sizes:
A is a $d \times n$ matrix, where n is the number of samples measured and d the number of wavelengths.
E is a $d \times p$ matrix, where p is the number of chemical components in a sample and
C is a $p \times n$ matrix. In chapter 4.2 we will see how Beer-Lambert Law is associated with the methods used for our problem solving.

The Beer-Lambert law is valid under the following conditions:

- The solutions are not dense. The preferable absorbance should range between 0.1 and 1.
- The only mechanism for the interaction between a dissolved substance and radiation is absorption.
- The incident radiation to a sample is monochromatic.
- The sample is in a cuvette (quartz glass in our case) with a uniform intersection.
- The absorbing molecules act individually and despite their number and kind ($A_{total} = A_1 + A_2 + ... + A_n$, where n is the number of dissolved substances in the mixture) [4].

These limitations appeared to be useful during the construction of the experimental design for this thesis(3.2.2).

2.6 Related Work-Applications

Spectral Imaging in general is a widely known "tool" with a large number of applications in various fields. Such fields are: analytical chemistry (qualitative and quantitative analysis of a chemical solution), biochemistry, in vitro analysis, cartography and material classification through geology and remote sensing, industrial applications (qualitative analysis of products) and medicine, especially optical biopsy as a non-destructive, non-invasive diagnostic method [1].

More specifically, going backwards, the problem of spectral decomposition originates in the field of *remote sensing* [12]. This field has been motivated by a desire to extract increasingly detailed information about the material properties of pixels in a scene for both civilian and military applications. While multispectral sensing has largely succeeded at classifying whole pixels, further analysis of the *constituent substances* that comprise a pixel is limited by a relatively low number of spectral measurements. The recognition that pixels of interest are frequently a combination of numerous disparate components has introduced a need to quantitatively decompose, or unmix these mixtures. Collecting data in hundreds of spectral bands, hyperspectral sensors have demonstrated the capability of performing spectral unmixing. In hyperspectral imagery, mixed pixels are a mixture of more than one distinct substance, and they exist for one of two reasons:

1. If the spatial resolution of a sensor is low enough that disparate materials can jointly occupy a single pixel, the resulting spectral measurement will be some composite of the individual spectra.

2. Second, mixed pixels can result when distinct materials are combined into a homogeneous mixture. This circumstance can occur independent of the spatial resolution of the sensor.

Analytical models for the mixing of disparate materials provide the foundation for developing techniques to recover estimates of the constituent substance spectra and their proportions from mixed pixels. The basic premise of mixture modelling is that within a given scene, the surface is dominated by a small number of distinct materials that have relatively constant spectral properties. These distinct substances (e.g., water, grass, mineral types) are called *endmembers*, and the fractions in which they appear in a mixed pixel are called fractional *abundances*. If the total surface area is considered to be divided proportionally according to the fractional abundances of the endmembers, then the reflected radiation will convey the characteristics of the associated media with the same proportions. In this sense, there exists a **linear relationship** between the fractional abundance of the substances comprising the area being imaged and the spectra in the reflected radiation. Hence, it is called the *linear mixing model(LMM)*. When M endmembers exist, each having L wavelengths, the LMM is expressed as:

$$\mathbf{x} = \sum_{i=1}^{M} \mathbf{a}_i \mathbf{s}_i = \mathbf{S}\mathbf{a} + \mathbf{w} \tag{2.3}$$

where $\mathbf{x}$ is the $L \times 1$ received pixel spectrum vector, $\mathbf{S}$ is the $L \times M$ matrix whose columns are the $L \times 1$ endmembers, $\mathbf{s}_i$, $i = (1, \ldots, M)$, $\mathbf{a}$ is the $M \times 1$ fractional abundance vector whose entries are $\mathbf{a}_i$, $i = 1, \ldots, M$ $\mathbf{w}$ is the $L \times 1$ additive observation noise vector. When N pixels are considered, block notation is utilized, such that $\mathbf{X} = \mathbf{S}\mathbf{A} + \mathbf{W}$, $\mathbf{X} = [x(1)...x(N)]$, $\mathbf{A} = [a(1)...a(N)]$, and $\mathbf{W} = [w(1)...w(N)]$.

The above linear model is precisely proportional to the Beer's Law model seen in the previous section, although instead of reflectance spectra(LMM) there are absorption spectra(Beer's Law), instead of endmembers we refer to absorptivities, and instead of fractional abundances we have absolute concentrations.

There are many other applications of quantitative analysis, except for remote sensing. In other words, hyperspectral imaging techniques provide an attractive solution for the microscopic and macroscopic analysis of biological, agricultural and environmental materials. The various applications outlined show the benefits of these techniques for sample characterization and chemical species distribution, for fruit and kernels and for food and feed mixtures. Furthermore, the application of optical photography, MRI (magnetic resonance imaging), and FTIR (Fourier transform infrared) imaging in the study of solvent diffusion in polymers, polymer dissolution, polymer crystallisation, and drug release from pharmaceutical tables have been reviewed in previous art. Specific applications of multivariate techniques, such as PLS and CLS, have been used to analyse imaging datasets obtained by ATR-FTIR (attenuated total reflectance-Fourier transform infrared) imaging of tablet dissolution [13]. PCA has mostly been used and applied to perform different statistical measurements or to perform quantitative analysis in combination with other studies. This shows the potential and possibilities of using PCA as a method that without kinetic assumptions can analyse the dynamic PET images/data aiming to explain kinetic behaviour of the administered tracer in different regions of the brain by observing its variation within the time sequence when it is accurately applied.

Considering forgone applications of chemometrics in tissues, we could mention the multispectral imaging application for the determination of astaxanthin concentration in salmonoids, especially rainbow trout fillets [14]. The contribution of this research,though, was not of biomedical significance but rather for quality evaluation of these products in order to fulfill customers' needs.

Another field of biomedical engineering where quantitative analysis has been applied is quantitative pathology [15]. In this paper, a novel spectral microscope system was presented together with a method for the quantitative assessment of the uptake by histologic samples of stains used in pathology to label tissue features of diagnostic importance.

From the bibliography ([44], [45] and [46]) it seems that one of the most relevant fields of biomedical engineering to quantitative analysis is immunohistochemistry. *Immunochistochemistry* it's a process where antibodies labeled/stained with a fluorescent or other stain can be bound to antigens (or biomarkers) which allow the detection and isolation of a particular cell type and, thus, lead to the detection of abnormal cells in tissues. Through this process and with the aid of deconvolution (or unmixing) of im-

munochistochemical stains, cells with a particular phenotype which tends to contribute to metastasis can be revealed([45]).

Last but not least, there has previously been made an investigation of lung cancer biomarkers by hyphenated separation techniques and chemometrics. In that work, quantitative analysis of Volatile organic compounds(VOCs) in the headspace of lung tissues revealed that cancer cells released higher concentrations of ethanol, acetone, carbon disulfide, dimethyl sulfide, 1-propanol, 2-propanol, 2-butanone and 2-pentanone than healthy tissues. The increase of concentration of the substances was observed in the breath of patients with lung cancer in comparison to breath from healthy non-smoking volunteers. In overall, this concluded to the fact that detection of lung cancer is possible by volatile biomarkers analysis in breath [16], with the use of solid phase microextraction and gas chromatography mass spectrometry. However, in this diploma thesis we are going to develop different techniques from this particular one, and we are going to concentrate upon absorption or transmission spectroscopy.

Chapter 3

Problem Specifications

In this diploma thesis, we have encountered the problem of the decomposition of optical mixture spectra of two, three (usually) or more chemical substances (later). In order to handle the problem stated, we transferred the problem of spectral unmixing or spectral decomposition from the domain of quantitative analysis in remote sensing to the domain of quantitative analysis in chemometrics, and then from chemometrics to hyperspectral imaging.

More specifically, the initial challenge was to create an experimental design and a whole series of experiments in order to gather the necessary mixture samples from which the individual concentration spectra would be identified. Subsequently, we encountered the problem of finding and applying the most appropriate algorithmic methods that could identify the concentrations of the individual substances (measured in Molarity, M), based on an equation of known absorbance or transmission spectra and known or unknown concentrations. As seen before, the mechanism that connects the absorbance of a sample with its concentrations is the Beer-Lambert Law, so the object was to find the algorithms that based on this law-equation would perform better for our problem . Thus, the need for creation of a calibration regression model occurred and in this direction we moved in order to implement the most appropriate algorithms. Finally, we encountered the need for some validation "tools" in order to decide which of these algorithms performed better.

In the following sections of this chapter we are going to develop the experimental set-up and design, the devices needed for the experiments and the characteristics of the chemical substances used and their exited individual and mixture spectra, after measured with these devices.

3.1 Acquisition System

For the purposes of this diploma thesis, two data acquisition systems have been used, a UV-VIS Spectrophotometer and a Microscope. The first was used for the initial experiments on 9-sample absorbance spectra datasets of two or three substance mixtures, whereas the later was used for the acquisition of hyperspectral images (hyperspectral data cubes 2.3.1).

3.1.1 Spectrophotometer configuration

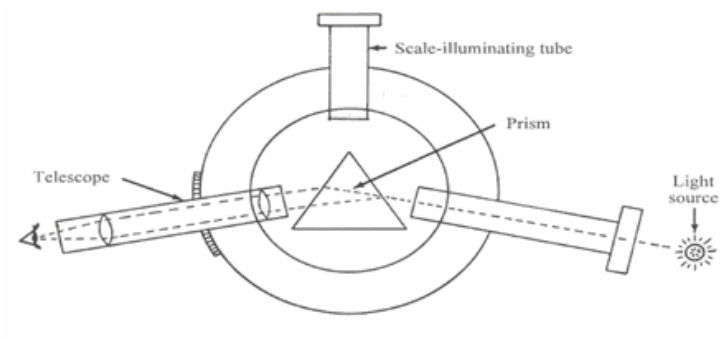


Figure 3.1: Diagram of a spectroscope

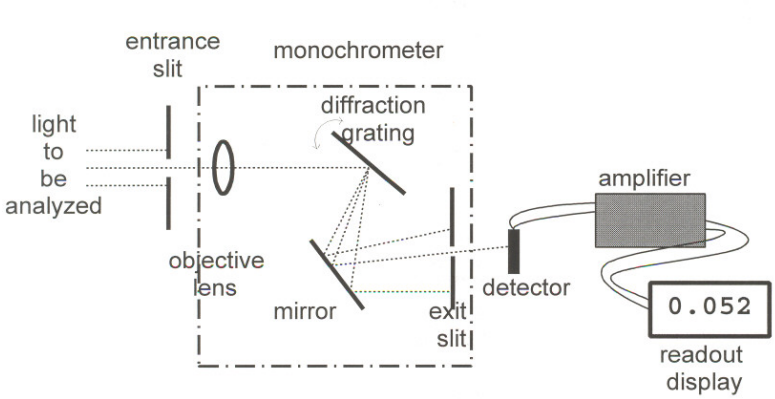


Figure 3.2: A schematic representation of a simple spectrometer

There is a large variety of instruments used to perform spectroscopy. They differ

greatly in the kinds of information they provide. What they all have in common is the ability to break light up into its component wavelengths. Such instruments are *spectroscopes*, *spectrometers* and *spectrophotometers*.

A *spectroscope* is the simplest of spectroscopic instruments. Its function is to take light from any source and spread it into a spectrum for viewing with the unaided eye.Figure 3.1 is a diagram of a simple spectroscope. The light from the source passes through the slit and into the prism where it is spread into a spectrum. The telescope is used to focus on the light coming out of the prism. The third arm contains a wavelength scale that can be superimposed over the spectrum by shining a white light into it.Spectroscopes are useful for determining what wavelengths of light are present in a light source, but they are not very useful for determining the relative amounts of light at different wavelengths. Spectroscopes are most commonly used for qualitative emission spectroscopy.

A *spectrometer* is a spectroscope that has some sort of meter attached that can measure the amount of light (number of photons) at specific wavelengths.Thus, it is designed to provide a numerical measure of the amount of light emitted or absorbed at a particular wavelength. It is constructed so that the wavelength can be varied by the operator and the amount of radiation absorbed or transmitted by the sample determined for each wavelength. In this way it is possible to learn which wavelengths of radiation are present and in what relative amounts. Spectrometers are common in astronomy where they are used to evaluate the light collected by telescopes. They are the only source of information we have about the chemical composition of the universe outside our own solar system. Figure 3.2is a schematic representation of a simple spectrometer.Light enters the spectrometer via the entrance slit and then passes through several parts: an objective lens, a grating, and an exit slit. This combination of parts functions as a *monochromator*, a device which selects only one color (actually, a narrow band of wavelengths) from all of the wavelengths/colors present in the source. A particular wavelength is selected, using the wavelength control, by adjusting the angle of the grating. This works because different wavelengths of light reflect off the grating at different angles. The net result is the separation of white light into a "rainbow" much like light transmitted through a prism of glass. The selected wavelength is at the center of the narrow band of wavelengths passing through the slit. The light then strikes a detector that generates a voltage in proportion to the intensity of the light hitting it. That voltage is then used to drive a read-out device that is designed to provide data in a useful fashion such as intensity.

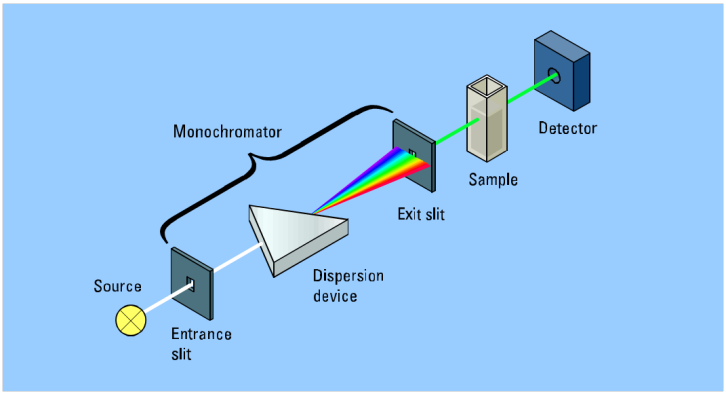


Figure 3.3: A schematic representation of a conventional spectrophotometer

Since spectrometers measure the amount of light entering the instrument, they are most often used for emission spectroscopy. In order to perform absorption spectroscopy, a light source of known intensity is required. An instrument that includes such a light source is known as a *spectrophotometer*. It is constructed so that the sample to be studied can be irradiated with light of known wavelength and intensity. The wavelength can be varied and the amount of radiation absorbed or transmitted by the sample determined for each wavelength. From this information, an absorption spectrum for a species can be obtained and used for both qualitative and quantitative determinations. Figure 3.3 is a schematic representation of a conventional spectrophotometer.

Usually a spectrophotometer is the best, but spectrometers are cheaper and faster in data acquisition. However, the spectrophotometer provided to us by the Analytical and Environmental Chemistry Laboratory had a stable set-up isolated from environmental noise, and also with a better resolution. For all the above reasons, the use of a **spectrophotometer** seemed the most appropriate for our first series of experiments. The type of Ultraviolet - Visible (UV-Vis.) spectrophotometer that we used was a model of the brand Cary, namely *Cary* $1E$, *Varian*, as it is illustrated in figure 3.4 and figure 3.5. The device in general terms follows the basic configuration of a spectrophotometer analyzed previously. For this instrument, the light source used inside is a Xenon lamp. There is also a dual cuvette plate inside the instrument. Of the two cuvette holders on it, the holder furthest inside the Cary is for a **reference solution**, which remains there throughout the experiment; the holder closest to the front of the Cary is for **establish-**

Figure 3.4: Representation of Cary 1E Varian UV-Vis. spectrophotometer

ing the blank baseline at the beginning of the experiments and for **the sample itself** later on (figure 3.6). Further details about the inside of the device and the software used for the data acquisition can be found in [17].

Experimental set-up

During the experimental procedure some steps are followed:

1. The essential samples for the experiments are prepared. The samples are put in glass or plastic cuvettes of $4ml$ capacity with the aid of a glass pipette. One important characteristic of such cuvettes is that regular glass is opaque below $350nm$. Hence, if measurements are to be made below $350nm$, they must be made of quartz glass [5].

2. The Cary Uv-Vis. spectrophotometer is put into function. Two cuvettes with water are placed in the baseline and reference cuvette holders, respectively.

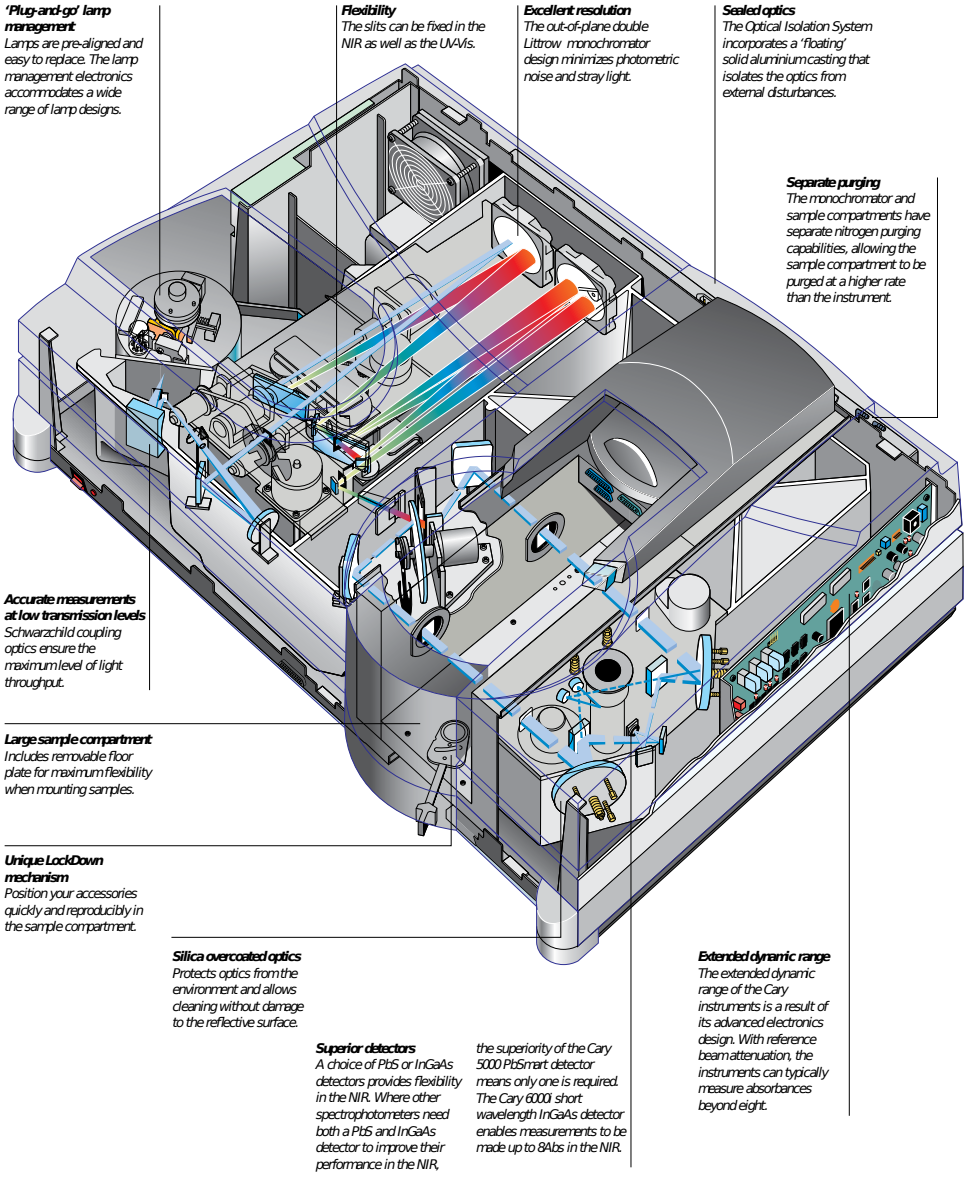


Figure 3.5: A schematic of the guts and the path the light takes through the instrument

3. The software program($Scan$)connected to the spectrophotometer gets started. Some regulations have to be made, according to the selection of Absorption or Transmission spectra gathering, the wavelengths' range, the intervals between wavelengths (for us $0.1nm$), average scanning time(s) and scanning rate(nm/min). Subsequently, there has to be made a baseline correction in order to correct for any

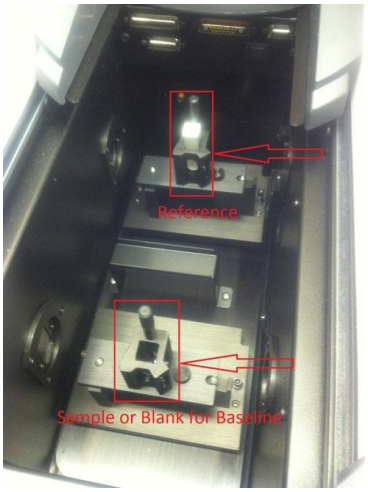


Figure 3.6: The cuvette holders of the spectrophotometer

absorbance that comes from the sample holder, substrate, solvent, etc., that is any additional background noise that causes unnecessary peaks.

4. At this point, we have the arrangement ready for the beginning of the experimental procedure. After establishing the blank baseline, we keep the cuvette in the reference holder, we remove the other and we begin putting our samples in order to obtain the absorbance spectra we need. Finally, we can save the acquired spectra in an file for further processing.

3.1.2 Microscope Configuration

Figure 3.7 illustrates a general overview of the components of the system used for hyperspectral images acquisition [18].

As the "powering" component of the tunable light source we used a **halogen lamp white light source**. In fact, it is a NAVITAR 1-60563 220 Volt fiber optic power supply with a 150 Watts EKE halogen lamp and a 0.720" fiber receptacle.

The white light emitted by the halogen lamp is filtered by a **linear filter** which is mounted on a ramp and moved by a NANOTEC - ST4118L1804 **bipolar stepper motor**. The filter selected is a Linear Variable Filter (LVF); Schott, VERIL BL200(figure 3.8). A LVF is a bandpass filter where the coating has been intention-

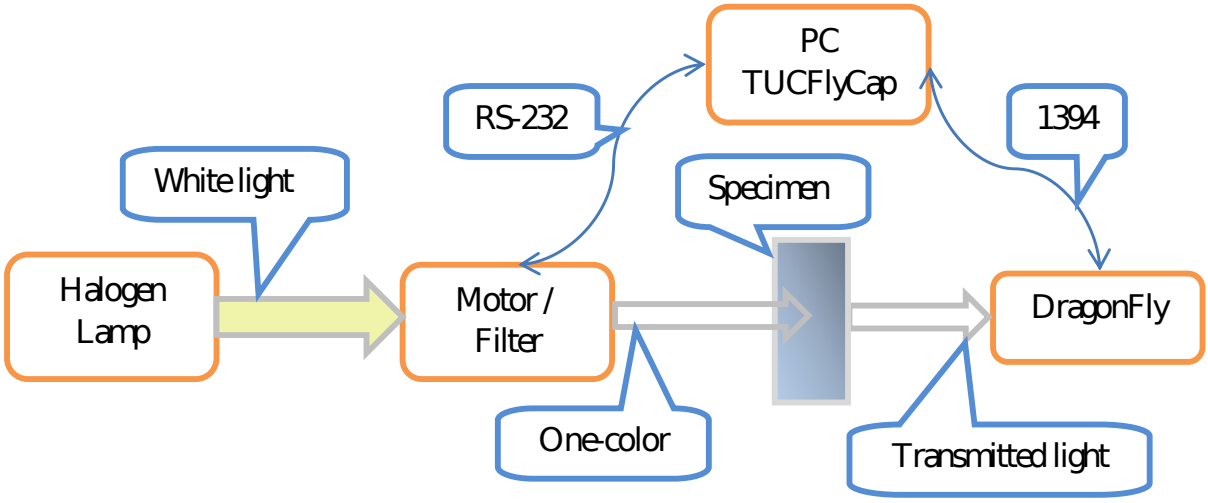


Figure 3.7: System Overview for the acquisition of hyperspectral images

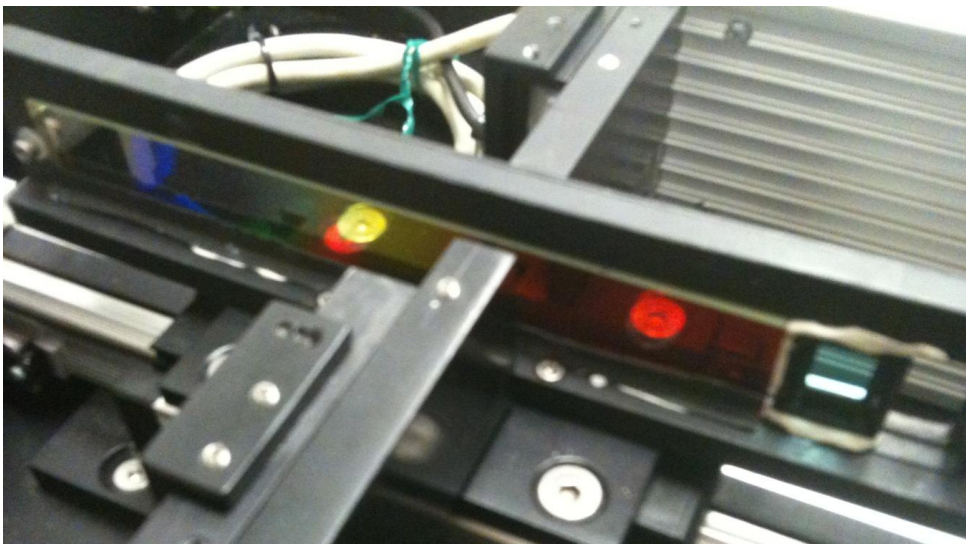


Figure 3.8: Linear variable bandpass filter, Schott - VERIL BL200

ally wedged in one direction. This wedge causes the center wavelength (nm) of the filter to shift linearly across the length of the filter. Adjusting the filter orientation allows a specific wavelength to be selected.

The filtered light, which ideally is *monochromatic* or with narrow bandwidth, passes through a **specimen of a microscope**. We use an upright microscope for transmission

Figure 3.9: OLYMPUS - BX51 for transmission/fluorescence microscopy

and fluorescence microscopy; OLYMPUS - BX51, which is depicted in Figure 3.9. This specific microscope has the ability of operating in two modes. The first is the common *transmission* microscopy where the light (generated from a halogen lamp located at the lower-backside of the microscope) is transmitted through the specimen and then it is depicted at PC monitor using a digital camera and/or seen by from the ocular (eyepiece) lens. The second mode is the *fluorescence* (or epifluorescence) mode, where the light is inserted using a "special" lamp (arc lamp; mercury or xenon lamp) by the lamphouse located at upper-backside of the microscope.

The transmitted light is captured and recorded as spectral data from a **spectral imager/camera**. The spectral camera is what eventually combines spectroscopy and imaging into spectral imaging.In this project we used POINT GREY RESEARCH (PGR) - DragonFly2-13S2C-CS. This device is a full featured IEEE 1394a (FireWire) digital spectral camera. Its CCD sensors can acquire both Black & White (BW/gray=-scale) and color images. Both the filter movement and the camera settings are controlled via

the same software using RS-232 communication for the filter and IEEE 1394 (FireWire) for the camera again as presented in the System Overview figure.

3.2 Experimental Data Description

3.2.1 Biomarkers and Biomedical Stains

A *biomarker*, or biological marker [19], is in general a substance used as an indicator of a biological state.It is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Especially a **Cancer Biomarker** refers to a substance or process that is indicative of the presence of cancer in the body. A Cancer Biomarker may be a molecule secreted by a tumor or a specific response of the body to the presence of cancer. Biomarkers are used in many scientific fields. Genetic, epigenetic, proteomic, glycomic, and imaging biomarkers can be used for cancer diagnosis, prognosis, and epidemiology. In many areas of medicine, biomarkers are limited to proteins identifiable or measurable in the blood or urine. In cancer research and medicine, biomarkers are used in three primary ways: [20]

1. To help diagnose conditions, as in the case of identifying early stage cancers (Diagnostic)

2. To forecast how aggressive a condition is, as in the case of determining a patient's ability to fare in the absence of treatment (Prognostic)

3. To predict how well a patient will respond to treatment (Predictive)

More specifically, in medicine, a biomarker can be a substance that is introduced into an organism as a means to examine organ function or other aspects of health. It can also be a substance whose detection indicates a particular disease state, for example, the presence of an antibody may indicate an infection.
It can also be a substance whose detection indicates a particular disease state, for example, the presence of an antibody may indicate an infection. More specifically, a biomarker indicates a change in expression or state of a protein that correlates with the risk or progression of a disease, or with the susceptibility of the disease to a given treatment.

A biomarker can also be used to indicate exposure to various environmental substances in epidemiology and toxicology. In these cases, the biomarker may be the external substance itself (e.g. asbestos particles or NNK (Nicotine-derived nitrosamine ketone) from tobacco), or a variant of the external substance processed by the body (a metabolite).
In cell biology, a biomarker is a molecule that allows for the detection and isolation of a particular cell type (for example, the protein Oct-4 is used as a biomarker to identify embryonic stem cells).
In genetics, a biomarker (identified as genetic marker) is a DNA sequence that causes disease or is associated with susceptibility to disease.
Biomarkers of diagnostic importance can be highlighted with the use of *biomedical stains* or *dyes*, and that is the challenge we are asked to confront. For this thesis, we performed a series of experiments with solutions of substances used as biomedical dyes or stains. The difference between a stain and a dye is that stains are temporary whereas dyes are permanent and can be removed after cell wall destruction. A dye is a coloring agent used for general purposes and a stain is used for any biological specimal staining. As stated in the introduction of this thesis, the use of biomedical stains aid the designation of biomarkers of diagnostic importance. The origin of these dyes or stains are solid substances dissolved in deionized water so that they result in chemical solutions. Such substances used are the following:

- **Methyl Orange** is a pH indicator frequently used in titrations (also known as volumetric analysis) because of its clear and distinct color change. Unlike a universal indicator, methyl orange does not have a full spectrum of color change, but has a sharper end point. Methyl orange has mutagenic properties [21].In a solution becoming less acidic, methyl orange moves from red to orange and finally to yellow with the reverse occurring for a solution increasing in acidity.

- **Copper(II) sulfate**, also known as cupric sulphate or copper sulphate, is the chemical compound with the chemical formula $\mathbf{CuSO_4}$. The pentahydrate form ($CuSO_4.5H_2O$), the most commonly encountered salt, is bright blue. It can be used as a herbicide,a fungicide or pesticide, but also has applications in medicine, art, etching and vegetable dyeing [22], [23], [24], [25].

- **Malachite Green** is an organic compound that is used as a dyestuff and has emerged as a controversial agent in aquaculture. Malachite green is traditionally used as a dye for materials such as silk, leather, and paper. Although called malachite green, the compound is not related to the mineral malachite - the name just comes from the similarity of color. MG is active against the oomycete Saprolegnia, which infects fish eggs in commercial aquaculture, and other fungi. Furthermore, MG is also used as a parasiticide and antibacterial. [26] It is a very popular treatment against ichthyophthirius in freshwater aquaria. The principal metabolite, LMG, is found in fish treated with malachite green, and this finding is the basis of controversy and government regulation.

- **Fast Green** FCF, also called Food green 3, FD& C Green No. 3, Green 1724, Solid Green FCF, and C.I. 42053, is a sea green triarylmethane food dye (intensive color dyes used as pH indicators). It is used as a quantitative stain for histones at alkaline pH after acid extraction of DNA. It is also used as a protein stain in electrophoresis. Its absorption maximum is at 625 nm. Fast Green FCF is poorly absorbed by the intestines. Its use as a food dye is prohibited in European Union and some other countries. It can be used for tinned green peas and other vegetables, jellies, sauces, fish, desserts, and dry bakery mixes at level of up to 100 $\frac{mg}{kg}$. In the United States, Fast Green FCF is the least used of the seven main FDA approved dyes [27].

- **Thymol Blue**(thymolsulphonephthalein) is a brownish-green or reddish-brown crystalline powder that is used as a pH indicator. It is insoluble in water but soluble in alcohol and dilute alkali solutions. It transitions from red to yellow at pH 1.2–2.8 and from yellow to blue at pH 8.0–9.6. It is usually a component of Universal indicator. For this thesis we have used the thymol blue solution at neutral acid-base conditions, whose color is yellow.

- **Cobalt(II) chloride** is an inorganic compound of cobalt and chlorine, with the formula $\mathbf{CoCl_2}$. It is usually supplied as the hexahydrate $CoCl_2 \bullet 6H_2O$, which is one of the most commonly used cobalt compounds in the laboratory. [28]The hexahydrate is deep purple in color, that yields a pink solution when dissolved in water, whereas the anhydrous form is sky blue. Because of the ease of the

hydration/dehydration reaction, and the resulting color change, cobalt chloride is used as an indicator for water in desiccants. Niche uses include its role in organic synthesis and electroplating objects with cobalt metal. Another of its uses is as invisible ink, therefore a good chemical dye.

- **Methylene Blue** (CI 52015) is a heterocyclic aromatic chemical compound with the molecular formula $C_16H_18N_3SCl$. It has many uses in a range of different fields, such as biology and chemistry.At room temperature it appears as a solid, odorless, dark green powder, that yields a blue solution when dissolved in water. The hydrated form has 3 molecules of water per molecule of methylene blue. Methylene blue should not be confused with methyl blue, another histology stain, new methylene blue, nor with the methyl violets often used as pH indicators. One of its medical uses is as a dye or stain, therefore appropriate for the purposes of this thesis. [29]

The diluted substances are illustrated in figure 3.10.

Figure 3.10: Second row (from left to right): Fast Green, Malachite Green, Methylene Blue. Third row (from left to right): Methyl Orange, $CuSO_4$, $CoCl_2$, Thymol

The substances above have been chosen in a way that no interaction could occur among them. Seemingly, most of them are ph indicators, therefore chemical substances that are added in small amounts to a solution so that the pH (acidity or basicity) of the solution can be determined visually. Another important factor for the choice of the substances was the peaks of their spectra. For this purpose, we collected a variety of substances with peaks on the right, middle or left part of the optical spectrum and a use of various combinations of them in the mixtures in order to allow the decomposition of close or distant spectra, and observe the algorithms' performance in both cases. Considering the fact that for overlapping (close to each other) spectra the decomposition would be more difficult, this was a challenging part of the problem we researched.

3.2.2 Experimental Design

Experimental Design or *Design of Experiments(DOE)* remains a core area of study in chemometrics and several monographs are specifically devoted to experimental design in chemical applications.

Although the actual significance of following one of the experimental designs mentioned in bibliography is about how the factors interact with each other, and what a factor can be, we adapted some types of already existing experimental designs and came up with exemplified ones, considering that, for example the existence of 2 factors means that there are 2 components in a mixtures, 3 factors mean 3 components, etc. Hence, each one of the designs for the construction of the experimental datasets is more of a symbolic experimental design, and not following word-for-word the designs described in bibliography.

For this reason, we are going to make some conventions. First of all, when the word *factors* is mentioned, we refer to the different constituents contained in a mixture (2 factors for 2 components, 3 factors for 3 components etc) and when mentioning *factor levels*, low or high, we refer to the lowest or highest admissible absorbance 2.5, and therefore concentration that a sample was obtained at, respectively.

Some types of experimental designs generally used are the following [30] [31] [32]:

- **Full factorial** design or complete factorial design is when the response variable (here Absorbance or transmission) is measured for all possible combinations of the chosen factor levels. In this type of design, the number of possible combinations,

that is possible samples or experiments in order to form an experimental dataset is 2^m, where $m \geq 2$ is the number of factors or chemical components, e.g. for 2 factors there are 4 combinations, for 3 factors there are 8 combinations, for 4 factors there are 16 combinations etc. A more representative example of full factorial design for 3 factors is displayed in table 3.1. For the highest possible concentration we set '1' and for the lowest admissible one we set '0'.

- **Fractional factorial** design, which is either half-fractional, 1-quarter fractional, 1-eighth fractional etc. The individual experiments in the fractional design must be carefully chosen to ensure that they give the maximum information. This type of design involves $2^(m-1)$ experiments (samples). A more representative example of a half-fractional factorial design for 4 factors is presented in table 3.2. Also, an example of a quarter-fractional factorial design for 5 factors is presented in table 3.3.

- **Face-centered** design is a type of design including a full factorial design but instead of two levels (low and high) it also includes a central point. As seen for the full factorial design, the first 2^m experiments are the same as for a Factorial Design. Then, the next $2m$ experiments are obtained by keeping all the variables except one at their central level (a Star Design). A more representative example of a half-fractional factorial design is presented in table 3.4. For the lowest possible concentration we set '-1', for the central one '0' and for the highest admissible we set '1'.

- **Mixture** design is a type of design usually applied in pharmaceutical industry, food products beverages, paintings, method optimization. The reason that it is not applicable in our diploma thesis is that in a mixture we must cope with the implicit constraint that the sum of all the components must be 1 (or 100%). The number of components under study in a mixture design is not higher than four [30].

When applied to single spectra, the above designs correspond to a dataset of single experiments, whereas for the Hyperspectral Image experiments they correspond to whole images, each one consisting of many repeats of the same (as far as possible) single spectrum. In other words, for example, for two factors the applied experimental design for

Table 3.1: Complete factorial design for three factors

Sample Number	A compound	B compound	C compound
1	1	1	1
2	1	1	0
3	1	0	1
4	1	0	0
5	0	1	1
6	0	1	0
7	0	0	1
8	0	0	0

spectrometer data provides a dataset of 9 samples or 9 spectra, whereas for the microscope data, it provides a dataset of 9 images, each image consisting of xy (size of the image) same spectra!

To summarize, for a particular number of factors we followed a specific design, considering the fact that we would like a sufficient number of experiments for each number of factors combination, but not too many, as shown below:

- **2 factors**: face-centered 9-sample dataset design
- **3 factors**: full factorial 8-sample dataset design
- **4 factors**: half-factorial 8-sample dataset design
- **5 factors**: quarter-factorial 8-sample dataset design

3.2.3 The individual components' spectra

As mentioned before in 2.4, the concentration of a chemical substance dissolved in water is measured in Molarity (units: $\frac{mol}{L}$ or M), which represents the number of moles of a solute per litre of solution. The concentration C is defined as follows:

$$C = \frac{n}{V} \tag{3.1}$$

Table 3.2: Half-factorial design for four factors

Sample Number	A compound	B compound	C compound	D compound
1	1	1	1	1
2	1	1	0	0
3	1	0	1	0
4	1	0	0	1
5	0	1	1	0
6	0	1	0	1
7	0	0	1	1
8	0	0	0	0

Table 3.3: Quarter-factorial design for five factors

Sample Number	A	B	C	D	E
1	0	0	0	0	0
2	0	0	1	0	1
3	0	1	0	1	1
4	0	1	1	1	0
5	1	0	0	1	1
6	1	0	1	1	0
7	1	1	0	0	0
8	1	1	1	0	1

Table 3.4: Face centered design for two factors

Sample Number	A compound	B compound
1	-1	-1
2	+1	-1
3	-1	+1
4	+1	+1
5	-1	0
6	+1	0
7	0	-1
8	0	+1
9	0	0

where n is the moles (molecular weight) of the solute(mol) and V is the volume of the solution(L).

For the purposes of this diploma thesis some solutions based on the substances analyzed above were provided. At first, we obtained an adequate quantity of chemical solutions prepared by the graduate students of the Analytical and Environmental Chemistry Laboratory of Technical University of Crete. These solutions were based on the solid substances mentioned in previous section with known molecular weight. We are not going to develop the preparation chemical procedure, because it is not concerning this thesis. Secondly, and before mixing the substances and prepare our experimental datasets, we measured with the spectrophotometer the absorbance spectra of the individual components with the initial concentrations given. Subsequently, we diluted the solutions further, adding 1 or $2ml$ of water in $3ml$ of a solute, according how "heavy" or "light" the substance was, in order to reach a solution whose absorbance was close to the upper absorption limit according to Beer-Lambert law 2.5. The term *close* means that in our experiments, and by extension for the whole problem solution, the mixtures were of a greater importance than the individual components solutions, as well as when preparing a mixture of 2, 3 or more components, the individual concentration of each is further diminished. Therefore, we could have an individual component spectrum with

an absorbance a little bit above 1 (1.1-1.2) at some wavelengths which mathematically does not have a big influence. Furthermore, after acquiring the first admissible solution, we continued diluting the solutions to a greater extent, so that we obtained a number of individual component absorbance spectra in different concentrations, until reaching the lower admissible spectrum limit according to Beer-Lambert law.

The equation from chemistry that connects initial and final concentration of a solutions is presented below:

$$C_1V_1 = C_2V_2 \tag{3.2}$$

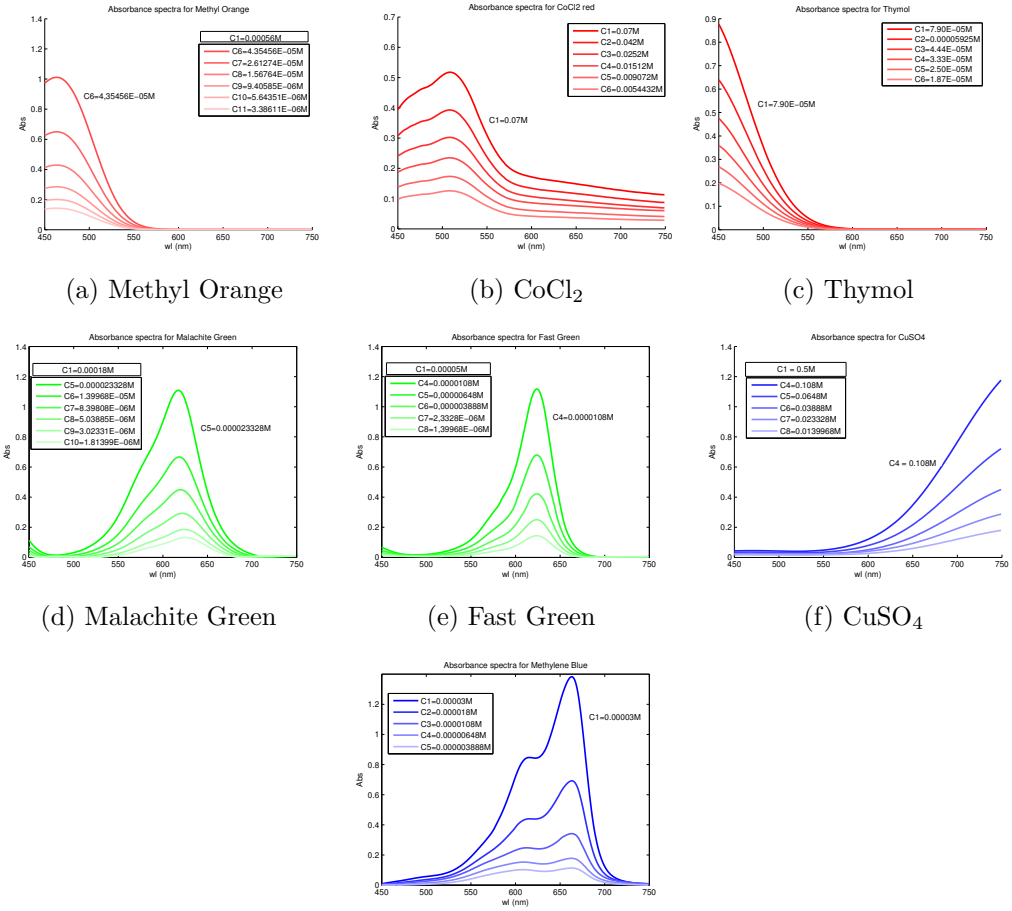
where C_1 is the initial concentration of the substance before the dilution, V_1 its initial volume, C_2 is the final concentration and V_2 is the final total volume of the solution (V_1 + V_{water}).

Further down(figure 3.11), the absorbance spectra of the individual component solutions are displayed. As stated above, the indexes next to the concentrations appearing below are connected to the first admissible concentration we kept for the specific substance. We also include in the diagrams the first concentration given beyond which the dilutions were made. It should be noted that the dilutions for Methyl Orange, $CuSO_4$, Malachite Green, Fast Green and Methylene blue were prepared with $3ml(V_1)$ of solution and $2ml(V_2{=}5ml)$ water, whereas for $CoCl_2$ and Thymol were prepared with $3ml(V_1)$ of solution and $1ml(V_2{=}4ml)$ water.

3.2.4 The mixtures

The chemical equation for the mixtures is the same as for the individual spectra (see equation (3.2)) except that it is calculated for each one of the substances individually. We should note that the mixtures of 2 components were prepared with $2ml$ of each component (total of $4ml$), whereas the mixtures of 3 components were prepared with $1ml$ of each component (total of $3ml$).

Further down(figures 3.12, 3.13, 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20), the absorbance spectra of the mixtures, which follow the experimental designs for 2 and 3 components, are displayed.



(a) Methyl Orange (b) $CoCl_2$ (c) Thymol

(d) Malachite Green (e) Fast Green (f) $CuSO_4$

(g) Methylene Blue

Figure 3.11: Individual Components' Spectra

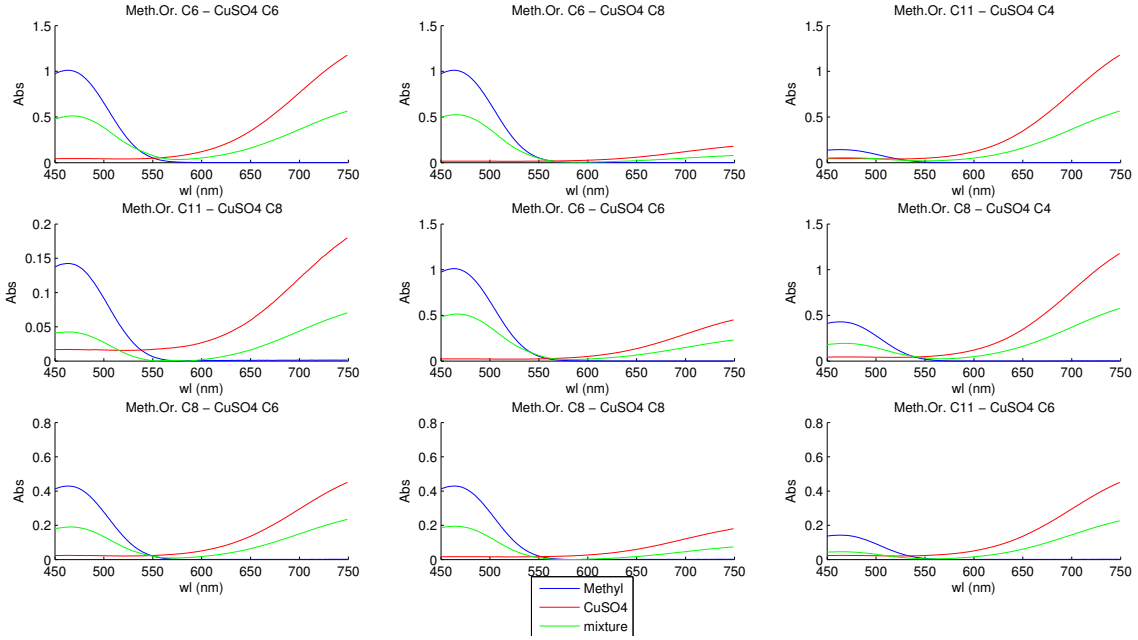


Figure 3.12: Methyl Orange - $CuSO_4$ mixtures absorbance spectra

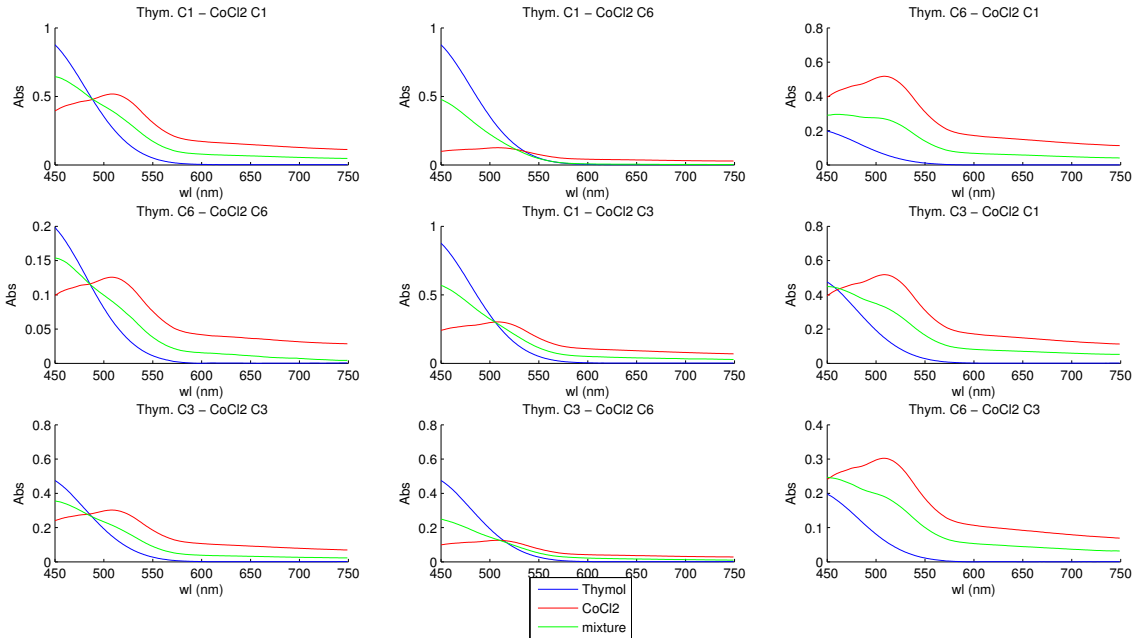


Figure 3.13: Thymol - $CoCl_2$ mixtures absorbance spectra

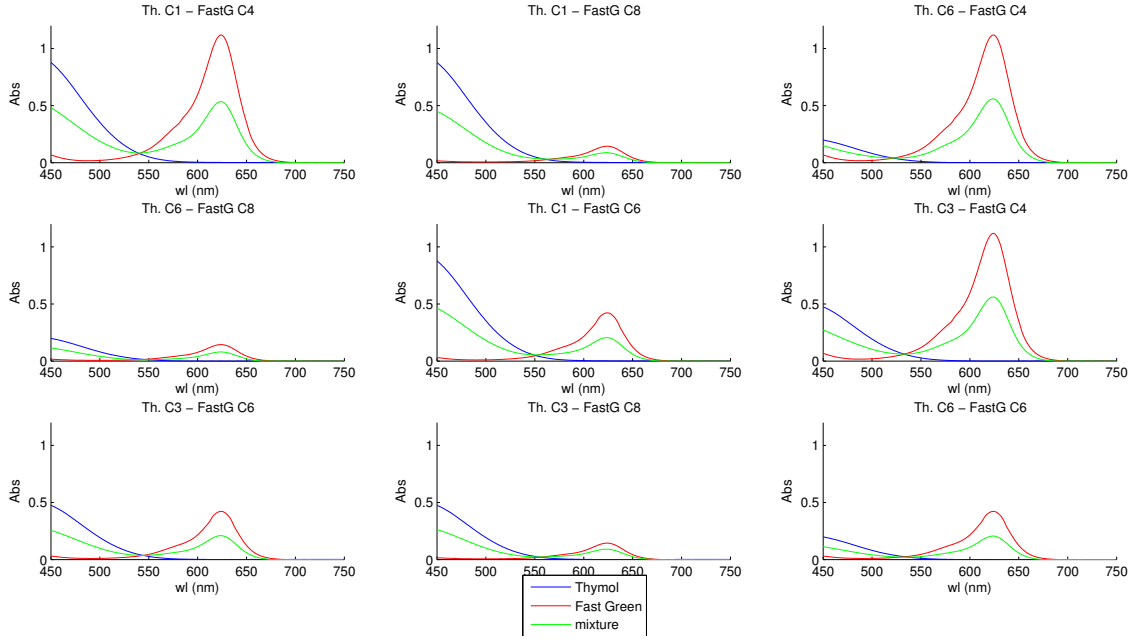


Figure 3.14: Thymol - Fast Green mixtures absorbance spectra

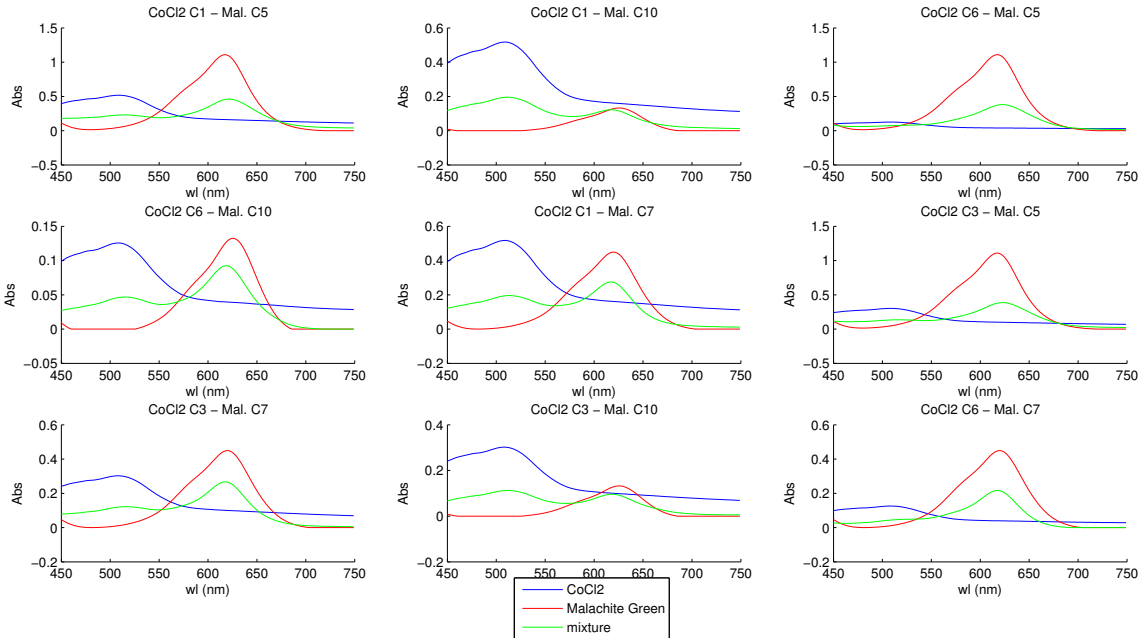


Figure 3.15: $CoCl_2$ - Malachite Green mixtures absorbance spectra

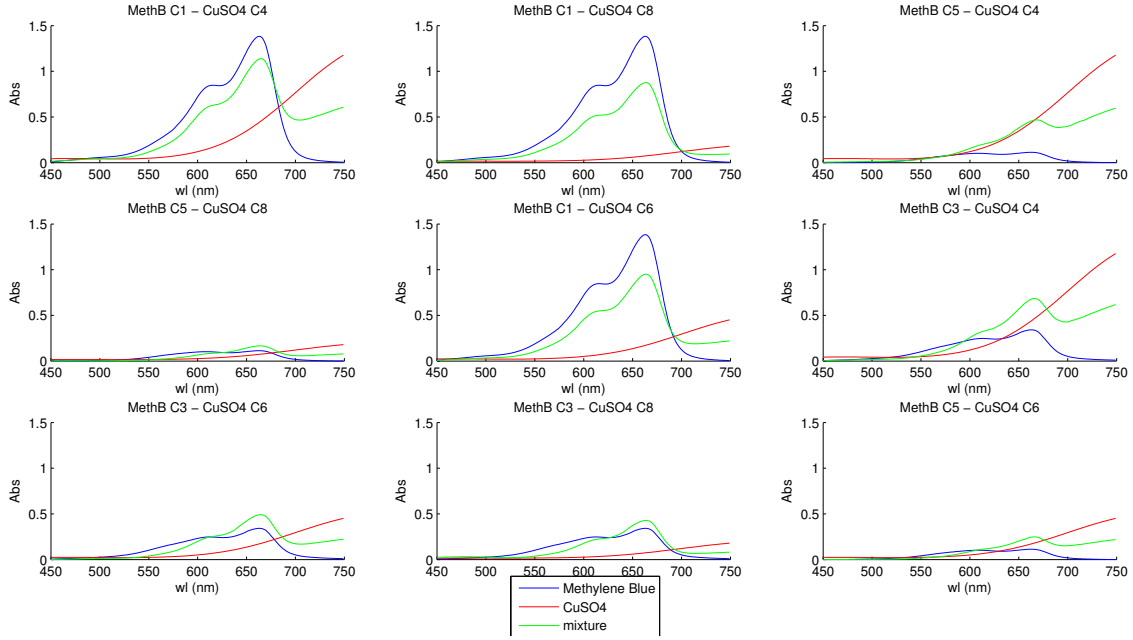


Figure 3.16: Methylene Blue - $CuSO_4$ mixtures absorbance spectra

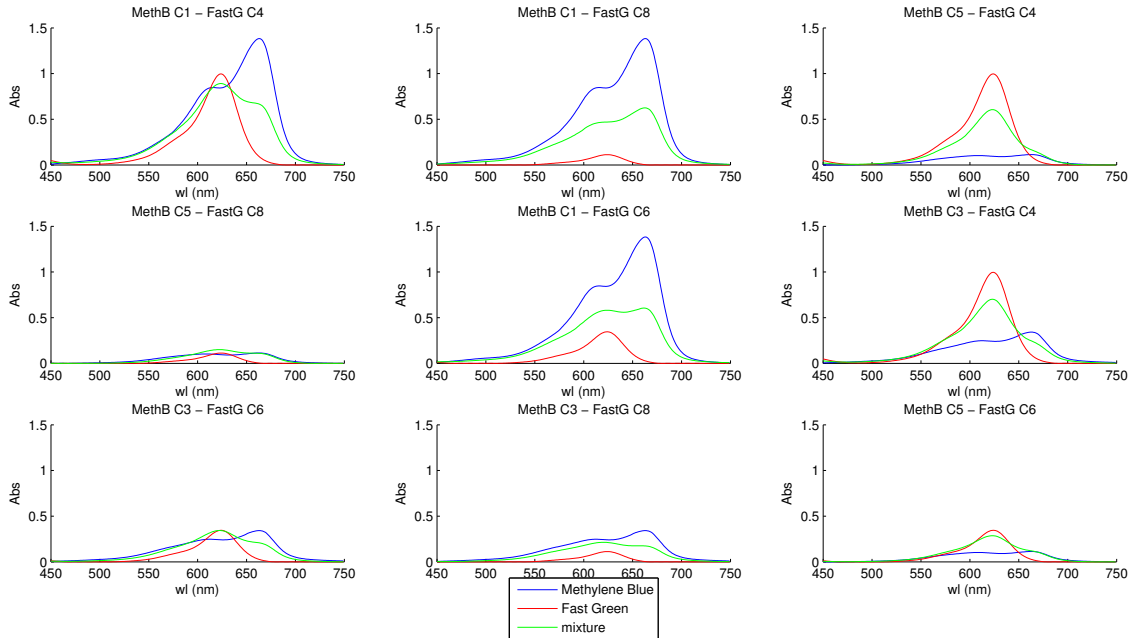


Figure 3.17: Methylene Blue - Fast Green mixtures absorbance spectra

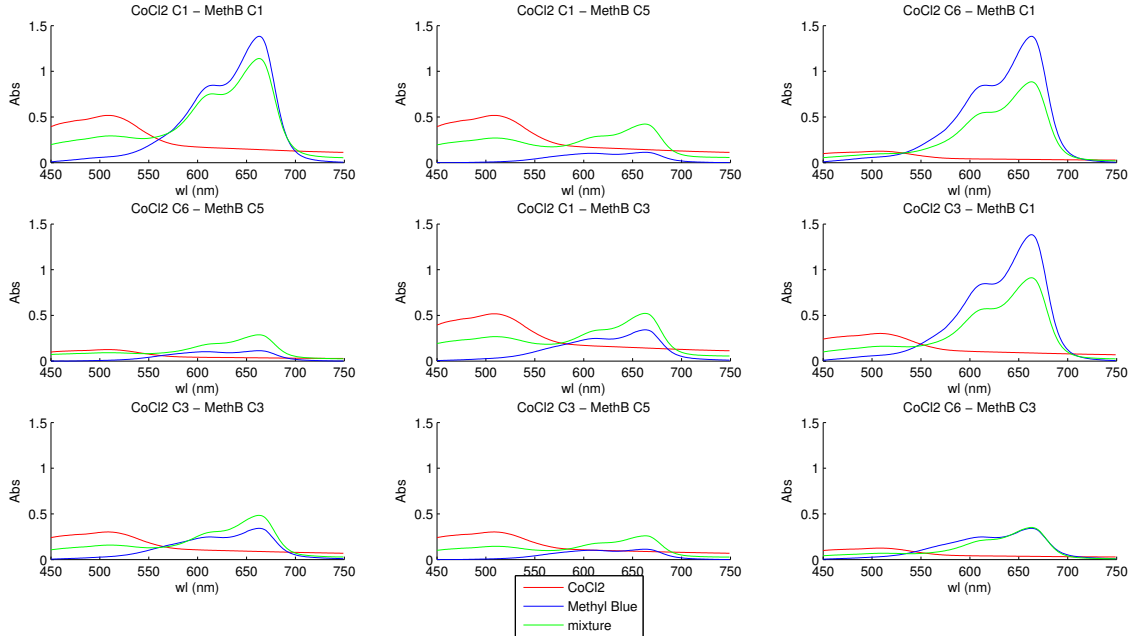


Figure 3.18: $CoCl_2$ - Methylene Blue mixtures absorbance spectra

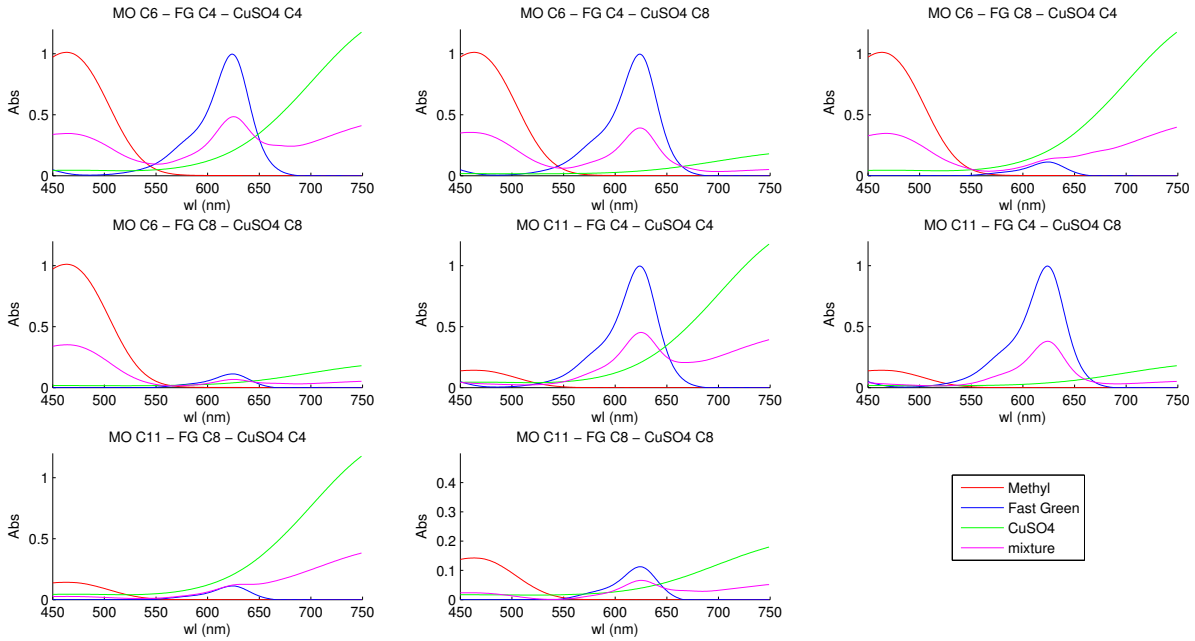


Figure 3.19: Methyl Orange - Fast Green - $CuSO_4$ mixtures absorbance spectra

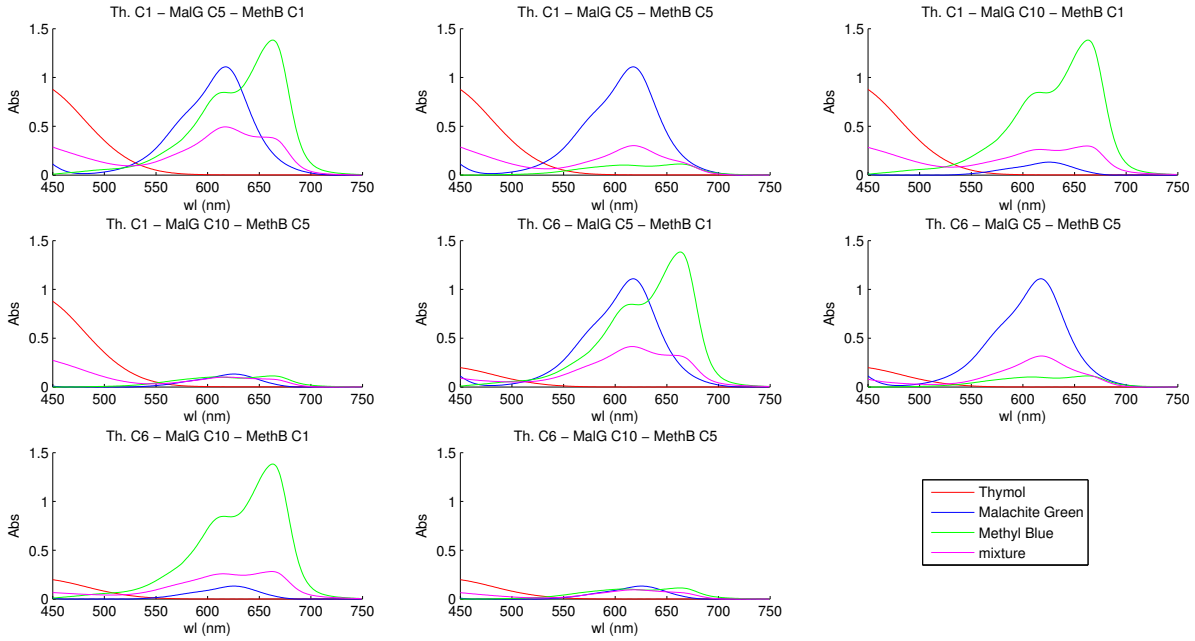


Figure 3.20: Thymol - Malachite Green - Methylene Blue mixtures absorbance spectra

3.2.5 Hyperspectral Image Data

For this part, we have collected two hyperspectral cubes using the microscope configuration. These cubes contained the images from the experimental designs for two and three components. For the two components dataset, we collected the mixtures of Methylene Blue and Fast Green compound solutions, whereas for the three components dataset, we collected mixtures of $CoCl_2$, Fast Green and $CuSO_4$. The first dataset spectra were gathered with a sampling of $5nm$, whereas the second were gathered every $10nm$.

It should be noted that the CCD sensors of the spectral imager/camera of the microscope might introduce some three main sources of noise which could lead to errors in the image acquisition. These are the following:

1.Readout Noise: It appears during the reading of the signal. This kind of noise is dependent to the inherent CCD preamplifier, which differs from one CCD model to another, and to the speed with which the charge is transferred and transmitted from the preamplifier. The bigger the transferring and transmission speed the greater the noise.

2.Dark Current Noise: CCDs build up "dark current" whether the CCD is being exposed to light or not. Dark current is caused by thermally generated electrons that build up in the pixels of all CCDs. The rate of dark current accumulation depends on the temperature of the CCD but will eventually completely fill every pixel in a CCD. Managing dark current is particularly important for astrophotography because of the long exposures typically required for night sky imaging. The pixels in a CCD are cleared before beginning an exposure, but dark current starts accumulating again immediately. The rate of dark current build up can be reduced by a factor of 100 or more by cooling the CCD. The remaining dark current is subtracted from an image using dark frames.

3.Photon Noise of Photon Shot Noise: Shot noise is caused by the random arrival of photons. This is a fundamental trait of light. Since each photon is an independent event, the arrival of any given photon cannot be precisely predicted; instead the probability of its arrival in a given time period is governed by a Poisson distribution. With a large enough sample, a graph plotting the arrival of photons will plot the familiar bell curve. Shot noise is most apparent when collecting a relatively small

number of photons. It can be reduced by collecting more photons, either with a longer exposure or by combining multiple frames.

3.3 Source of Error

There are three types of error that occur from mistakes in data acquisition [9], ***gross error, systematic error*** and ***random error***.

The first type of error, ***gross error***, is usually connected to an analyst's mistake in weighing, calibration or even calculation during the experiment. Repeating the experiment might show up this error. It may be possible to identify such an error and remove that result from further consideration, but there is no other way we can usefully employ the result once this error has occurred. It must be noted that careless and unrecorded expunging of results could amount to scientific fraud. However, because of its unique nature, gross error cannot guide our future actions.

This second type of error, ***systematic error***, is a permanent deviation from the true result. When applied to an instrument, systematic error is known as *bias*. The reason why we obtain such an error may be due to a flawed measurement method. Systematic error can be estimated by measuring a reference material a large number of times. The difference between the average of the measurements and the value of the reference material is the systematic error. It is always desirable to know the sources of systematic error in an experiment and to correct them in measurements.

As seen in the paragraph above, in order to estimate the systematic error, it is suggested that the experiment be repeated a large number of times. This is necessary because of the contribution of another source of error, namely ***random error***. Despite your best efforts, having considered and removed or corrected for sources of systematic error, having ironed out gross errors, repeating experiments always seems to give slightly different answers. There are a myriad of factors that can contribute to random error: the inability of the analyst to exactly reproduce conditions, fluctuations in the environment (temperature, pressure), rounding of arithmetic calculations, brief gusts of wind, or a shake of the analyst's hand. What do not contribute to random error are changes in conditions such as the regular drift in baseline of an instrument and the aging of a chromatography column.

Apart from all of the above, we must consider an important factor, the clarity of the solutions used. Samples containing solid material, or which are cloudy, are difficult to analyze using a spectrophotometer,so anyone must be very careful before beginning the absorption analysis.

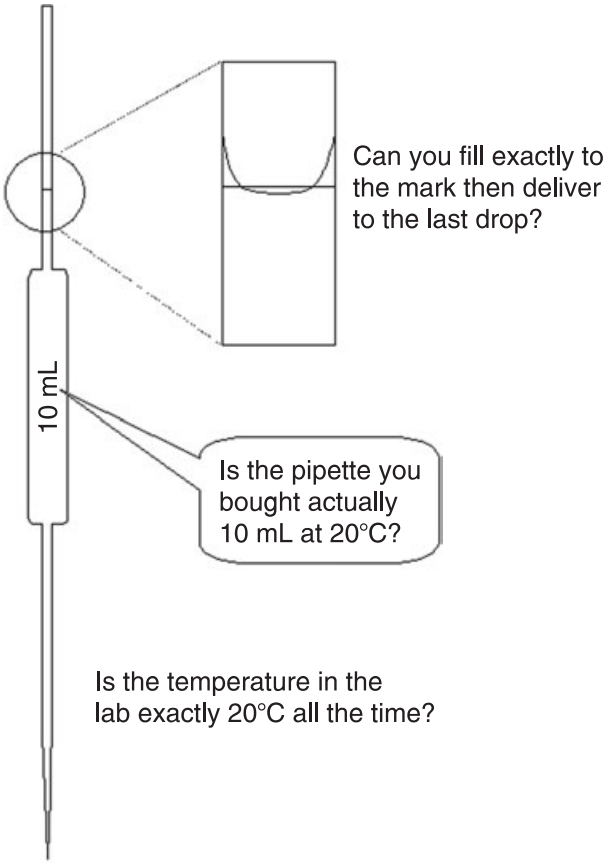


Figure 3.21: Uncertainties and errors in delivering 10 ml by a pipette.

An example - pipetting

Considering why we might not deliver exactly 10 ml using a 10 ml pipette is instructive (figure 3.21).We shall identify three contributing factors to the problem.

1. The manufacturer will admit that the pipette used,when filled properly to the mark at 20°C, is only guaranteed to have a volume somewhere between 9.98 and 10.02 ml. Luckily,perhaps an analyst will have a 10.00 ml pipette, but perhaps not. Any error of this type is a **systematic error**.

2. When you use a pipette, it may be difficult or even impossible to really fill it exactly to the same mark each time. A series of 10 experiments of filling a pipette with distilled water and weighing what runs out, gives a range of values from 9.95 to 10.04 ml. Thus, the analyst's contribution to the error is definitely **random**.

3. It should be noted that during the experiments the temperature in the laboratory fluctuates between 19.2 and 23.1°C, and the volume of 10 ml of water will increase by 0.0021 ml approximately for every degree centigrade rise in temperature. If the experiments take long enough to allow the temperature to change in a **random** fashion about some average, then these changes will be included in the results. In addition, unless the average temperature during the experiments was exactly 20°C there will also be a **systematic error** arising from the difference.

3.4 Simulated Data

For the requirements of this diploma thesis, except for the experimental data, the use of some sets of simulated data has been employed. The ultimate reason for this, was that the spectrophotometer could not allow the use of larger cuvettes, thus, in order to check what happens for mixtures of 4 or 5 constituents we needed to use simulated data from an excel file. Another reason was to observe what happens when the data are devoid of any gross, systematic or random errors analyzed in the section above, and if the algorithmic methods deliver in the same way.

The simulated datasets were obtained from the *RegressionDemo.ods* file from the following website [33]. The individual and mixture spectra in the file are following a normal distribution and random noise can be added in them. The file also contains the pure spectra (spectra at unit concentration) for each component used.

Further down(figure 3.22, 3.23, 3.24, 3.25, 3.26, 3.27, 3.28) the absorbance spectra of the simulated data mixtures, which follow the experimental designs for 2,3,4 and 5 components, are displayed.

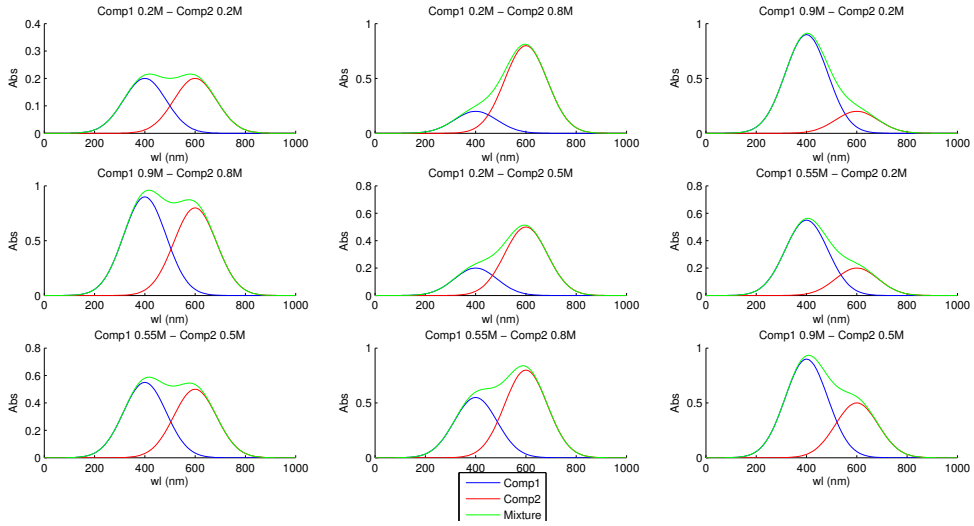


Figure 3.22: Absorbance mixture spectra for 2 components (Simulated data)

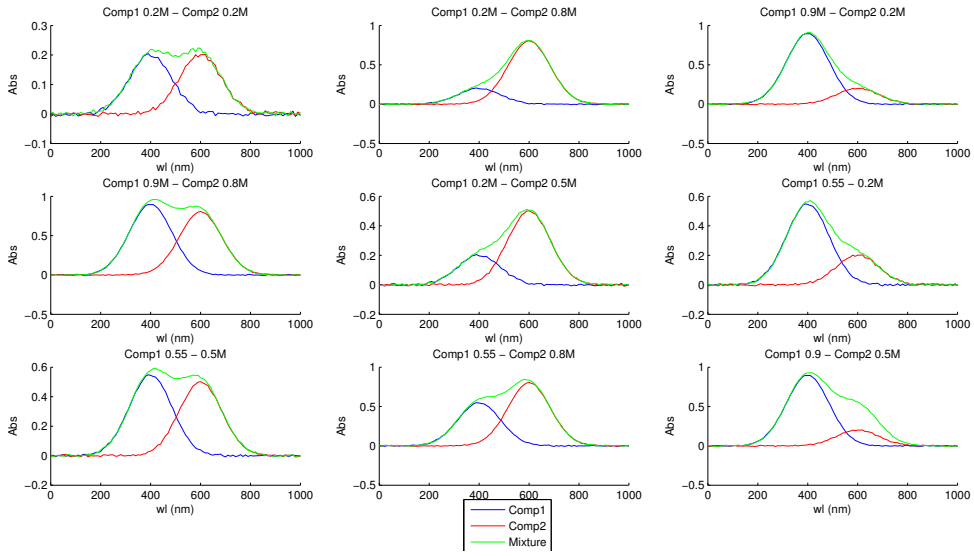


Figure 3.23: Absorbance mixture spectra for 2 components with random noise=0.01 (Simulated data)

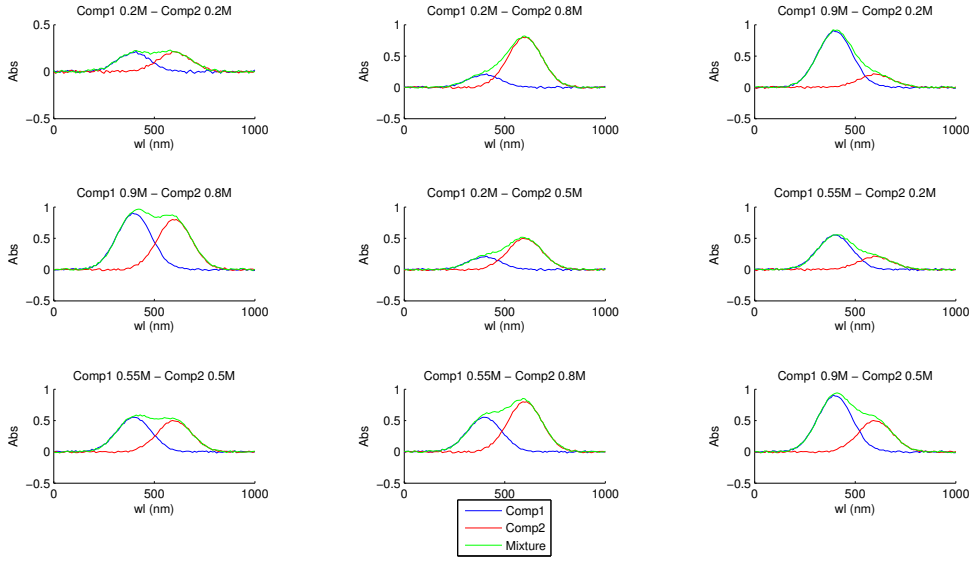


Figure 3.24: Absorbance mixture spectra for 2 components with random noise=0.02 (Simulated data)

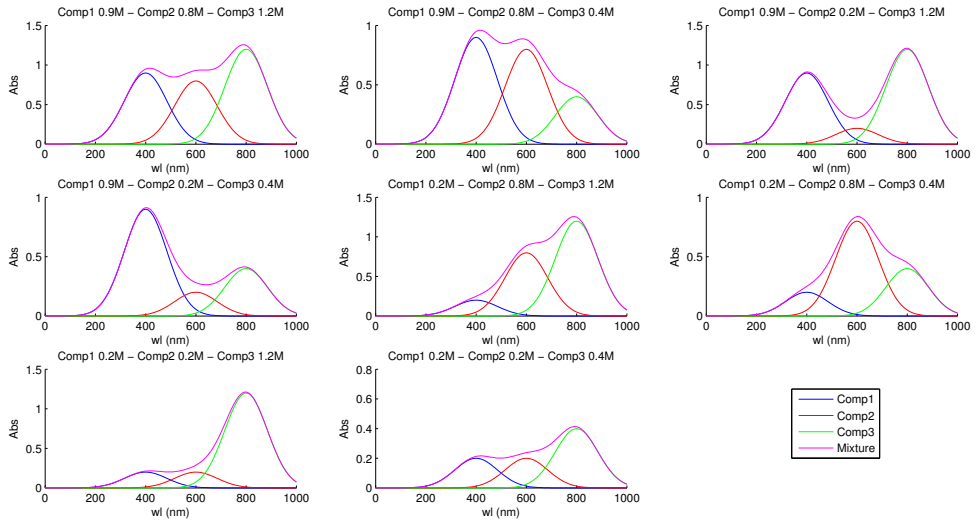


Figure 3.25: Absorbance mixture spectra for 3 components (Simulated data)

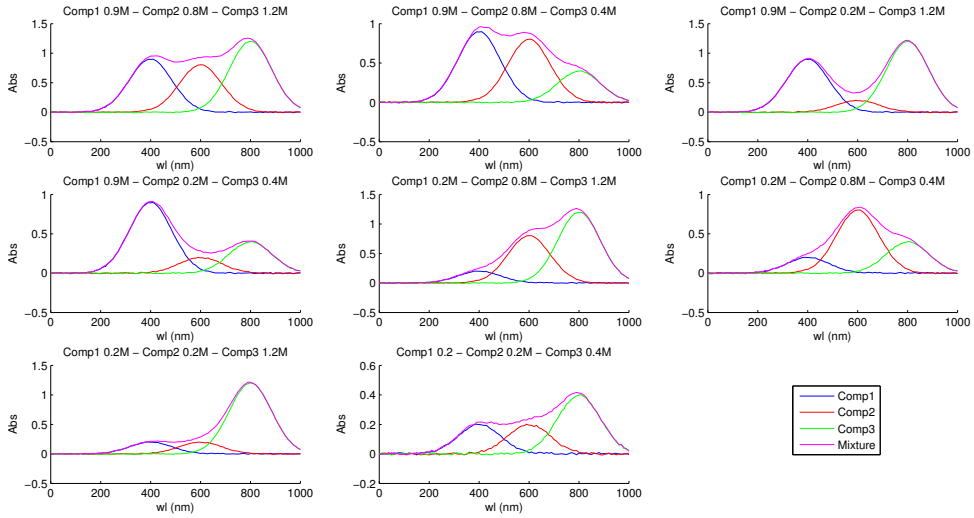


Figure 3.26: Absorbance mixture spectra for 3 components with noise=0.01 (Simulated data)

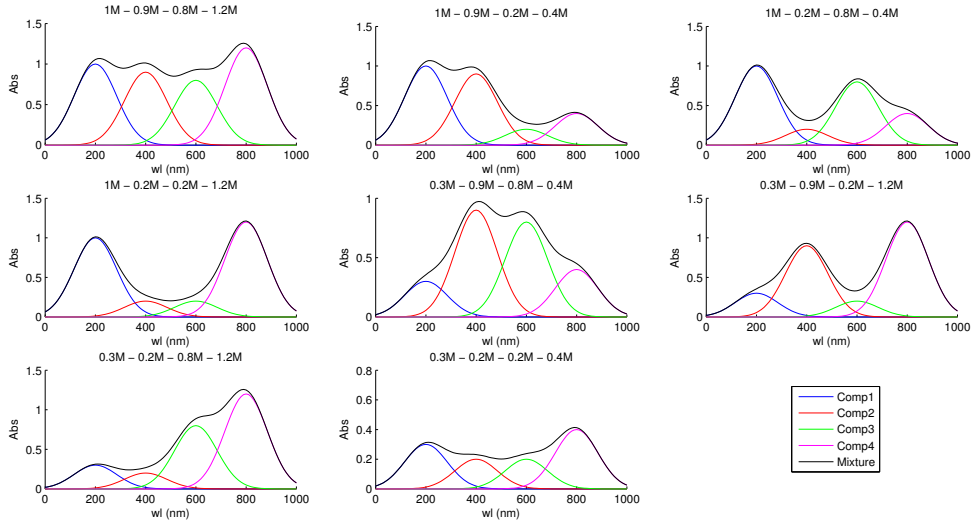


Figure 3.27: Absorbance mixture spectra for 4 components (Simulated data)

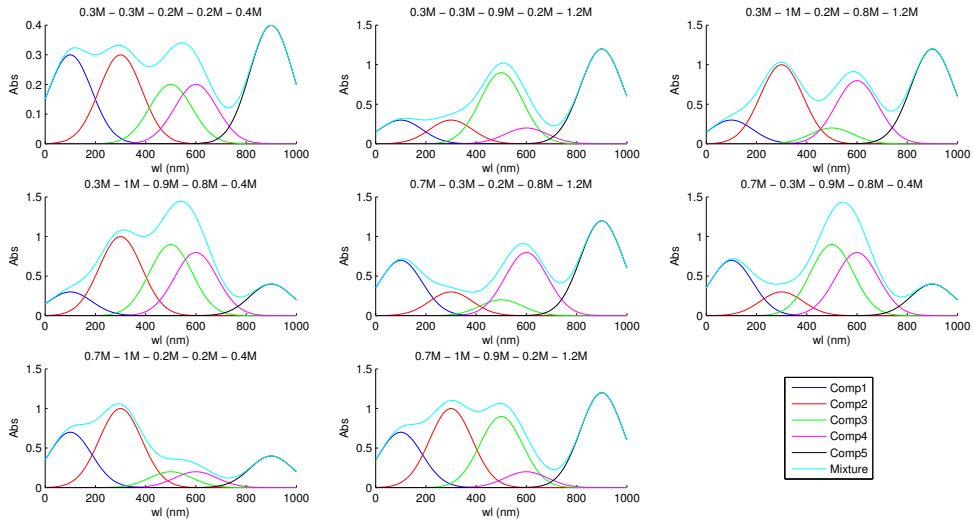


Figure 3.28: Absorbance mixture spectra for 5 components (Simulated data)

Chapter 4

Methods and System Validation

4.1 Calibration

All scientific activities, and more specifically many chemical problems and applications of chemometrics involve calibration. This is to make sure that the reported measurement results are reliable and to make an estimate of accuracy and precision. The calibration activity in a laboratory pertains to a wide range of actions. Examples are: obtaining pure water at the correct temperature and pure dry chemicals to make standard solutions with the aid of an analytical balance (calibrated and with all necessary corrections made) and calibrated flasks [13]. Furthermore,the calibration of an instrument is associated with the selection of reference standards with known values to cover the range of interest.

Calibration is a concept used in the context of chemometrics: calibration data and calibration models are used to allow the prediction of dependent(response) variable values from related independent (predictor) variables. Actually, it is the creation of a mathematical model in order to relate the output of an instrument (eg absorbance) to the properties of a sample(eg concentration of the analytes) [34]. Whereas with prediction, we use the model to predict properties of a sample, given the instrument output.

Of course, as we are going to observe further down in this chapter, we have created calibration models not only for classical few-sample datasets but for whole hyperspectral images. The importance of image calibration over classical calibration is that because in hyperspectral imaging we can have many more sample, it enables more opportunities for statistical testing, for example by making histograms of residuals. Another advantage of images is that all samples as object have spatial coordinates; this makes it possible to

construct images from prediction or residual values enabling additional visual inspection and interpretation.

The main advantages of the use of multivariate calibration techniques is that fast, cheap, or non-destructive analytical measurements (such as optical spectroscopy) can be used to estimate sample properties which would otherwise require time-consuming, expensive or destructive testing.

Univariate Calibration

The simplest form of a linear calibration model is $y_i = b_1 x_i + e_i$, where y_i represents the response (dependent) variable, normally the concentration of the i_{th} calibration sample, x_i denotes the corresponding instrument reading, the explanatory (independent) variable, normally the absorbance for the i_{th} sample; , b_1 symbolizes the calibration coefficient (slope of the fitted line), and e_i signifies the error associated with the i_{th} calibration sample, assumed to be normal distributed random, $N(0, 1)$. A **single instrument response**, e.g., absorbance at a **single wavelength**, is measured for each calibration sample [35]. In matrix algebra notation, the model is expressed as:

$$\mathbf{y} = \mathbf{x}b_1 + \mathbf{e} \tag{4.1}$$

and depicted as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} b_1,$$

where n is the number of calibration samples. The above is the simplest calibration model that has a slope and no intercept. It is a simple linear regression, in which a single x -variable (because we have one wavelength) is linked to a single y -variable. However, when there are more than one response variables in the model, e.g. the absorbance of a spectrum at **many wavelengths**, then we are talking about a ***Multiple Linear Regression***(MLR) model, which in matrix form is expressed as [36]:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{4.2}$$

and (without an intercept depicted as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix},$$

where n is the number of calibration samples and d is the number of wavelengths.

Multivariate Calibration

Univariate calibration is specific to situations where the instrument response depends only on the target analyte concentration. With multivariate calibration, model parameters can be estimated where responses depend on the target analyte in addition to other chemical or physical variables and, hence, multivariate calibration corrects for these interfering effects [35].

For this diploma thesis, multivariate calibration would be helpful when the subject of concern is two or more target analytes. In fact, the type of calibration models we have dealt with are based on a **multiple multivariate calibration** model (multiple = for many wavelengths, multivariate = for many components). The goal of a multivariate calibration is to establish a connection between a multivariate signal $\boldsymbol{X}$ and one or more physical or chemical properties $\boldsymbol{Y}$. In other words, in a multiple multivariate regression system, with n calibration samples, d wavelengths and m constituents, we have data matrices $\boldsymbol{X}$ ($n \times d$) and $\boldsymbol{Y}$ ($n \times m$) for the equation:

$$\mathbf{Y} = \mathbf{XB} \tag{4.3}$$

In general, multivariate analysis is appropriate when the spectra of the constituents overlap so that their concentrations cannot be determined without previous chemical separation. In order to calibrate the system a number of specimens containing different mixtures of the analytes are taken and the spectrum is measured for each specimen [31].

Non-zero intercepts and Mean-Centering

Equation 4.2 assumes that the instrument response provides a value of zero when the analyte concentration is zero. Nonetheless, in reality the instrument response (e.g. absorbance) usually has no good correlation with the components, so a non-zero intercept

should be added to the model. This constant term when added provides the calibration model a place to discard the "garbage" from this bad correlation [37]. All is needed to do to add a non-zero intercept is inserting a column of 1's to the original $\mathbf{X}$ matrix, as follows:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix}$$

An intercept of zero for a model can be obtained if $\mathbf{y}$ and $\mathbf{x}$ are mean-centered to respective means before using equation 4.2. It should be noted that while the calibration line for mean-centered $\mathbf{y}$ and $\mathbf{x}$ as an intercept of zero, inherently, a nonzero intercept is generally involved. The nonzero intercept is removed by the mean-centering process. Thus, mean-centering $\mathbf{y}$ and $\mathbf{x}$ to generate a zero intercept is not the same as using the original data and constraining the model to have an intercept of zero. In the absence of mean centering, it is possible to include a nonzero intercept, b_0, in a calibration model, by expressing the model as:

$$y_i = b_0 + x_{i1}b_1 + ... + x_{ik}b_d \tag{4.4}$$

In the following lines a representation of the centering for the response and predictor variables is shown.

Let $\mathbf{x}$ be a $n \times 1$ given column of $\mathbf{X}$, $\mathbf{y}$ a $n \times 1$ given column of $\mathbf{Y}$ and $\mathbf{1}$ a $n \times 1$ vector of 1's. We can center $\mathbf{x}$ and $\mathbf{y}$ by subtracting their means:

$$\dot{\mathbf{x}} = \mathbf{x} - \mathbf{1}\bar{x} \qquad \Rightarrow \qquad \mathbf{x} = \dot{\mathbf{x}} - \mathbf{1}\bar{x} \tag{4.5}$$

$$\dot{\mathbf{y}} = \mathbf{y} - \mathbf{1}\bar{y} \qquad \Rightarrow \qquad \mathbf{y} = \dot{\mathbf{y}} - \mathbf{1}\bar{y} \tag{4.6}$$

The column-wise centered $\mathbf{X}$ and $\mathbf{Y}$ are:

$$\dot{\mathbf{X}} = \mathbf{X} - \mathbf{1}\bar{x} \tag{4.7}$$

$$\dot{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\bar{y} \tag{4.8}$$

4.2 Classical Calibration Methods

There are two main categories associated with *multivariate calibration analysis*, **Classical Least Squares** model and **Inverse Least Squares** Model. The first one is based on Beer-Lambert law 2.5, whereas the second is based on the inverse Beer-Lambert. A graphical representation of the methods related to these models and the reason to choose each of theme, is illustrated in figure 4.1 [34]. Classical Least Squares is an individual method itself whereas Inverse Least Squares is distinguished in MLR model, Principal Component Regression(PCR), Partial Least Squares(PLS), as shown in figure 4.1.

Although, due to our experimental datasets, we know all of the analytes, for some methods (like PCR, PLS in the sections below) according to 4.1 this is not necessary. Each one of these calibration methods is divided in further algorithmic approaches, depending on whether the data are mean-centered or not or whether there is a non-zero intercept in the calibration model.

4.2.1 Classical Least Squares (CLS)

The first spectroscopic quantification method analysed and applied is **Classical Least Squares** (CLS) method, also known as *Direct Least Squares*, *Beer's law method*, **K**-*matrix* and is more closely related to the way chemists & spectroscopists think about spectra, and not so much how mathematicians or statisticians perceive them.

This method is based on Beer-Lambert law 2.2, although it is a multi-component Beer-Lambert law. Matrix **B** (or **K**) emerges from $A{=}\varepsilon bC$ where $\varepsilon{=}b{=}1$, thus $A{=}\varepsilon$, which means that for unit concentration and unit pathlength the absorbance response is equal to the "absorptivities" [38]. More specifically, the columns of **B** (from MLR model 4.2)are the pure component spectra at unit concentration and unit pathlength (absorptivity-pathlength products). Thus, if the pathlength is kept constant and concentration matrix **C** is not equal to 1, then the CLS model can be expressed as:

$$\mathbf{X} = \mathbf{YA}, \tag{4.9}$$

where **X** is the $n \times d$ matrix of absorbance mixture spectra,where n is the number of samples and d is the number of wavelengths **Y** is the $n \times p$ matrix of concentrations (for p components) and **A** is the $p \times d$ **K**-matrix of absorbances for pure component spectra or "absorptivities" or calibration coefficients.

Constraints: A constraint that must be counted for the CLS method is $n \geq p$, that is, the number of mixture calibration samples must be equal or larger than the number of components in the mixture. Otherwise, the system is underdetermined and no unique solution is possible.

Advantages and Disadvantages of CLS [10]: Classical least squares is a fast full-spectrum technique, meaning that it does not require wavelength selection. As long as the number of wavelengths exceeds the number of constituents, any number can be used, even the entire spectrum. In addition, using a large number of wavelengths tends to give an averaging effect to the solution, making it less susceptible to noise in the spectra. However, this technique requires knowing the complete composition (concentration of every constituent) of the calibration mixtures and is not useful for mixtures with constituents that interact. CLS is also known as "total calibration" because all components are evaluated simultaneously.

Direct Classical Least Squares (DCLS)

In this type of Classical Least Squares method, the calibration coefficients should be already known, either by being measured directly or by mathematical calculations. The first case is almost impossible to achieve at most times, because the coefficients should be measured for unit concentrations of the analytes as previously declared, to whit, solutions may be very dense, at a non-liquid state,making it impossible to be prepared. Thus, in that case we have in our disposal the simulated data. An implementation of DCLS, with the absorbancies and pure spectra already available is applied only for the simulated datasets and not for the other ones.

However, using mathematical calculations, we tried to measure the pure spectra of our components. For this reason, we obtained some absorbance spectra from the spectrophotometer, each one related to a different concentration (all for the same compound). We did the same for the other components we used in the mixtures. Subsequently, we created a fitted line between the absorbances of a compound and its concentrations and from the slope of this regression line, we found the "absorptivities", also known as regression coefficients or pure spectra.

Indirect Classical Least Squares (ICLS)

In this approach of CLS, the regression coefficients (pure spectra) are not available, thus, they must be estimated from the model itself. Therefore, we solve equation (4.9) for **A** (**K**-matrix):

$$\widehat{\boldsymbol{A}} = \mathbf{Y}^{+}\mathbf{X} \tag{4.10}$$

Other solutions for this system, apart from using the pseudo-inverse, can be seen in section 5.3. Once this equation is solved, it can be used to predict concentrations of unknown samples, as follows:

$$\widehat{\boldsymbol{Y}} = \mathbf{X}/\mathbf{A} \tag{4.11}$$

If **X**, **Y** are mean-centered, the model is expressed as follows [36]:

$$\dot{\boldsymbol{X}} = \dot{\boldsymbol{Y}}\boldsymbol{A} \tag{4.12}$$

The concentrations of the individual spectra can be predicted through the following procedure:

$$\widehat{\boldsymbol{A}} = \dot{\boldsymbol{Y}}^{+}\dot{\boldsymbol{X}} \Rightarrow \tag{4.13}$$

$$\dot{\boldsymbol{Y}} = \dot{\boldsymbol{X}}\boldsymbol{A}^{\vdash} \tag{4.14}$$

We know that (4.14) = (4.8), thus for concentration prediction we have:

$$\widehat{\boldsymbol{Y}} = \dot{\boldsymbol{X}}\boldsymbol{A}^{\vdash} + \overline{\boldsymbol{y}} \tag{4.15}$$

where **X** is the $n \times d$ matrix of absorbance mixture spectra,where n is the number of samples and d is the number of wavelengths **Y** is the $n \times p$ matrix of concentrations (for p components) and **A** is the $p \times d$ **K**-matrix of absorbances for pure component spectra or "absorptivities" or calibration coefficients.

Overall, CLS is an extremely useful model for spectroscopic analysis because it provides quantitative and interpretative chemical information.

4.2.2 Inverse Least Squares (ILS)

This method is known as *Inverse Least Squares* (ILS), or *Multiple Linear Regression* (MLR, here as Multivariate Multiple Linear Regression) or **P**-*matrix*.

Inverse Least Squares is generally based on sequential feature (wavelength) selection, in Multiple Linear Regression via **Step-wise Regression** and in Multivariate Linear Regression via **Forward Selection** or **Backward Elimination**. Step-wise regression is a sequential process for fitting the least squares model, where at each step a single explanatory variable is either added to or removed from the model in the next fit(see [39]). For the purposes of our problem, forward selection and backward elimination are applied on the data with the existence or not of a non-zero intercept.

Forward selection

Forward selection procedure begins with no explanatory variable in the model and sequentially adds a variable at each step, using a so-called *wrapper* method implementing a learning algorithm and applying cross-validation in order to select features. The goodness of features (criterion function) is either measured by the Mahalanobis distance or by the **Euclidean distance**, (or the square root of the *Residual Sum of Squares*, as in our case) [35]. This procedure has two limitations. Some of the variables never get into the model and hence their importance is never determined. Another limitation is that a variable once included in the model remains there throughout the process, even if it loses its stated significance, after the inclusion of other variable(s) [40].

Backward elimination

The backward elimination procedure begins with all the variables in the model and proceeds by eliminating the least useful variable at a time using a similar method as this applied for forward selection and the Euclidean distance criterion. The disadvantage of Backward selection is that it requires more computation time than Forward selection.

The ILS method is actually expressed in the same way as the MLR model (see equation (4.3)). Thus, if we solve (4.3) for **B**, we will get:

$$\widehat{\boldsymbol{B}} = \boldsymbol{X}^{+}\boldsymbol{Y} \tag{4.16}$$

where $\mathbf{X}$ is the $n \times d$ matrix of absorbance mixture spectra,where n is the number of samples and d is the number of wavelengths $\mathbf{Y}$ is the $n \times p$ matrix of concentrations (for p components) and $\mathbf{B}$ is the $d \times p$ $\mathbf{P}$-matrix of absorbances for pure component spectra or "absorptivities" or calibration coefficients. Once this equation is solved, it can be used to predict concentrations of unknown samples, as follows:

$$\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\mathbf{B}} \tag{4.17}$$

If $\mathbf{X}$, $\mathbf{Y}$ are mean-centered, the model is expressed as follows [36]:

$$\dot{\boldsymbol{Y}} = \dot{\boldsymbol{X}}\boldsymbol{B} + \boldsymbol{F} \tag{4.18}$$

The concentrations of the individual spectra can be predicted through the following procedure: First, we solve for $\mathbf{B}$:

$$\widehat{\boldsymbol{B}} = \dot{\boldsymbol{X}}^{+}\dot{\boldsymbol{Y}} \tag{4.19}$$

Thus,

$$\dot{\boldsymbol{Y}} = \dot{\boldsymbol{X}}\widehat{\boldsymbol{B}} \tag{4.20}$$

We can easily observe that (4.14) = (4.20) $\Rightarrow \widehat{\boldsymbol{B}} = \boldsymbol{A}^{\vdash}$
Thus, from (4.20) = (4.8), we have for concentrations prediction $\Rightarrow$

$$\widehat{\boldsymbol{Y}} = \dot{\boldsymbol{X}}\boldsymbol{B} + \overline{\boldsymbol{y}} \tag{4.21}$$

Constraints: Unlike the classical least squares methods, inverse least squares is not a full-spectrum method, but requires careful selection of the wavelengths of the absorbances that are used in the calibration model. The number of wavelengths selected should not exceed the number of calibration samples and usually it is smaller than the number of chemical components in the mixtures, that is $n >= d$ if there is a zero intercept and $n >= (d+1)$ if a non-zero intercept exists.

Advantages and Disadvantages of ILS: A disadvantage of wavelength selection over full-spectrum techniques is the ability to detect unusual samples because of the elimination of variables. Moreover, wavelength selection can be difficult and time consuming. However, ILS is relatively fast and allows calibration of very complex mixtures since only knowledge of constituents of interests (and not of all components) is required.

4.2.3 Principal Components Regression (PCR)

The **Principal Components Regression** method combines the *Principal Components Analysis* (PCA) spectral decomposition with an *Inverse Least Squares* (ILS) regression method to create a quantitative model for complex samples. Unlike quantification methods based directly on Beer's Law, which attempt to calculate the absorbtivity coefficients for the constituents of interest from a direct regression of the constituent concentrations onto the spectroscopic responses, the PCR method regresses the concentrations on the PCA scores. We consider the principal components of $\boldsymbol{X}^\top \boldsymbol{X}$.

PCA: There are several ways of finding the principal components of the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix. One possibility is to apply the SVD method to $\boldsymbol{X}$, writing the reduced form of SVD as follows:

$$\mathbf{X} = \mathbf{UDP^T}, \tag{4.22}$$

where $\mathbf{U}$ is the $n \times f$ left singular values (LSV) matrix, $\mathbf{D}$ is the $f \times f$ singular values (SV) matrix and $\mathbf{P}$ is the $d \times f$ right singular values (RSV) matrix, or **loadings** matrix, where f is the number of PCA eigenvectors. Let the **scores** matrix be defined by

$$\boldsymbol{T} = \boldsymbol{UD}, \tag{4.23}$$

a matrix with orthogonal, but not necessarily orthonormal columns [41]. In fact

$$\boldsymbol{T}^\top \boldsymbol{T} = \boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{D} = \boldsymbol{D}^2 = \boldsymbol{\Lambda}_r, \tag{4.24}$$

where $\boldsymbol{\Lambda}_r = \text{diag}\{\lambda_1, \ldots, \lambda_r\}$ contains the non-zero eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ in its diagonal. We assume that the eigenvalues are in decreasing order, $\lambda_1 \geq \cdots \geq \lambda_r > 0$. In this diploma thesis, we performed the singular value decomposition on the covariance matrix of $\boldsymbol{X}$. It should be noted that when $\boldsymbol{X}$ is a centered data matrix, then $\boldsymbol{X}^\top \boldsymbol{X}/(n-1)$ is the covariance matrix of $\boldsymbol{X}$.

Since

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^\top \tag{4.25}$$

we find that $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{P}\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{P}^\top = \boldsymbol{P}\boldsymbol{\Lambda}_r \boldsymbol{P}^\top$, which is the spectral decomposition for $\boldsymbol{X}^\top \boldsymbol{X}$, except that columns of $\boldsymbol{P}$ corresponding to zero eigenvalues have been left out. By using that $\boldsymbol{P}$ is orthogonal, we may also write (4.25) as follows:

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}, \tag{4.26}$$

which follows by noting that $\boldsymbol{XP} = \boldsymbol{TP}^\top \boldsymbol{P} = \boldsymbol{T}$. To sum up, the columns of $\boldsymbol{T}$ are known as *scores*, and those of $\boldsymbol{P}$ as *loadings*.
We can find the principal components of $\boldsymbol{X}^\top \boldsymbol{X}$ or the eigenvalues, choosing the first g columns of $\boldsymbol{T}$ and the first g columns of $\boldsymbol{P}$, such as we form matrices $\boldsymbol{T}_g$ and $\boldsymbol{P}_g$ respectively.

In order to help rationalize the choice of g , the relative size of the eigenvalues are expressed as a percentage of the sum of all eigenvalues,

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_r} \times 100 \tag{4.27}$$

and this percentage is interpreted as the percent variation explained by the corresponding principal component. Often, the accumulated percentages are used, so that the percent variation explained by the first g components is

$$\frac{\lambda_1 + \cdots + \lambda_g}{\lambda_1 + \cdots + \lambda_r} \times 100 \tag{4.28}$$

As a rule, g should be chosen so that at least about 80-90 percent of the variation is explained.

The basic idea in Principal Components Regression (PCR) is that after choosing a suitable value for g, the important features of $\boldsymbol{X}$ have been retained by $\boldsymbol{T}_g$. We then perform the MLR with $\boldsymbol{T}_g$ in place of $\boldsymbol{X}$ for an $n \times m$ calibration data matrix $\boldsymbol{Y}$,

$$\boldsymbol{Y} = \boldsymbol{T}_g \boldsymbol{C} + \boldsymbol{F}. \tag{4.29}$$

The least squares method then gives

$$\widehat{\boldsymbol{C}} = \left(\boldsymbol{T}_g^\top \boldsymbol{T}_g\right)^{-1} \boldsymbol{T}_g^\top \boldsymbol{Y}, \tag{4.30}$$

where $\boldsymbol{T}_g^\top \boldsymbol{T}_g$, being diagonal, is easy to invert. The fact that we have left out the loadings matrix $\boldsymbol{P}_g$ in (4.29) is of no consequence for prediction, because the scores are linear combinations of the columns of $\boldsymbol{X}$, and the PCR method amounts to singling out those linear combinations that are best for predicting $\boldsymbol{Y}$.

For prediction with PCR, it is necessary to turn to $\boldsymbol{X}$ again, and using (4.26) we may write the regression equation as follows:

$$\boldsymbol{Y} = \boldsymbol{T}_g \boldsymbol{C} + \boldsymbol{F} = \boldsymbol{X} \boldsymbol{P}_g \boldsymbol{C} + \boldsymbol{F} \tag{4.31}$$

Thus, the concentrations are predicted using the following equation:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{P}_g\widehat{\boldsymbol{C}} + \overline{\boldsymbol{y}}, \tag{4.32}$$

where $\boldsymbol{P}_g\widehat{\boldsymbol{C}}$ is called the regression matrix, and may be compared with the $\widehat{\boldsymbol{B}}$ matrix of MLR.

Advantages and Disadvantages of PCR: This method does not require wavelength selection. Any number can be used; usually the whole spectrum. A larger number of wavelengths gives averaging effect, making model less susceptible to spectral noise. PCR can be used for very complex mixtures since only knowledge of constituents of interest is required and can sometimes be used to predict samples with constituents (contaminants) not present in the original calibration mixtures. Nonetheless, PCR calculations are slower than most classical methods. Generally, a large number of samples are required for accurate calibration, but collecting calibration samples can be difficult, because collinear constituent concentrations must be avoided.

4.2.4 Partial Least Squares (PLS)

The PCR method from the previous module represents a considerable improvement over MLR and CLS. By using latent variables (scores), it is possible to use a large number of variables (frequencies), just as in CLS, but without having to know about all interferences.

Problems may arise, however, if there is a lot of variation in $\boldsymbol{X}$ that is not due to the analyte as such. PCR finds, somewhat uncritically, those latent variables that describe as much as possible of the variation in $\boldsymbol{X}$. But sometimes the analyte itself gives rise to only small variations in $\boldsymbol{X}$, and if the interferences vary a lot, then the latent variables found by PCR may not be particularly good at describing $\boldsymbol{Y}$. In the worst case important information may be hidden in directions in the $\boldsymbol{X}$-space that PCR interprets as noise, and therefore leaves out.

Partial Least Squares Regression (PLS) is able to cope better with this problem, by forming variables that are relevant for describing $\boldsymbol{Y}$ [36] [42] [36].

PLS1

The PLS1 algorithm starts with the initialization $j = 1$, $\boldsymbol{X}_1 = \boldsymbol{X}$ and $\boldsymbol{y}_1 = \boldsymbol{y}$. The algorithm then proceeds through the following steps to find the first g latent variables:

1. Let $\boldsymbol{w}_j = \boldsymbol{X}_j^\top \boldsymbol{y}_j / \left\| \boldsymbol{X}_j^\top \boldsymbol{y}_j \right\|$.
2. Let $\boldsymbol{t}_j = \boldsymbol{X}_j \boldsymbol{w}_j$.
3. Let $\widehat{c}_j = \boldsymbol{t}_j^\top \boldsymbol{y}_j / \boldsymbol{t}_j^\top \boldsymbol{t}_j$.
4. Let $\boldsymbol{p}_j = \boldsymbol{X}_j^\top \boldsymbol{t}_j / \boldsymbol{t}_j^\top \boldsymbol{t}_j$.
5. Let $\boldsymbol{X}_{j+1} = \boldsymbol{X}_j - \boldsymbol{t}_j \boldsymbol{p}_j^\top$ and $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j - \boldsymbol{t}_j \widehat{c}_j$.
6. Stop if $j = g$; otherwise let $j = j + 1$ and return to Step 1.

Now form the two $d \times g$ matrices $\boldsymbol{W}$ and $\boldsymbol{P}$ and $n \times g$ matrix $\boldsymbol{T}$ with columns $\boldsymbol{w}_j$, $\boldsymbol{p}_j$ and $\boldsymbol{t}_j$, respectively, and form a column vector $\widehat{\boldsymbol{c}}$ ($g \times 1$) with elements $\widehat{c}_j$. Let

$$\widehat{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{P}^\top = \sum_{j=1}^{g} \boldsymbol{t}_j \boldsymbol{p}_j^\top \tag{4.33}$$

and

$$\widehat{\boldsymbol{y}} = \boldsymbol{T}\widehat{\boldsymbol{c}} = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{P}^\top \boldsymbol{W})^{-1}\widehat{\boldsymbol{c}}, \tag{4.34}$$

which are the predicted values of $\boldsymbol{X}$ and $\boldsymbol{y}$, respectively. The matrix $\boldsymbol{W}$ is orthogonal, and $\boldsymbol{T}$ has orthogonal columns.

It should be noted that, in spite of the similarities with the NIPALS algorithm, the PLS1 algorithm is recursive and requires exactly g steps, whereas the NIPALS algorithm is iterative, the number of iterations cannot be determined in advance, and is dependent on the choice of a stopping criterion. In this sense, the PLS1 algorithm is simpler than the NIPALS algorithm.
After the g runs have been completed in step 6, the following relations hold:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^\top + \boldsymbol{X}_{g+1}\boldsymbol{y} = \boldsymbol{T}\widehat{\boldsymbol{c}} + \boldsymbol{y}_{g+1}. \tag{4.35}$$

Prediction for the PLS1 method is slightly more complicated, than for PCR, in spite of the algorithm being simpler. Consider a new prediction sample $\boldsymbol{z}$ ($1 \times d$ vector) and predicted value $\overrightarrow{\boldsymbol{y}}$ (both uncentered). Note the new notation for the predicted value. Let $\overline{\boldsymbol{x}}$ ($d \times 1$) and $\overline{y}$ be the calibration sample averages. The prediction is performed by essentially retracing the steps of the algorithm, letting the row vector $\boldsymbol{z} - \overline{\boldsymbol{x}}$ follow the same steps as a row of the $\boldsymbol{X}$ matrix.

Let $\boldsymbol{W}$, $\boldsymbol{T}$, $\boldsymbol{P}$ and $\widehat{\boldsymbol{c}}$ be the matrices and vector formed after applying the PLS1 algorithm to the calibration data. Initialize by taking $j = 1$ and $\boldsymbol{x}_j = \boldsymbol{z} - \overline{\boldsymbol{x}}$. Then proceed through the following steps:

1. Let $t_j = \boldsymbol{x}_j \boldsymbol{w}_j$.
2. Let $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j - t_j \boldsymbol{p}_j^\top$.
3. Let $j = j + 1$, and repeat Steps 1 to 3 until $j = g$.

Now form the row vector $\widehat{\boldsymbol{t}} = (t_1, \ldots, t_g)$, and complete the prediction as follows:

$$\overrightarrow{\boldsymbol{y}} = \overline{y} + \widehat{\boldsymbol{t}}\widehat{\boldsymbol{c}}. \tag{4.36}$$

It is possible, though, to summarize the prediction in a matrix formula as follows:

$$\overrightarrow{\boldsymbol{y}} = \overline{y} + (\boldsymbol{z} - \overline{\boldsymbol{x}})^\top \widehat{\boldsymbol{b}}, \tag{4.37}$$

where $\widehat{\boldsymbol{b}}$, the so-called regression vector, is

$$\widehat{\boldsymbol{b}} = \boldsymbol{W} \left(\boldsymbol{P}^\top \boldsymbol{W}\right)^{-1} \widehat{\boldsymbol{c}}. \tag{4.38}$$

PLS2

As already mentioned, one may use PLS1 separately for each analyte ($\boldsymbol{Y}$ -column), which allows a separate optimal model to be constructed for each analyte. It may, however, be advantageous to include information from other analytes when predicting any specific analyte. This may be done by constructing an overall model describing $\boldsymbol{Y}$ as a function of $\boldsymbol{X}$, and for this purpose we may use the PLS2 method.

When several analytes are to be predicted simultaneously, the situation becomes more complicated than for the PLS1 algorithm. Suffice it to say that separate application of the PLS1 algorithm to each column of $\boldsymbol{Y}$ would lead to different sets of scores being formed for each $\boldsymbol{Y}$ -column. In PLS2, these separate scores are in effect reconciled into a single set of scores, but this extra constraint implies a more complex algorithm. Note that such a complication does not arise in connection with PCR, because PCR does not take $\boldsymbol{Y}$ into account when forming the scores.

The principle behind the PLS2 algorithm may be outlined as follows. Similar to PLS1, we form a model for $\boldsymbol{X}$ namely

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^\top + \boldsymbol{X}_{g+1} \tag{4.39}$$

when g scores are to be used. The scores (columns of $\boldsymbol{T}$) is the single set of scores alluded to above. But now we form a similar model for $\boldsymbol{Y}$, namely

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^\top + \boldsymbol{Y}_{g+1}. \tag{4.40}$$

This includes a second set of scores for $\boldsymbol{Y}$, namely the columns of $\boldsymbol{U}$. These two equations are linked by an inner relationship,

$$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{C} + \boldsymbol{U}_{g+1}, \tag{4.41}$$

meaning a relationship that holds between latent, rather than observed variables. The two matrices $\boldsymbol{U}$ and $\boldsymbol{T}$ are both $n \times g$, and $\boldsymbol{T}$ has orthogonal columns. $\boldsymbol{P}$ is a $d \times g$ matrix, $\boldsymbol{Q}$ is an $p \times g$ matrix whose columns are unit vectors, and $\boldsymbol{C}$ is a $g \times g$ diagonal matrix of regression coefficients. Similar to PLS1, we will also need the $d \times g$ orthogonal matrix $\boldsymbol{W}$.

The three 'error' terms $\boldsymbol{X}_{g+1}$ $\boldsymbol{Y}_{g+1}$ and $\boldsymbol{U}_{g+1}$ are supposed to represent noise. Hence g should be chosen large enough to make the term $\boldsymbol{X}_{g+1}$ useless for predicting $\boldsymbol{Y}_{g+1}$; in other words, $\boldsymbol{X}_{g+1}$ and $\boldsymbol{Y}_{g+1}$ should be approximately uncorrelated. Ignoring the error terms in (4.40), (4.41) and using the estimated value of $\boldsymbol{C}$, we obtain the predicted value of $\boldsymbol{Y}$ as follows:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}\widehat{\boldsymbol{C}}\boldsymbol{Q}^\top. \tag{4.42}$$

We now proceed to describe the actual PLS2 algorithm, which, like the NIPALS algorithm, is iterative, rather than just recursive. The algorithm starts with the initialization $j = 1$, $\boldsymbol{X}_1 = \boldsymbol{X}$ and $\boldsymbol{Y}_1 = \boldsymbol{Y}$, and then proceeds through the following steps to find the first g terms:

1. The vector $\boldsymbol{u}_j$ is initialized to be an arbitrary column of $\boldsymbol{Y}_j$.
2. Let $\boldsymbol{w}_j = \boldsymbol{X}_j^\top \boldsymbol{u}_j / \left\| \boldsymbol{X}_j^\top \boldsymbol{u}_j \right\|$.
3. Let $\boldsymbol{t}_j = \boldsymbol{X}_j \boldsymbol{w}_j$.

4. Let $\boldsymbol{q}_j = \boldsymbol{Y}_j^\top \boldsymbol{t}_j / \left\| \boldsymbol{Y}_j^\top \boldsymbol{t}_j \right\|$.

5. Let $\boldsymbol{u}_j = \boldsymbol{Y}_j \boldsymbol{q}_j$.

6. If $\boldsymbol{u}_j$ is unchanged continue with Step 7; otherwise go back to Step 2.

7. Let $\widehat{c}_j = \boldsymbol{t}_j^\top \boldsymbol{u}_j / \boldsymbol{t}_j^\top \boldsymbol{t}_j$.

8. Let $\boldsymbol{p}_j = \boldsymbol{X}_j^\top \boldsymbol{t}_j / \boldsymbol{t}_j^\top \boldsymbol{t}_j$.

9. Let $\boldsymbol{X}_{j+1} = \boldsymbol{X}_j - \boldsymbol{t}_j \boldsymbol{p}_j^\top$ and $\boldsymbol{Y}_{j+1} = \boldsymbol{Y}_j - \widehat{c}_j \boldsymbol{t}_j \boldsymbol{q}_j^\top$.

10. Stop if $j = g$; otherwise let $j = j + 1$ and return to Step 1.

Now form the matrices $\boldsymbol{W}$, $\boldsymbol{T}$, $\boldsymbol{Q}$, $\boldsymbol{U}$ and $\boldsymbol{P}$ with columns $\boldsymbol{w}_j$, $\boldsymbol{t}_j$, $\boldsymbol{q}_j$, $\boldsymbol{u}_j$ and $\boldsymbol{p}_j$, respectively, and form the $g \times g$ diagonal coefficient matrix $\widehat{\boldsymbol{C}}$ with diagonal elements $\widehat{c}_j$. After g runs through the algorithm, the relations (4.39), (4.40), (4.41), (4.42) are satisfied. In the special case $p = 1$, PLS2 reduces to the PLS1, because then $\boldsymbol{q}_j$ in Step 4 is 1, and $\boldsymbol{u}_j = \boldsymbol{y}_j$ in Step 5.

Prediction for the PLS2 method is quite similar to the case of PLS1, as long as we take the extra elements of the PLS2 algorithm into account. Consider, as before, a new prediction sample $\boldsymbol{z}$ $(1 \times k)$ and predicted value $\widehat{\boldsymbol{y}}(\boldsymbol{z})$ $(1 \times m)$ (both uncentered), and let $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{y}}$ be the calibration sample averages.

In order to follow the steps of the PLS2 algorithm, now for the purpose of prediction, we initialize by taking $j = 1$ and $\boldsymbol{x}_1 = \boldsymbol{z} - \overline{\boldsymbol{x}}$. The prediction then proceeds through the following steps:

1. Let $t_j = \boldsymbol{x}_j \boldsymbol{w}_j$.

2. Let $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j - t_j \boldsymbol{p}_j^\top$.

3. Stop if $j = g$; otherwise let $j = j + 1$, and go back to Step 1.

Now form the row vector $\widehat{\boldsymbol{t}} = (t_1, \ldots, t_g)$, and complete the prediction as follows:

$$\widehat{\boldsymbol{y}}(\boldsymbol{z}) = \overline{\boldsymbol{y}} + \widehat{\boldsymbol{t}} \widehat{\boldsymbol{C}} \boldsymbol{Q}^\top. \tag{4.43}$$

It is also possible to write the prediction in matrix form, as follows:

$$\widehat{\boldsymbol{y}}(\boldsymbol{z}) = \overline{\boldsymbol{y}} + (\boldsymbol{z} - \overline{\boldsymbol{x}})\, \widehat{\boldsymbol{B}}, \tag{4.44}$$

where the regression matrix $\widehat{\boldsymbol{B}}$ is

$$\widehat{\boldsymbol{B}} = \boldsymbol{W}\left(\boldsymbol{P}^{\top}\boldsymbol{W}\right)^{-1}\widehat{\boldsymbol{C}}\boldsymbol{Q}^{\top}. \tag{4.45}$$

Differences between the two methods: For PLS1 the first equation remains as it is, but in the second equation **c** and **f** are vectors and we obtain a set of equations for each constituent in the mixture. This makes PLS1 more **time-consuming** in calculations, since a separate set of eigenvectors and scores must be generated for every constituent of interest. For training sets with a large number of samples and constituents, the increased time of calculation can be significant. PLS1 may have the largest advantage when analyzing systems that have constituent concentrations that are widely varied. If the concentration ranges of the constituents are approximately the same, PLS1 may have less of an advantage over PLS2 and will definitely take longer to calculate. However, like the PCR method, PLS2 calibrates for all constituents simultaneously. In other words, the results of the spectral decomposition for both of these techniques give one set of scores and one set of eigenvectors for calibration. Therefore, the calculated vectors are not optimized for each individual constituent. This may sacrifice some accuracy in the predictions of the constituent concentrations, especially for complex sample mixtures. In PLS1, a separate set of scores and loading vectors is calculated for each constituent of interest. In this case, the separate sets of eigenvectors and scores are specifically tuned for each constituent, and therefore, should give **more accurate** predictions than PCR or PLS2.

Advantages and Disadvantages of PLS: PLS method provides a single step decomposition and regression; thus, eigenvectors are directly related to constituents of interest. Also, calibrations are generally more robust provided that the calibration set accurately reflects range of variability expected in the unknown samples. An additional advantage to this method is that it can be used for very complex mixtures since only knowledge of constituents of interest is required. Moreover, it can sometimes be used to predict samples with constituents (contaminants) not present in the original calibration mixtures. However, PLS has its disadvantages, too. One of them is that the calculations are slower that most Classical methods, especially in PLS1. The models are more abstract, thus more difficult to understand and interpret. Generally, a large number of samples are required for accurate calibration, but collecting calibration samples can be difficult, because collinear constituent concentrations must be avoided.

4.3 Performance Evaluation

In order to evaluate the algorithms' results and be able to highlight the best, some error metrics had to be used, which we are going to analyze further down in this section.

4.3.1 Error Estimation

There are two types of errors associated with performance validation of calibration (regression) models for concentration estimation. The first category is about **Root Mean Squared Errors** (RMSEs) and the second one is the **standard error of regression**, $s_{y/x}$. The later can be used only in CLS and ILS methods, whereas the first one can be used in every method used in this diploma thesis.

RMSEs as used in our problem solving are associated with the concentrations because we want to know how close to the initial concentrations are the concentrations estimated. RMSEs can be absolute values or can be expressed in percentages. The second way of expressing them is easier to interpret from an analyst's perspective, because it indicates how much (in terms of %) we are close to the desirable result. A low percentage of an RMSE means the desired estimated concentration is closer to the initial one, than a high percentage of RMSE. However, the standard error of regression is only expressed as an absolute value and the goal is to be as close to '0' as possible, because the smaller this error is, the better the regression fit is.

The first RMSE error applied for the validation of our system was a training error, which we name *Root Mean Squared Error of Prediction* (RMSEP), because it emerges from our training data, which in our case are all the calibration samples used in a specific experimental design dataset and it is the following:

$$RMSEP = \sqrt{\frac{\sum_n (\mathbf{C} - \mathbf{C}_{pred})^2}{n}} \tag{4.46}$$

where n are the samples in the calibration model and the relative % RMSEP is:

$$rRMSEP = \sqrt{\frac{\sum_n (\frac{\mathbf{C}-\mathbf{C}_{pred}}{C})^2}{n}} 100\% \tag{4.47}$$

The second RMSE error, *Root Mean Squard Error of Calibration* (RMSEC) describes the degree of agreement between the calibration model estimated concentration values

for the calibration samples and the accepted true values for the calibration samples [35] and it is the following:

$$RMSEC = \sqrt{\frac{\sum_n (\mathbf{C} - \mathbf{C}_{pred})^2}{dof}} \tag{4.48}$$

and the relative % RMSEC is:

$$rRMSEC = \sqrt{\frac{\sum_n (\frac{\mathbf{C} - \mathbf{C}_{pred}}{C})^2}{dof}} 100\% \tag{4.49}$$

where dof are the degrees of freedom.

In statistics, the number of **Degrees of Freedom (d.o.f.)** is the number of values in the final calculation of a statistic that are free to vary. More specifically, in chemometrics the degrees of freedom are the number of data minus the number of parameters calculated from them. For example,in multiple regression with p independent variables, the standard error has $n-p-1$ degrees of freedom. This happens because the degrees of freedom are reduced from n by $p+1$ numerical constants $b_0, b_1, b2, \ldots, b_p$, that have been estimated from the sample [39]. This happens when a non-zero intercept exist in the equation or the data are mean-centered, thus b_0 exists. Otherwise the degrees of freedom are $n-p$. More specifically, when referring to RMSEC, we take into account the number of chemical components in the model or the number of any factors existing, depending on the algorithmic method applied to the data. Therefore, the correct number of degrees of freedom for each of the four calibration methods are the following [37]:

- For CLS the number of dof is equal to the number of samples,n, minus the number of components modeled, c, minus 1 if there isn't a non-zero intercept or minus 2 if there is a non-zero intercept or the data are mean-centered.

$$dof = n - c \qquad or \qquad dof = n - c - 1$$

- For ILS the number of dof is equal to the number of samples minus the number of wavelengths, w, used in the calibration (i.e. the number of columns in the $\mathbf{P}$ matrix) minus 1 if there isn't a non-zero intercept or minus 2 if there is a non-zero intercept or the data are mean-centered.

$$dof = n - w \qquad or \qquad dof = n - w - 1$$

- For PCR the number of dof is equal to the number of samples,n, minus the number of the factors, f, used for the basis space minus 1 if there isn't a non-zero intercept or minus 2 if there is a non-zero intercept or the data are mean-centered.

$$dof = n - f \qquad or \qquad dof = n - f - 1$$

- For PLS the number of dof (approximately) is the number of samples, n, minus the number of factors (latent variables), f, minus 1 if there isn't a non-zero intercept or minus 2 if there is a non-zero intercept or the data are mean-centered.

$$dof = n - f \qquad or \qquad dof = n - f - 1$$

The last RSME error is a more complicated one and it's called the *Root Mean Squared Error of cross-Validation* (RMSECV or RMSEV). This kind of error metric emerges from a cross-validation procedure, also known as *Leave-One-Out Cross-Validation* (LOOCV). During this procedure, the n-sample calibration dataset breaks into a $n-1$-sample training set and an 1-sample test set. The procedure is iterative and it finishes when the calibration set has been split in every possible combination. At each step, an RMSEP error is calculated, as well as a relative RMSEP% 100 and then the RMSEV is the mean of all these errors divided by the size of the dataset, n.

As seen before, another error metric is the **standard error of regression**, $s_{y/x}$, which is a statistic for the estimation of the random errors in the y-direction of a general $y = ax$ calibration model [31] [9] and it is expressed as follows:

$$s_{y/x} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{dof}} \tag{4.50}$$

Apart from the above, other validation tools can be diagrams associated with the depiction of the differences between initial and estimated concentration C (meaning the estimated concentration for each compound individually), as well as initial and estimated Absorbance (using the estimated coefficients $\mathbf{A}_{est} = \mathbf{C}_{init}\mathbf{coeff}_{est}$) spectra. The smaller the differences, the better the performance of the particular algorithm. In the next chapter 5 we will discuss about these differences in absorbance spectra, which concluded in discarding candidate "unwanted" samples from our experimental datasets.

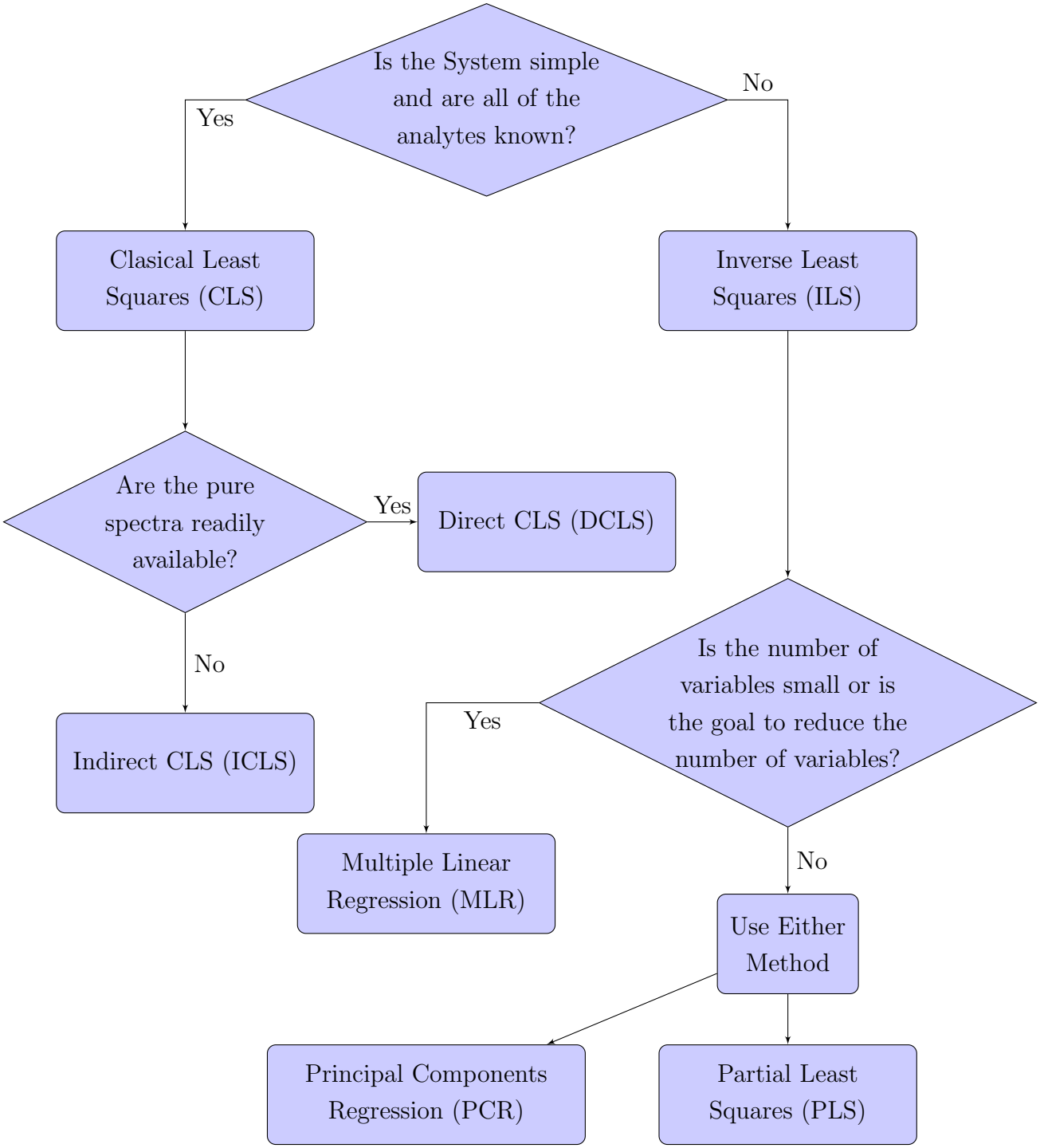


Figure 4.1: Multivariate Calibration Methods Decision Tree

Chapter 5

Implementation and Results

5.1 Software tools

The software tool used to process the experimental, simulated, and hyperspectral image data, implement the algorithmic methods and extract the results was MATLAB® 7.9 (release name R2009b). Also, for the processing of the microscope hyperspectral images and the initial extraction of the hyperspectral cubes we used the ElGreco GUI environment, a software tool developed by George Epitropou in his Diploma Thesis [3].

5.2 Preprocessing

Before applying the algorithmic methods on the data, we insert the useful matrices with the absorbance or transmission mixture spectra and their concentrations into a MATLAB script and perform preprocessing. This preprocessing step includes wavelength selection, which for our thesis was from $450nm$ to $750nm$ (visible spectrum area), curve fitting in order to reduce noise and sampling per $1nm$. Subsequently, we run the algorithms and obtain the results with an accuracy of 4 significant digits. Similarly, for the simulated data, the spectra extend from $0nm$ to $1000nm$, which in real world wouldn't be possible, but for this kind of data it is used for testing reasons. Finally, for the hyperspectral image data, a Wiener filter was applied before the application of the algorithms in order to achieve a smoothing effect (noise suppression) on the images, the gathered spectra were divided by 255 in order to obtain the normalized spectra from the grayscale ones

and the initial transmission spectra were converted into absorbance spectra through 2.2 equation.

5.3 Implementation Details

In this section we are going to briefly describe the implementation methods used for the applications of the various algorithms described in the previous chapter. First of all, as stated before (4.2) the CLS method can be approached through DCLS or ICLS.

More specifically, for DCLS, apart from the simple equation used, another implementation had to do with using the slope of the regression fit between Absorbance and Concentration matrices. For this reason, the Matlab *creatFit.m* custom function was used in order to estimate the matrix of absorptivities (regression coefficients **K**).

As for ICLS, many implementation approaches have been applied, such as CLS with or without a non-zero intercept, data mean-centering, Matlab function mvregress.m and for all of these and especially for the non-mean-centered data or those without a non-zero intercept, the coefficients' (**K**-matrix) estimation could be achieved either through the simple Matlab equation solving ($\mathbf{K} = \mathbf{C}\backslash\mathbf{A}$) or using the *lsqnonneg.m* Matlab function for NNLS (Non-negative Least Squares) or even the following formula to compute the least squares approximate solution [43]:

$$K = (C^\top C)^{-1} C^\top A \qquad or \qquad K = inv(C'C)C'A$$

mathematically and in Matlab, respectively, which translates into the following, using the pseudo-inverse:

$$K = C^+ A \qquad or \qquad K = pinv(C)A$$

In this case, the pseudo-inverse is needed because C may not be a square matrix; therefore if it's a singular (non-square) matrix, its inverse cannot be calculated.

The above implementation approaches have been used for the other methods, too. However, for time-consuming reasons it must be noted that mean-centered data and data with a non-zero intercept export the same results (because in mean-centered data the non-zero intercept is hidden) and for the non-mean-centered data the least squares equation solving also exports the same results either using "pinv", or "nnls, or "\".

Another thing to notice is that we applied only the PLS1 algorithm on the obtained datasets, and not PLS2. The reason for this was that the so-called PLS2 algorithm may be used for the case of more than one column in $\boldsymbol{Y}$. The PLS2 algorithm, however, is more complicated than PLS1 and even when several columns are available in $\boldsymbol{Y}$, it may be preferable to apply PLS1 separately to each column of $\boldsymbol{Y}$ [36].

In the following section, where the results for the various algorithmic methods are presented, some abbreviations are used for brevity, such as **DCLS** for Direct Classical Least Squares, **DCLS(T.S.)** for DCLS through slope, **ICLS** for plain Indirect Classical Least Squares, **ICLS(W.I.)** for ICLS with non-zero intercept intercept (or if the data are mean-centered), **ILS(F.S.)** for plain Inverse Least Squares through forward selection, **ILS(F.S.W.I.)** for ILS through forward selection with non-zero intercept, **ILS(B.E.)** for plain ILS through backward elimination, **ILS(B.E.W.I.)** for ILS through backward elimination with non-zero intercept, **PCR** for plain Principal Components Regression, **PCR(M.C.)** with mean-centered data, **PLS** for Partial Least Squares regression of type 1 and 2, and **PLS(M.C.)** for the mean-centered version of PLS.
For more implementation details and more a analytical explanation of the algorithms look up [41] and [36].

Last but not least, because the data from the microscope are extracted in the form of transmission spectra, before applying the algorithmic methods associated with Absoption spectroscopy we need to make a conversion from transmission to absorbance, using the following equation:

$$Abs = log\frac{1}{\%T} \tag{5.1}$$

5.4 Data from Cary: Results

In the following sections, the performance results for the various methods and their algorithms are presented. It should be noted that before extracting the results using the estimation errors, the simple CLS algorithm was applied on the datasets obtained from the spectrophotometer and the diagrams for the Absorbance versus the estimated Absorbance spectra for each sample of each dataset were presented. In doing so, it could be possible to understand whether the absorbance spectra were estimated correctly for each sample and which sample of the specific dataset enclosed the bigger error (during data acquisition).

Thus, this sample was removed. In figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9,the diagrams for the A-A_{est} are presented. Observing these diagrams we extracted the following samples: sample number 8 from Methylene Blue - Fast Green dataset, sample number 4 from $CoCl_2$ - Methylene Blue dataset and sample number 8 from Thymol - Malachite Green - Methylene Blue dataset. It might be interesting to note that although removing these specific samples the results for the corresponding dataset were improved, it is not so clear if the biggest error is connected with this sample specifically or with the fact that the datasets contain a few samples.

Following the figures, the tables with the results from the estimation errors for each dataset are presented (tables 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9), along with some histograms(5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18) that better represent the algorithms' performance and the ones tha yield the smallest errors.

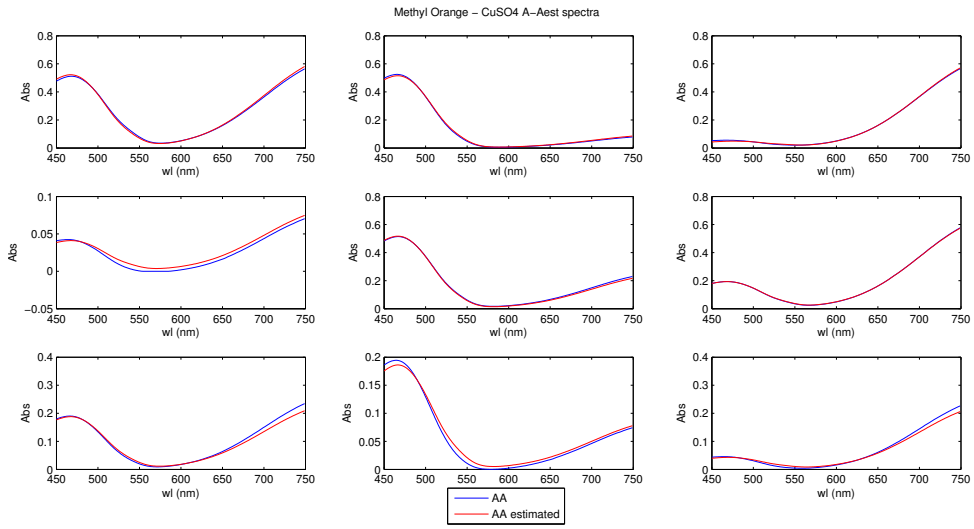


Figure 5.1: Methyl Orange - $CuSO_4$, A-A_{est} spectra

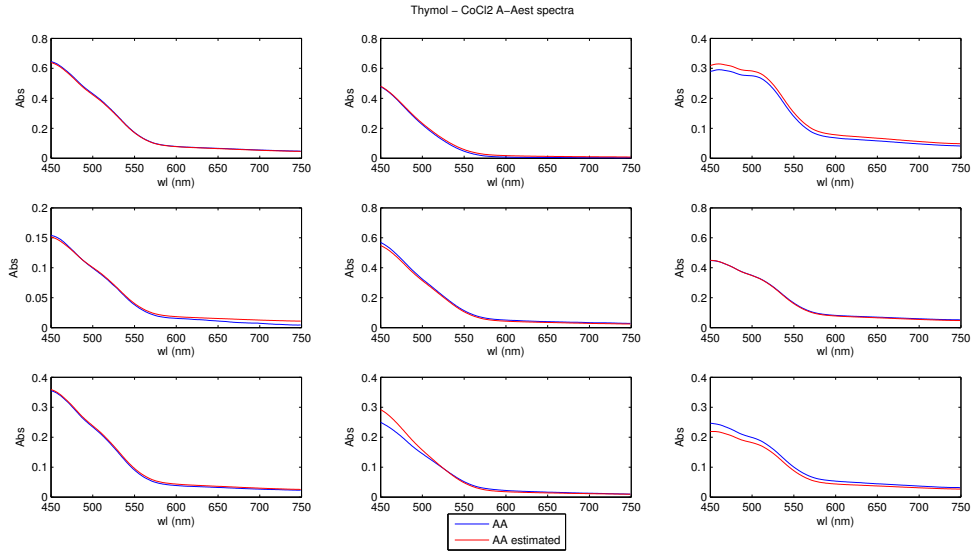


Figure 5.2: Thymol - $CoCl_2$, A-A_{est} spectra

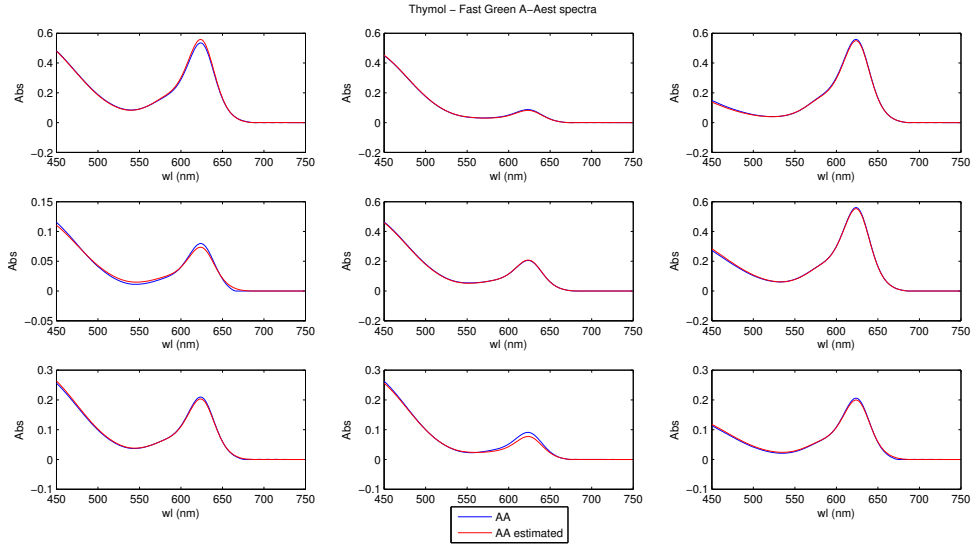


Figure 5.3: Thymol - Fast Green, A-A_{est} spectra

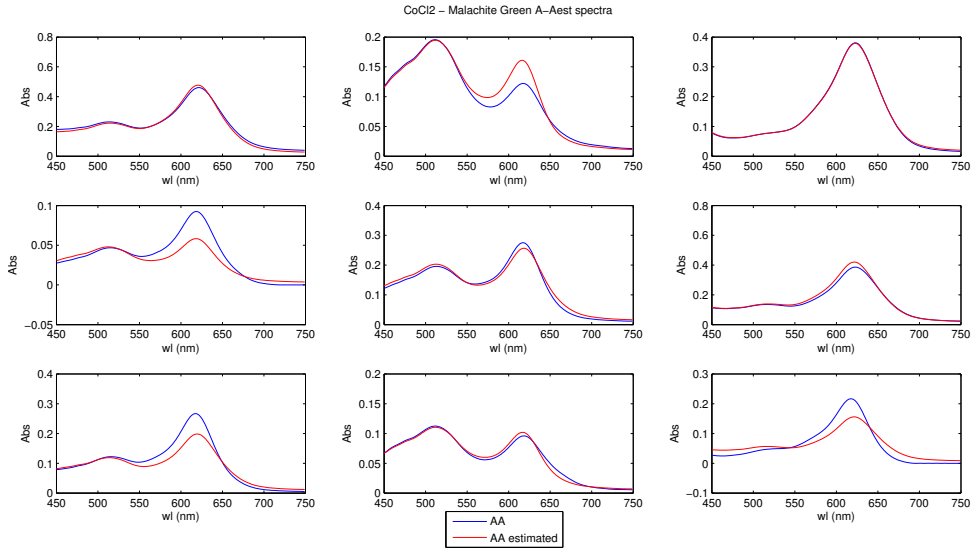


Figure 5.4: $CoCl_2$ - Malachite Green, A-A_{est} spectra

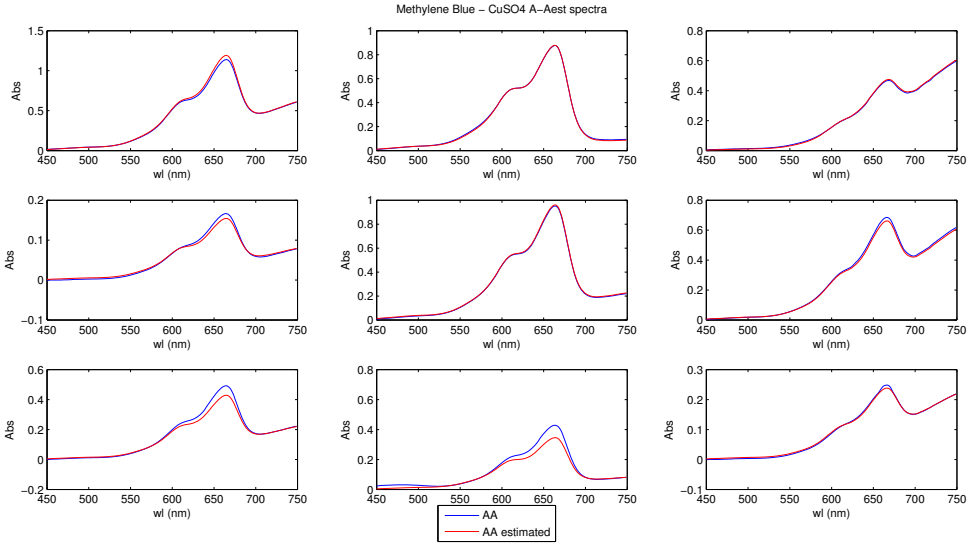


Figure 5.5: Methylene Blue - $CuSO_4$, A-A_{est} spectra

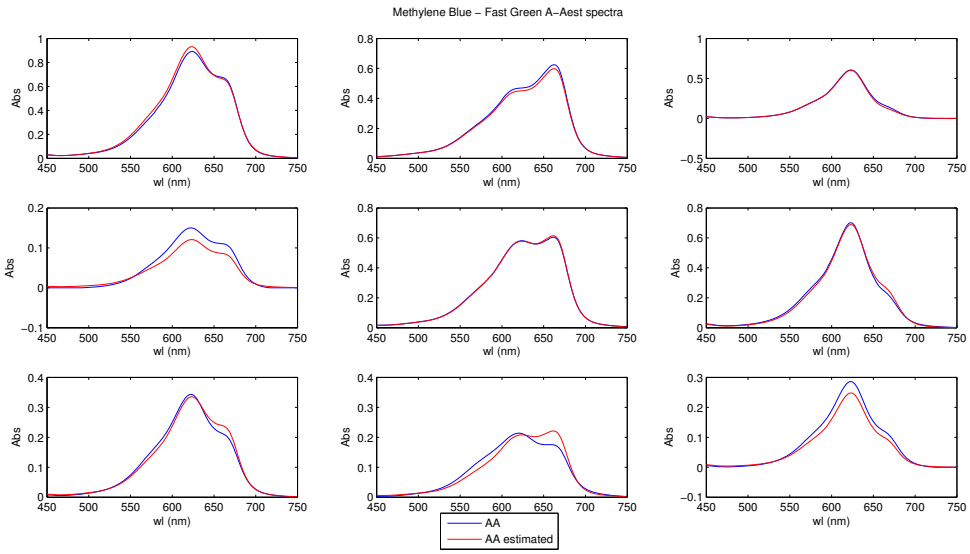


Figure 5.6: Methylene Blue - Fast Green, A-A_{est} spectra

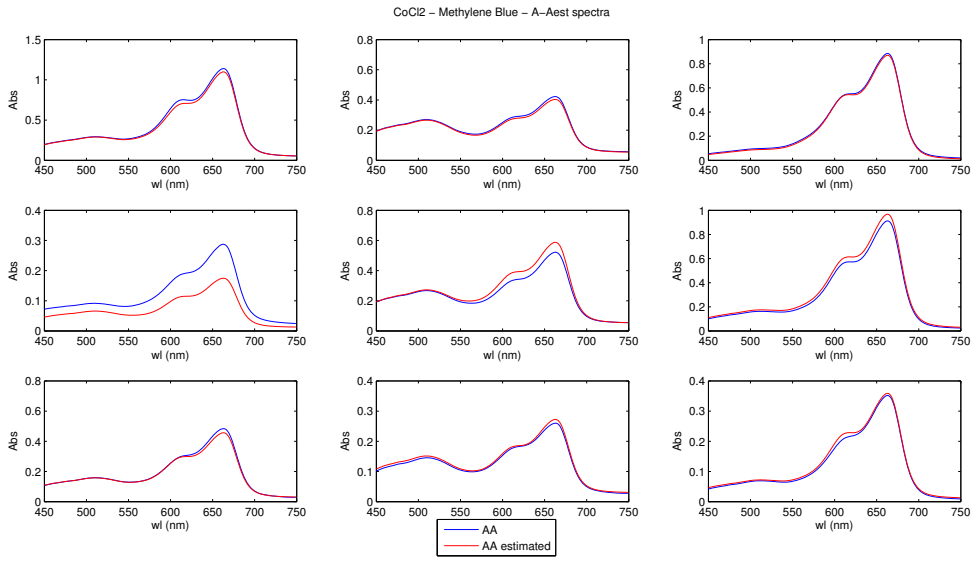


Figure 5.7: $CoCl_2$ - Methylene Blue, A-A_{est} spectra

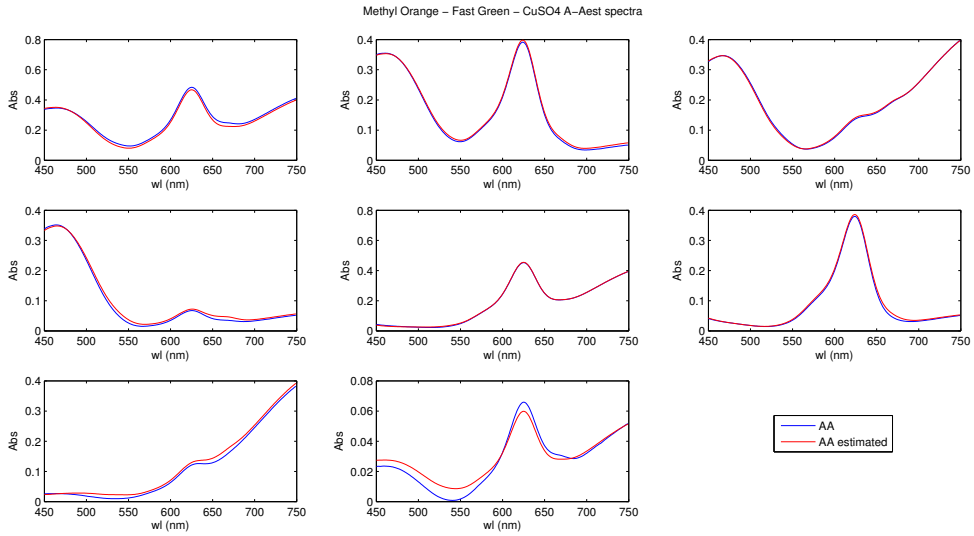


Figure 5.8: Methyl Orange - Fast Green - $CuSO_4$, A-A_{est} spectra

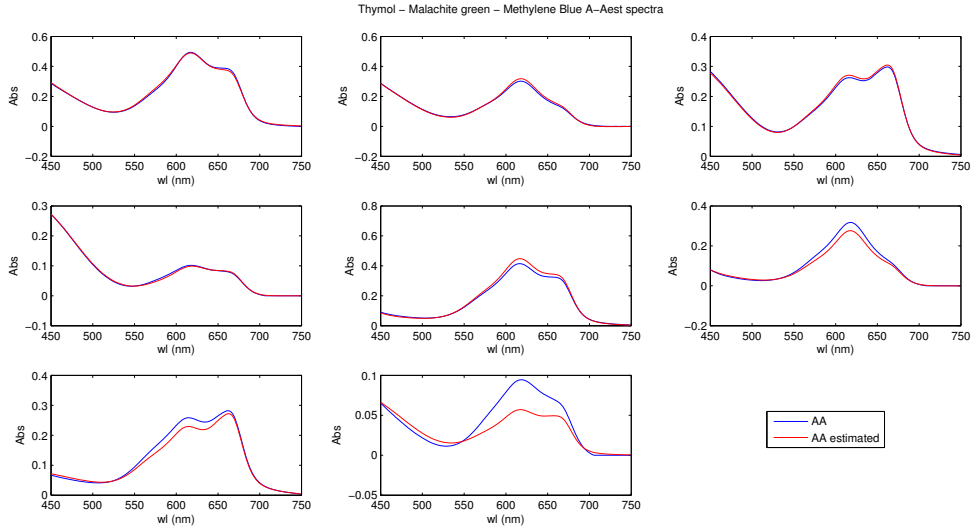


Figure 5.9: Thymol - Malachite Green - Methylene Blue, A-A_{est} spectra

Table 5.1: Dataset 1, 2 factors - RMSE and standard deviation errors

Methyl Or.-$CuSO_4$	RMSEP		RMSEC		RMSEV	
	C1	C2	C1	C2	C1	C2
ICLS	1.1×10^{-7}	0.001228	1.248×10^{-7}	0.001392	1.347×10^{-7}	0.001323
ICLS(W.I.)	1.062×10^{-7}	0.001175	1.301×10^{-7}	0.001439	1.477×10^{-7}	0.001583
DCLS(T.S.)	2.095×10^{-6}	0.001346	2.565×10^{-6}	0.001649		
ILS(F.S.)	1.091×10^{-7}	2.117×10^{-8}	3.274×10^{-7}	6.35×10^{-8}	1.487×10^{-6}	2.884×10^{-7}
ILS(F.S.W.I.)	7.483×10^{-8}	4.507×10^{-5}	2.245×10^{-7}	0.0001352	7.113×10^{-7}	0.0002009
ILS(B.E.)	8.243×10^{-8}	3.543×10^{-6}	2.473×10^{-7}	1.063×10^{-5}	2.304×10^{-6}	9.904×10^{-5}
ILS(B.E.W.I.)	5.409×10^{-8}	0.000514	1.623×10^{-7}	0.001542	2.551×10^{-7}	0.002444
PCR	1.1×10^{-7}	0.001226	1.247×10^{-7}	0.00139	1.344×10^{-7}	0.001319
PCR(M.C.)	1.062×10^{-7}	0.001173	1.301×10^{-7}	0.001437	4.19×10^{-7}	0.001837
PLS	1.099×10^{-7}	0.001226	1.247×10^{-7}	0.00139	1.340×10^{-7}	0.001319
PLS(M.C.)	1.062×10^{-7}	0.001173	1.3×10^{-7}	0.001437	8.983×10^{-8}	0.001061

Methyl Or.-$CuSO_4$	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
	C1	C2	C1	C2	C1	C2	$s_{y/x}$
ICLS	4.189	7.726	4.75	8.761	3.557	8.167	0.1268
ICLS(W.I.)	4.343	10.55	5.319	12.92	4.208	12.15	0.1133
DCLS(T.S.)	27.83	7.528	34.09	9.22			0.5374
ILS(F.S.)	3.017	0.0001038	9.05	0.0003113	37.22	0.002819	
ILS(F.S.W.I.)	2.082	0.3283	6.245	0.985	9.917	1.508	
ILS(B.E.)	3.631	0.02185	10.89	0.06556	47.77	0.7001	
ILS(B.E.W.I.)	0.873	4.926	2.619	14.78	7.184	13.64	
PCR	4.247	7.565	4.816	3.552	8.02	2.673	
PCR(M.C.)	4.409	10.06	5.4	12.32	9.826	8.718	
PLS	4.246	7.569	4.815	8.582	3.543	8.08	
PLS(M.C.)	4.406	10.06	5.397	12.32	2.648	8.018	

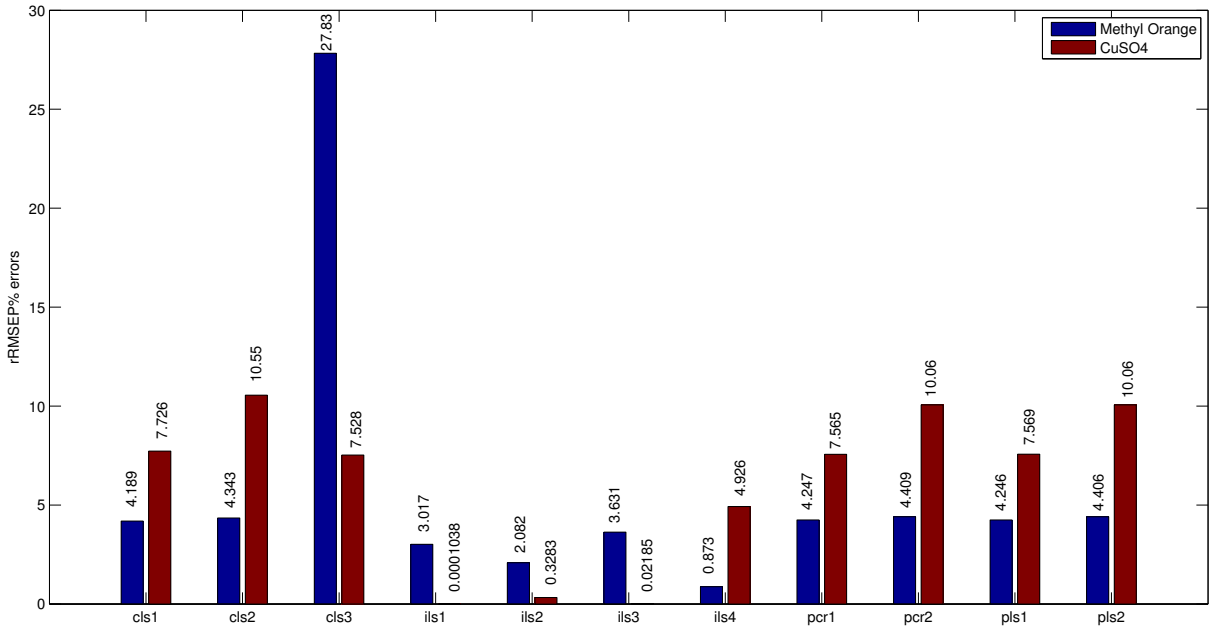


Figure 5.10: rRMSEP(%) error performance representation for Methyl Orange-$CuSO_4$ dataset

Table 5.2: Dataset 2, 2 factors - RMSE and standard deviation errors

Thymol-$CoCl_2$	RMSEP		RMSEC		RMSEV	
	C1	C2	C1	C2	C1	C2
ICLS	1.789×10^{-6}	0.001932	2.028×10^{-6}	0.00219	1.536×10^{-6}	0.002483
ICLS(W.I.)	1.76×10^{-6}	0.001911	2.156×10^{-6}	0.00234	1.877×10^{-6}	0.002619
DCLS(T.S.)	1.893×10^{-6}	0.00192	2.319×10^{-6}	0.002352		
ILS(F.S.)	6.732×10^{-7}	4.255×10^{-7}	2.02×10^{-6}	1.276×10^{-6}	8.963×10^{-6}	5.665×10^{-6}
ILS(F.S.W.I.)	2.194×10^{-7}	3.594×10^{-5}	6.581×10^{-7}	0.0001078	1.159×10^{-6}	0.0001644
ILS(B.E.)	1.121×10^{-6}	1.558×10^{-5}	3.362×10^{-6}	4.674×10^{-5}	2.41×10^{-5}	0.0003351
ILS(B.E.W.I.)	1.077×10^{-6}	0.0001255	3.23×10^{-6}	0.0003765	9.007×10^{-6}	0.0005762
PCR	1.771×10^{-6}	0.001915	2.008×10^{-6}	0.002172	1.436×10^{-6}	0.002386
PCR(M.C.)	1.741×10^{-6}	0.001889	2.132×10^{-6}	0.002313	1.977×10^{-6}	0.002206
PLS	1.77×10^{-6}	0.001897	2.007×10^{-6}	0.002151	1.427×10^{-6}	0.002369
PLS(M.C.)	1.739×10^{-6}	0.001871	2.13×10^{-6}	0.002291	1.237×10^{-6}	0.001573

Thymol-$CoCl_2$	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		$s_{y/x}$
	C1	C2	C1	C2	C1	C2	
ICLS	8.762	14.64	9.935	16.6	7.637	16.73	0.1709
ICLS(W.I.)	9.549	14.53	11.69	17.8	10.66	18.07	0.1809
DCLS(T.S.)	11.84	12.08	14.5	14.79			0.2307
ILS(F.S.)	3.305	0.003187	9.916	0.00956	67.12	0.03261	
ILS(F.S.W.I.)	1.85	0.1426	5.551	0.4279	5.895	0.9594	
ILS(B.E.)	8.873	0.1583	26.62	0.4749	154.5	1.949	
ILS(B.E.W.I.)	7.402	0.673	22.21	2.019	36.48	3.527	
PCR	9.181	13.15	10.41	14.91	7.881	15.89	
PCR(M.C.)	10.31	13.19	12.63	16.16	13.03	13.74	
PLS	9.171	13.06	10.4	14.81	7.871	15.8	
PLS(M.C.)	10.3	13.1	12.62	16.05	7.683	10.31	

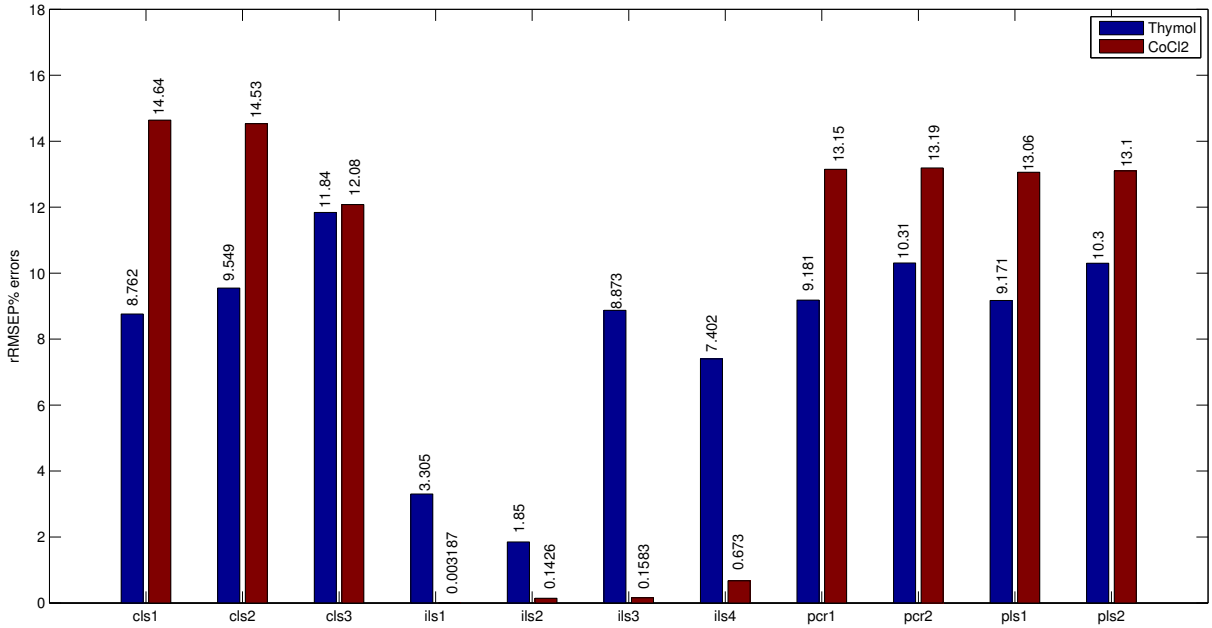


Figure 5.11: rRMSEP(%) error performance representation for Thymol-$CoCl_2$ dataset

Table 5.3: Dataset 3, 2 factors - RMSE and standard deviation errors

	RMSEP		RMSEC		RMSEV	
Thymol-Fast Green	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
ICLS	5.833×10^{-7}	9.896×10^{-8}	6.614×10^{-7}	1.122×10^{-7}	7.264×10^{-7}	1.087×10^{-7}
ICLS(W.I.)	5.72×10^{-7}	7.38×10^{-8}	7.005×10^{-7}	9.038×10^{-8}	7.985×10^{-7}	1.046×10^{-7}
DCLS(T.S.)	1.473×10^{-6}	1.195×10^{-7}	1.804×10^{-6}	1.464×10^{-7}		
ILS(F.S.)	1.941×10^{-7}	7.04×10^{-8}	5.824×10^{-7}	2.112×10^{-7}	2.486×10^{-5}	1.724×10^{-6}
ILS(F.S.W.I.)	4.147×10^{-8}	4.879×10^{-9}	1.244×10^{-7}	1.464×10^{-8}	2.011×10^{-7}	3.572×10^{-8}
ILS(B.E.)	8.491×10^{-10}	2.568×10^{-9}	2.547×10^{-9}	7.704×10^{-9}	1.378×10^{-8}	4.168×10^{-8}
ILS(B.E.W.I.)	3.971×10^{-8}	2.433×10^{-8}	1.191×10^{-7}	7.3×10^{-8}	3.031×10^{-7}	1.558×10^{-7}
PCR	5.829×10^{-7}	9.885×10^{-8}	6.61×10^{-7}	1.121×10^{-7}	7.304×10^{-7}	1.067×10^{-7}
PCR(M.C.)	5.714×10^{-7}	7.374×10^{-8}	6.998×10^{-7}	9.032×10^{-8}	9.563×10^{-7}	1.175×10^{-7}
PLS	5.829×10^{-7}	9.882×10^{-8}	6.609×10^{-7}	1.121×10^{-7}	7.304×10^{-7}	1.064×10^{-7}
PLS(M.C.)	5.713×10^{-7}	7.374×10^{-8}	6.997×10^{-7}	9.031×10^{-8}	5.176×10^{-7}	6.383×10^{-8}

	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
Thymol-Fast Green	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**	$\mathbf{s_{y/x}}$
ICLS	3.904	7.308	4.426	8.287	4.25	6.19	0.0831
ICLS(W.I.)	3.903	6.294	4.781	7.709	4.905	6.879	0.06986
DCLS(T.S.)	10.95	13.54	13.42	16.58			0.2869
ILS(F.S.)	1.305	5.202	3.914	15.61	2.259	8.616	
ILS(F.S.W.I.)	0.1863	0.4613	0.5589	1.384	0.9159	1.205	
ILS(B.E.)	0.006095	0.249	0.01829	0.7471	0.06516	2.619	
ILS(B.E.W.I.)	0.2208	1.396	0.6623	4.189	1.748	12.79	
PCR	3.964	7.535	4.494	8.544	4.303	6.3	
PCR(M.C.)	3.974	6.362	4.867	7.792	6.449	7.587	
PLS	3.963	7.532	4.494	8.54	4.301	6.27	
PLS(M.C.)	3.973	6.361	4.866	7.791	3.13	4.539	

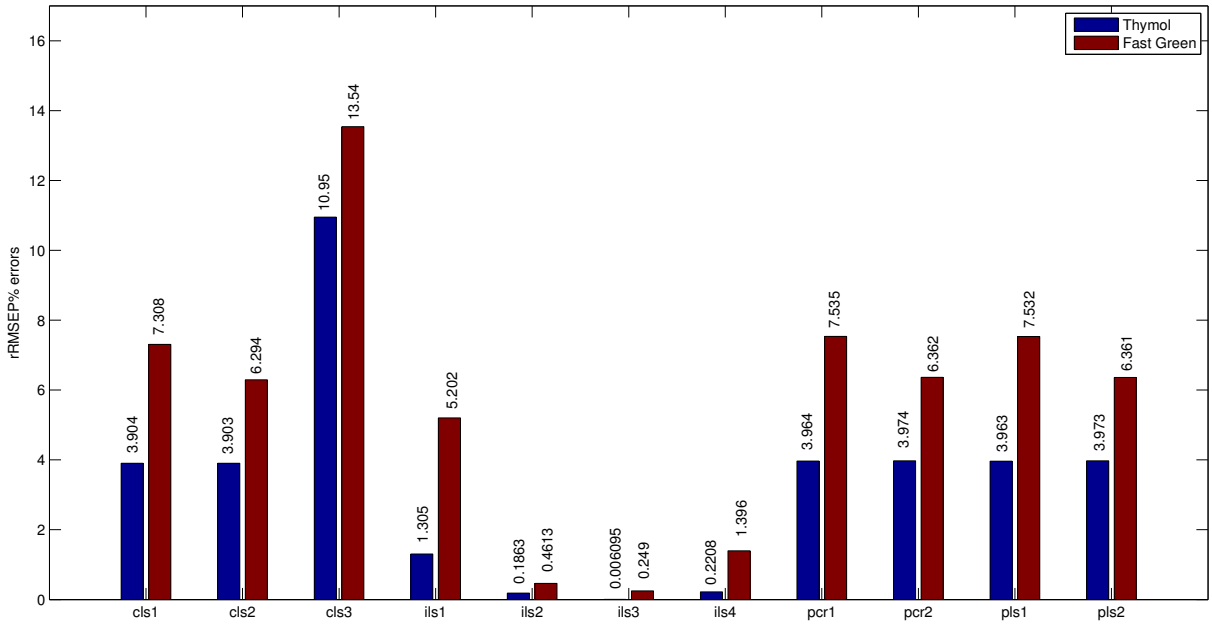


Figure 5.12: rRMSEP(%) error performance representation for Thymol-Fast Green dataset

Table 5.4: Dataset 4, 2 factors - RMSE and standard deviation errors

$CoCl_2$-Malachite Green	RMSEP		RMSEC		RMSEV	
	C1	C2	C1	C2	C1	C2
ICLS	0.001073	8.619×10^{-7}	0.001217	9.773×10^{-7}	0.0009893	8.525×10^{-7}
ICLS(W.I.)	0.001058	6.901×10^{-7}	0.001295	8.452×10^{-7}	0.001319	8.741×10^{-7}
DCLS(T.S.)	0.006491	2.291×10^{-6}	0.00795	2.805×10^{-6}		
ILS(F.S.)	4.159×10^{-6}	8.264×10^{-8}	1.248×10^{-5}	2.479×10^{-7}	7.802×10^{-5}	1.55×10^{-6}
ILS(F.S.W.I.)	2.771×10^{-5}	3.014×10^{-7}	8.314×10^{-5}	9.043×10^{-7}	0.0001239	1.911×10^{-6}
ILS(B.E.)	2.378×10^{-5}	1.753×10^{-7}	7.134×10^{-5}	5.26×10^{-7}	0.00079	5.824×10^{-6}
ILS(B.E.W.I.)	8.328×10^{-5}	9.318×10^{-8}	0.0002498	2.795×10^{-7}	0.0004189	5.833×10^{-7}
PCR	0.001254	8.987×10^{-7}	0.001422	1.019×10^{-6}	0.001087	9.626×10^{-7}
PCR(M.C.)	0.001147	7.056×10^{-7}	0.001405	8.642×10^{-7}	0.001309	8.344×10^{-7}
PLS	0.001156	8.07×10^{-7}	0.00131	9.151×10^{-7}	0.000985	8.726×10^{-7}
PLS(M.C.)	0.00106	6.476×10^{-7}	0.001298	7.932×10^{-7}	0.0008697	5.671×10^{-7}

$CoCl_2$-Malachite Green	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
	C1	C2	C1	C2	C1	C2	$s_{y/x}$
ICLS	9.621	52.88	10.91	59.96	6.315	41.3	0.2978
ICLS(W.I.)	8.134	42.05	9.962	51.5	8.477	38.3	0.2591
DCLS(T.S.)	30.24	35.85	37.04	43.91			1.212
ILS(F.S.)	0.03732	8.002	0.112	24.01	0.4831	55.74	
ILS(F.S.W.I.)	0.2655	21.27	0.7966	63.81	0.7462	108.9	
ILS(B.E.)	0.2316	18.72	0.6947	56.17	4.74	128.1	
ILS(B.E.W.I.)	0.6875	9.686	2.062	29.06	2.68	23.6	
PCR	11.21	48.72	12.72	55.24	7.444	41.2	
PCR(M.C.)	8.801	36.06	10.78	44.17	10.2	33.04	
PLS	10.26	46.97	11.63	53.26	6.449	38.85	
PLS(M.C.)	8.076	34.51	9.89	42.27	5.917	23.51	

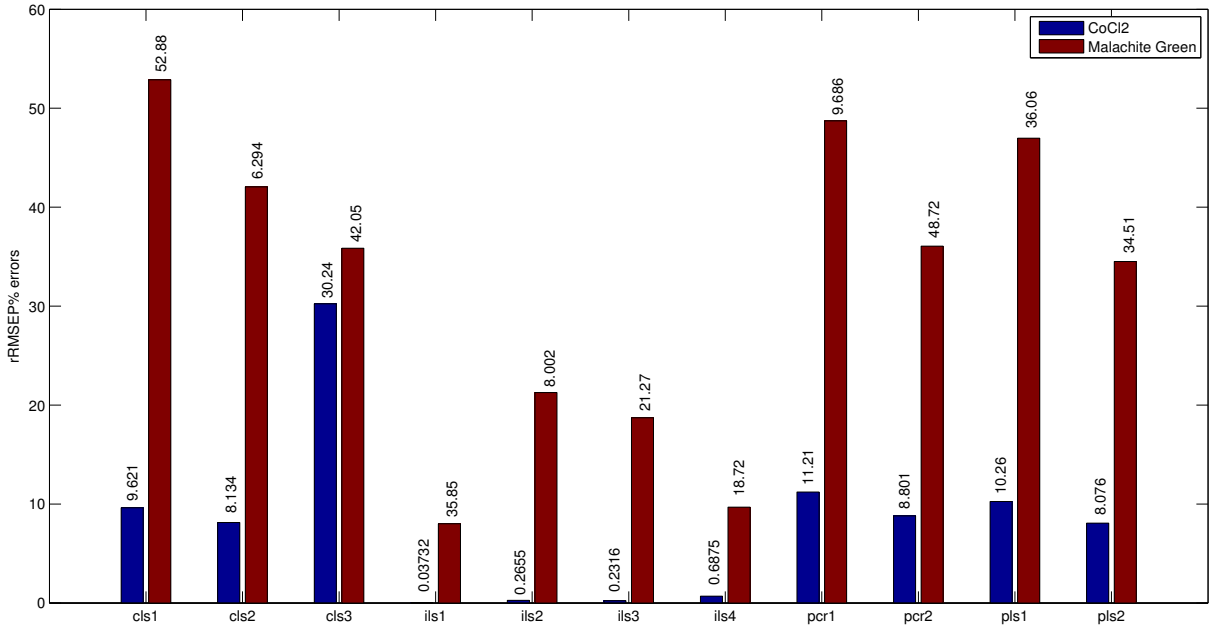


Figure 5.13: rRMSEP(%) error performance representation for $CoCl_2$-Malachite Green dataset

Table 5.5: Dataset 5, 2 factors - RMSE and standard deviation errors

Meth. Blue-$CuSO_4$	RMSEP		RMSEC		RMSEV	
	C1	C2	C1	C2	C1	C2
ICLS	5.629×10^{-7}	0.0005446	6.383×10^{-7}	0.0006175	4.59×10^{-7}	0.000644
ICLS(W.I.)	4.119×10^{-7}	0.0005424	5.044×10^{-7}	0.0006643	5.002×10^{-7}	0.0006828
DCLS(T.S.)	2.684×10^{-6}	0.00281	3.288×10^{-6}	0.003442		
ILS(F.S.)	2.865×10^{-7}	2.253×10^{-6}	8.595×10^{-7}	6.76×10^{-6}	6.458×10^{-6}	5.079×10^{-5}
ILS(F.S.W.I.)	1.359×10^{-7}	0.000163	4.077×10^{-7}	0.0004889	5.451×10^{-7}	0.0006227
ILS(B.E.)	4.827×10^{-8}	4.062×10^{-7}	1.448×10^{-7}	1.219×10^{-6}	8.464×10^{-7}	7.121×10^{-6}
ILS(B.E.W.I.)	5.975×10^{-8}	1.766×10^{-5}	1.793×10^{-7}	5.297×10^{-5}	4.664×10^{-7}	0.000178
PCR	5.622×10^{-7}	0.0005441	6.375×10^{-7}	0.000617	4.994×10^{-7}	0.0006429
PCR(M.C.)	4.111×10^{-7}	0.0005421	5.035×10^{-7}	0.0006639	6.417×10^{-7}	0.001178
PLS	5.598×10^{-7}	0.000544	6.348×10^{-7}	0.0006169	4.979×10^{-7}	0.000643
PLS(M.C.)	4.101×10^{-7}	0.000542	5.022×10^{-7}	0.0006638	3.538×10^{-7}	0.0004226

Meth. Blue-$CuSO_4$	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
	C1	C2	C1	C2	C1	C2	$s_{y/x}$
ICLS	9.965	3.232	11.3	3.665	7.606	3.322	0.2893
ICLS(W.I.)	14.16	3.305	17.34	4.048	14.36	3.636	0.2291
DCLS(T.S.)	64.74	16.92	79.3	20.72			1.306
ILS(F.S.)	10.66	0.02162	31.98	0.06487	157.3	0.4784	
ILS(F.S.W.I.)	4.03	1.431	12.09	4.293	15.59	5.356	
ILS(B.E.)	1.441	0.003527	4.322	0.01058	26.36	0.05095	
ILS(B.E.W.I.)	2.659	0.1027	7.976	0.3081	14.16	1.521	
PCR	9.832	3.237	11.15	3.67	8.26	3.324	
PCR(M.C.)	13.4	3.267	16.41	4.002	13.23	6.8	
PLS	9.786	3.238	11.1	3.671	7.70	3.325	
PLS(M.C.)	13.36	3.267	16.37	4.002	9.703	2.233	

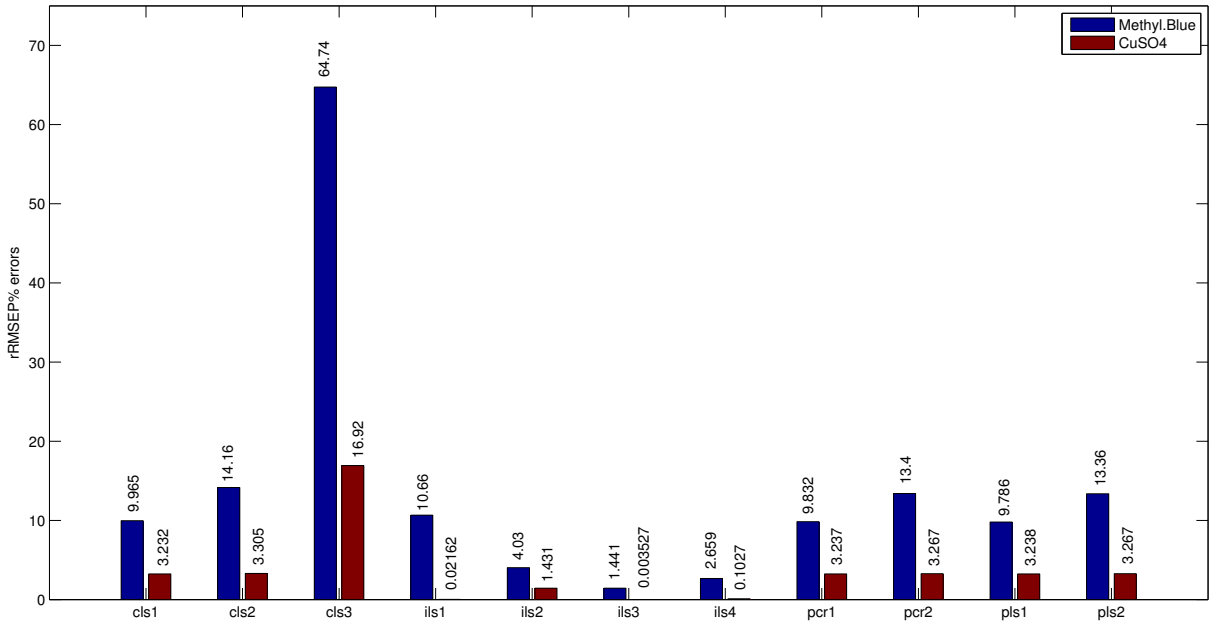


Figure 5.14: rRMSEP(%) error performance representation for Methylene Blue-$CuSO_4$ dataset

Table 5.6: Dataset 6, 2 factors - RMSE and standard deviation errors

Meth. Blue-Fast Green	RMSEP		RMSEC		RMSEV	
	C1	C2	C1	C2	C1	C2
ICLS	5.355×10^{-7}	2.891×10^{-7}	6.184×10^{-7}	3.338×10^{-7}	7.08×10^{-7}	3.467×10^{-7}
ICLS(W.I.)	5.187×10^{-7}	2.139×10^{-7}	6.561×10^{-7}	2.706×10^{-7}	7.638×10^{-7}	2.898×10^{-7}
DCLS(T.S.)	1.646×10^{-6}	6.041×10^{-7}	2.082×10^{-6}	7.641×10^{-7}		
ILS(F.S.)	3.166×10^{-8}	2.591×10^{-10}	8.954×10^{-8}	7.328×10^{-10}	4.458×10^{-7}	3.648×10^{-9}
ILS(F.S.W.I.)	2.167×10^{-8}	2.074×10^{-8}	6.13×10^{-8}	5.866×10^{-8}	9.134×10^{-8}	8.055×10^{-8}
ILS(B.E.)	2.759×10^{-9}	2.002×10^{-9}	7.803×10^{-9}	5.663×10^{-9}	3.944×10^{-8}	2.863×10^{-8}
ILS(B.E.W.I.)	5.671×10^{-8}	4.248×10^{-8}	1.604×10^{-7}	1.202×10^{-7}	2.751×10^{-7}	1.835×10^{-7}
PCR	5.369×10^{-7}	2.892×10^{-7}	6.2×10^{-7}	3.339×10^{-7}	7.077×10^{-7}	3.66×10^{-7}
PCR(M.C.)	5.195×10^{-7}	2.15×10^{-7}	6.572×10^{-7}	2.719×10^{-7}	7.162×10^{-7}	3.333×10^{-7}
PLS	5.327×10^{-7}	2.864×10^{-7}	6.151×10^{-7}	3.307×10^{-7}	7.034×10^{-7}	3.38×10^{-7}
PLS(M.C.)	5.135×10^{-7}	2.12×10^{-7}	6.495×10^{-7}	2.681×10^{-7}	4.747×10^{-7}	1.79×10^{-7}

Meth. Blue-Fast Green	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
	C1	C2	C1	C2	C1	C2	$s_{y/x}$
ICLS	17.97	13.05	20.75	15.06	18.76	14.79	0.2528
ICLS(W.I.)	14.74	11.69	18.64	14.78	18.54	12.67	0.1887
DCLS(T.S.)	21.55	44.63	27.26	56.46			0.8809
ILS(F.S.)	1.328	0.01927	3.755	0.05451	10.37	0.2136	
ILS(F.S.W.I.)	0.9563	1.722	2.705	4.871	1.714	6.611	
ILS(B.E.)	0.1179	0.189	0.3334	0.5347	0.7398	1.373	
ILS(B.E.W.I.)	2.389	1.943	6.756	5.497	5.04	6.12	
PCR	19.01	13.86	21.95	16	19.32	16.61	
PCR(M.C.)	15.98	11.53	20.21	14.59	23.11	15.12	
PLS	18.87	13.78	21.79	15.92	19.28	16.53	
PLS(M.C.)	15.78	11.37	19.96	14.39	12.14	8.301	

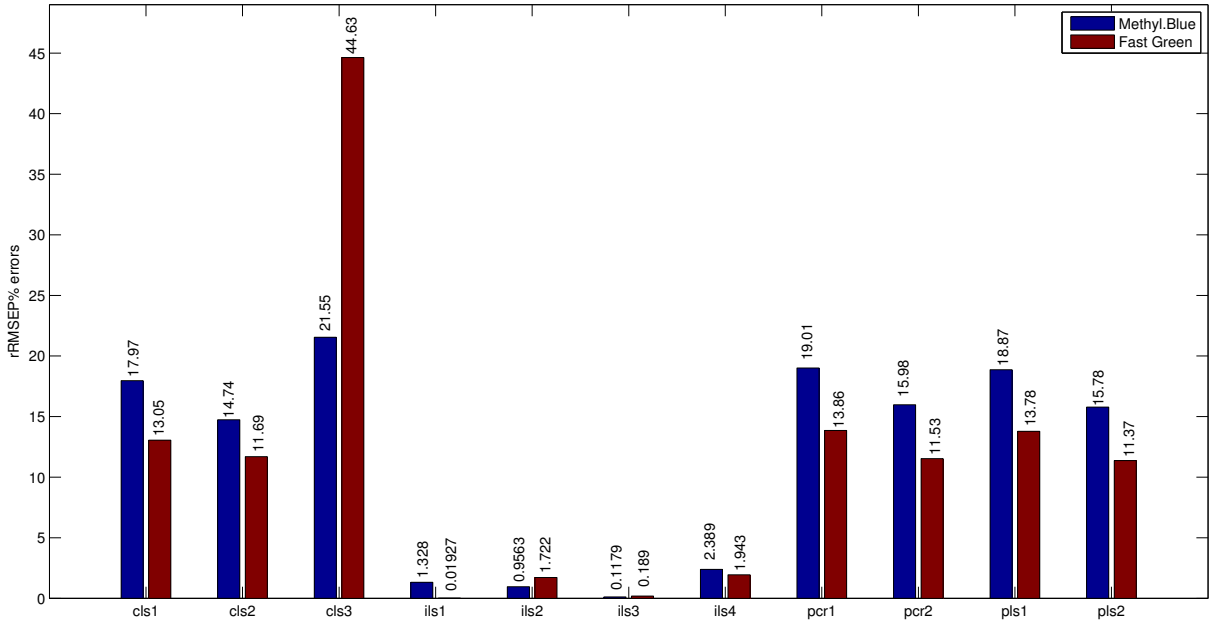


Figure 5.15: rRMSEP(%) error performance representation for Methylene Blue-Fast Green dataset

Table 5.7: Dataset 7, 2 factors - RMSE and standard deviation errors

$CoCl_2$-Meth.Blue	RMSEP		RMSEC		RMSEV	
	C1	C2	C1	C2	C1	C2
ICLS	0.0009304	6.507×10^{-7}	0.001074	7.514×10^{-7}	0.001136	7.176×10^{-7}
ICLS(W.I.)	0.0008868	6.413×10^{-7}	0.001122	8.111×10^{-7}	0.001141	8.4×10^{-7}
DCLS(T.S.)	0.002362	3.977×10^{-6}	0.002988	5.03×10^{-6}		
ILS(F.S.)	1.588×10^{-8}	2.837×10^{-8}	4.493×10^{-8}	8.023×10^{-8}	2.085×10^{-7}	3.723×10^{-7}
ILS(F.S.W.I.)	0.0001862	6.51×10^{-7}	0.0005266	1.841×10^{-6}	0.001046	2.247×10^{-6}
ILS(B.E.)	3.56×10^{-7}	2.476×10^{-7}	1.007×10^{-6}	7.002×10^{-7}	8.719×10^{-6}	6.063×10^{-6}
ILS(B.E.W.I.)	0.0001828	8.047×10^{-8}	0.0005169	2.276×10^{-7}	0.0009089	4.586×10^{-7}
PCR	0.0009293	6.491×10^{-7}	0.001073	7.495×10^{-7}	0.001119	7.198×10^{-7}
PCR(M.C.)	0.0008818	6.356×10^{-7}	0.001115	8.039×10^{-7}	0.001306	8.195×10^{-7}
PLS	0.000928	6.48×10^{-7}	0.001072	7.482×10^{-7}	0.001123	7.1969×10^{-7}
PLS(M.C.)	0.0008805	6.345×10^{-7}	0.001114	8.026×10^{-7}	0.0006367	5.13×10^{-7}

$CoCl_2$-Meth.Blue	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
	C1	C2	C1	C2	C1	C2	$s_{y/x}$
ICLS	7.274	11.23	8.4	12.97	7.676	11.96	0.3534
ICLS(W.I.)	7.157	11.39	9.052	14.4	7.851	13.41	0.3814
DCLS(T.S.)	19.67	113.3	24.88	143.3			1.838
ILS(F.S.)	0.000112	0.9701	0.0003167	2.744	0.001442	6.169	
ILS(F.S.W.I.)	1.042	23.72	2.948	67.1	8.215	62.47	
ILS(B.E.)	0.003265	4.48	0.009234	12.67	0.03665	233	
ILS(B.E.W.I.)	1.571	3.155	4.442	8.925	6.147	8.361	
PCR	7.207	11.53	8.322	13.31	7.509	12.41	
PCR(M.C.)	7.16	12.19	9.056	15.42	9.72	18.07	
PLS	7.197	11.51	8.311	13.3	7.499	12.38	
PLS(M.C.)	7.147	12.16	9.04	15.39	4.252	9.516	

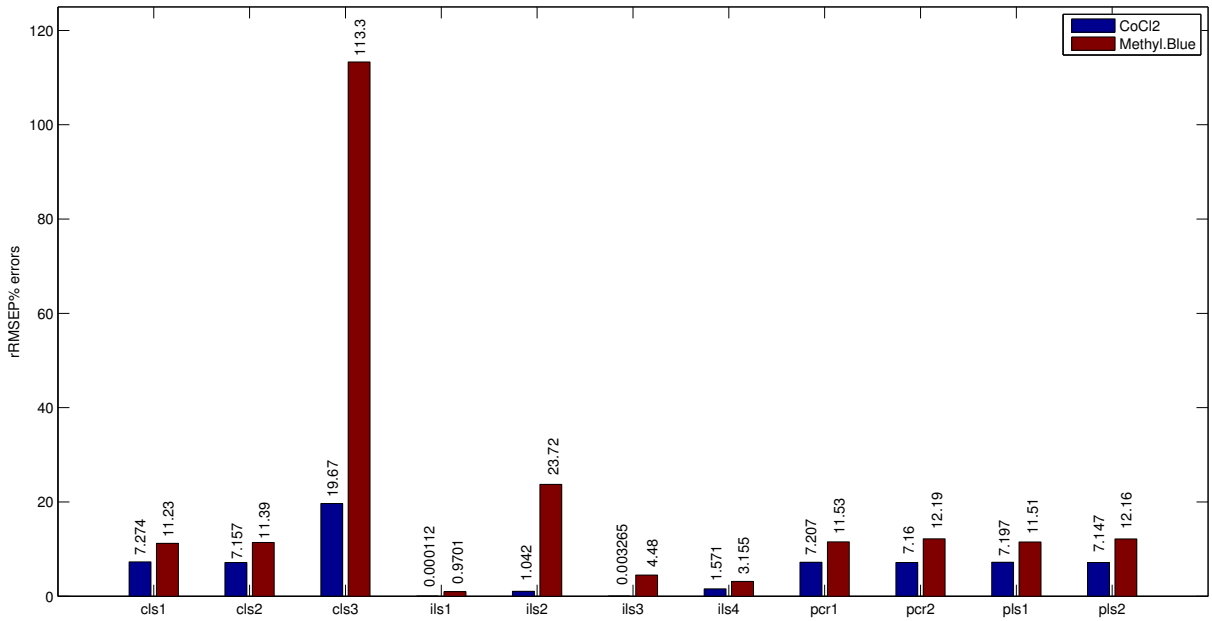


Figure 5.16: rRMSEP(%) error performance representation for $CoCl_2$-Methylene Blue dataset

Table 5.8: Dataset 8, 3 factors - RMSE and standard deviation errors

Methyl Or.-Fast Green-$CuSO_4$	RMSEP			rRMSEP(%)		
	C1	C2	C3	C1	C2	C3
ICLS	1.606×10^{-7}	9.586×10^{-8}	0.0008746	11.95	10.83	9.416
ICLS(W.I.)	7.79×10^{-8}	8.217×10^{-8}	0.0006648	4.817	12.04	10.14
DCLS(T.S.)	1.641×10^{-6}	4.786×10^{-7}	0.002068	25.59	38.88	11.94
ILS(F.S.)	7.051×10^{-8}	4.441×10^{-7}	8.789×10^{-8}	5.504	52.48	0.001481
ILS(F.S.W.I.)	4.155×10^{-6}	9.064×10^{-8}	4.373×10^{-5}	198.7	12.92	0.7464
ILS(B.E.)	7.408×10^{-8}	3.136×10^{-8}	1.629×10^{-6}	5.106	5.216	0.03224
ILS(B.E.W.I.)	7.351×10^{-8}	1.534×10^{-8}	0.0001469	5.228	2.525	2.849
PCR	1.603×10^{-7}	9.588×10^{-8}	0.0008746	12.02	10.87	9.497
PCR(M.C.)	7.79×10^{-8}	8.211×10^{-8}	0.000664	4.873	12.36	10.45
PLS	1.592×10^{-7}	9.507×10^{-8}	0.0008656	11.95	10.78	9.402
PLS(M.C.)	7.781×10^{-8}	8.176×10^{-8}	0.0006607	4.865	12.29	10.39

Methyl Or.-Fast Green-$CuSO_4$	RMSEC			rRMSEC(%)		
	C1	C2	C3	C1	C2	C3
ICLS	2.032×10^{-7}	1.213×10^{-7}	0.001106	15.11	13.7	11.91
ICLS(W.I.)	1.102×10^{-7}	1.162×10^{-7}	0.0009402	6.812	17.03	14.34
DCLS(T.S.)	2.321×10^{-6}	6.768×10^{-7}	0.002925	36.19	54.98	16.88
ILS(F.S.)	1.994×10^{-7}	1.256×10^{-6}	2.486×10^{-7}	15.57	148.4	0.004188
ILS(F.S.W.I.)	1.175×10^{-5}	2.564×10^{-7}	0.0001237	562.1	36.55	2.111
ILS(B.E.)	2.095×10^{-7}	8.87×10^{-8}	4.609×10^{-6}	14.44	14.75	0.0912
ILS(B.E.W.I.)	2.079×10^{-7}	4.34×10^{-8}	0.0004155	14.79	7.142	8.058
PCR	2.027×10^{-7}	1.213×10^{-7}	0.001106	15.2	13.75	12.01
PCR(M.C.)	1.102×10^{-7}	1.161×10^{-7}	0.000939	6.891	17.47	14.77
PLS	2.013×10^{-7}	1.203×10^{-7}	0.001095	15.12	13.63	11.89
PLS(M.C.)	1.1×10^{-7}	1.156×10^{-7}	0.0009343	6.88	17.38	14.7

Methyl Or.-Fast Green-$CuSO_4$	RMSEV			rRMSEV(%)			$s_{y/x}$
	C1	C2	C3	C1	C2	C3	
ICLS	2.104×10^{-7}	1.321×10^{-7}	0.001087	9.768	13.37	11.49	0.1653
ICLS(W.I.)	1.271×10^{-7}	1.409×10^{-7}	0.001039	5.837	16.22	12.27	0.1404
DCLS(T.S.)							0.7179
ILS(F.S.)	7.078×10^{-7}	4.458×10^{-6}	8.823×10^{-7}	26.89	588.5	0.01127	
ILS(F.S.W.I.)	1.91×10^{-5}	3.734×10^{-7}	0.0001427	979.6	35.65	1.641	
ILS(B.E.)	8.026×10^{-7}	3.398×10^{-7}	1.765×10^{-5}	34.31	41.79	0.1465	
ILS(B.E.W.I.)	2.24×10^{-7}	6.692×10^{-8}	0.0006225	10.93	8.264	10.98	
PCR	2.077×10^{-7}	1.358×10^{-7}	0.001112	9.764	13.65	11.76	
PCR(M.C.)	6.588×10^{-7}	1.798×10^{-7}	0.001587	29.8	19.41	18.34	
PLS	2.066×10^{-7}	1.278×10^{-7}	0.001012	9.757	13.56	11.01	
PLS(M.C.)	6.29×10^{-8}	6.982×10^{-8}	0.0005146	2.989	8.101	6.222	

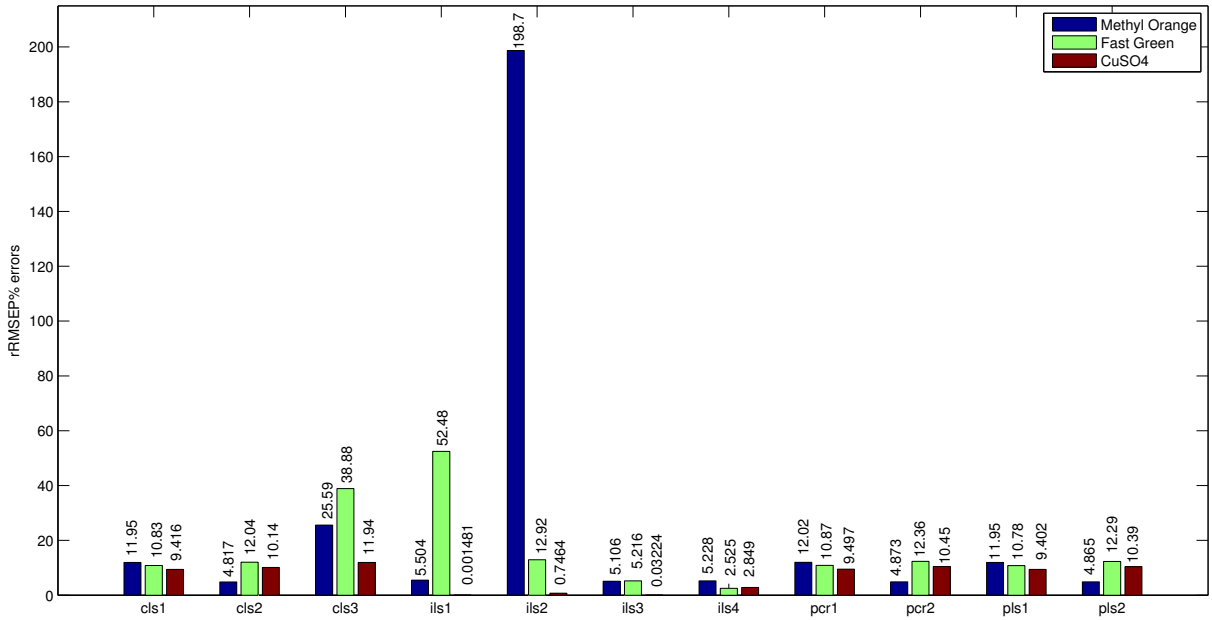


Figure 5.17: rRMSEP(%) error performance representation for Methyl Orange-Fast Green-$CuSO_4$ dataset

Table 5.9: Dataset 9, 3 factors - RMSE and standard deviation errors

Thymol-Mal Green	RMSEP			rRMSEP(%)		
Meth. Blue	C1	C2	C3	C1	C2	C3
ICLS	5.155×10^{-7}	7.278×10^{-7}	2.487×10^{-7}	6.401	76.64	10.1
ICLS(W.I.)	4.642×10^{-7}	4.972×10^{-7}	2.495×10^{-7}	4.43	22.01	10.82
DCLS(T.S.)	1.496×10^{-6}	1.75×10^{-6}	2.735×10^{-6}	15.6	153.9	37.3
ILS(F.S.)	9.769×10^{-8}	1.66×10^{-7}	1.94×10^{-8}	0.9533	19.47	1.244
ILS(F.S.W.I.)	1.115×10^{-7}	3.005×10^{-7}	1.246×10^{-7}	1.27	34.78	9.246
ILS(B.E.)	1.079×10^{-7}	2.548×10^{-8}	8.938×10^{-8}	0.972	2.928	5.234
ILS(B.E.W.I.)	4.208×10^{-8}	2.251×10^{-7}	1.73×10^{-7}	0.5294	24.43	10.7
PCR	5.146×10^{-7}	7.183×10^{-7}	2.471×10^{-7}	6.419	79.94	9.284
PCR(M.C.)	4.639×10^{-7}	4.94×10^{-7}	2.468×10^{-7}	4.801	22.44	10.64
PLS	5.006×10^{-7}	7.086×10^{-7}	2.404×10^{-7}	6.225	78.74	9.049
PLS(M.C.)	4.465×10^{-7}	4.847×10^{-7}	2.383×10^{-7}	4.607	22.45	10.36

Thymol-Mal Green	RMSEC			rRMSEC(%)		
Meth. Blue	C1	C2	C3	C1	C2	C3
ICLS	6.82×10^{-7}	9.628×10^{-7}	3.29×10^{-7}	8.468	101.4	13.36
ICLS(W.I.)	7.091×10^{-7}	7.596×10^{-7}	3.812×10^{-7}	6.767	33.62	16.53
DCLS(T.S.)	2.285×10^{-6}	2.674×10^{-6}	4.177×10^{-6}	23.83	235.1	56.98
ILS(F.S.)	2.585×10^{-7}	4.392×10^{-7}	5.132×10^{-8}	2.522	51.52	3.291
ILS(F.S.W.I.)	2.951×10^{-7}	7.951×10^{-7}	3.297×10^{-7}	3.36	92.03	24.46
ILS(B.E.)	2.855×10^{-7}	6.741×10^{-8}	2.365×10^{-7}	2.572	7.748	13.85
ILS(B.E.W.I.)	1.113×10^{-7}	5.956×10^{-7}	4.578×10^{-7}	1.401	64.63	28.32
PCR	6.808×10^{-7}	9.502×10^{-7}	3.269×10^{-7}	8.492	105.7	12.28
PCR(M.C.)	7.086×10^{-7}	7.546×10^{-7}	3.77×10^{-7}	7.334	34.28	16.25
PLS	6.622×10^{-7}	9.374×10^{-7}	3.181×10^{-7}	8.235	104.2	11.97
PLS(M.C.)	6.82×10^{-7}	7.404×10^{-7}	3.639×10^{-7}	7.038	34.29	15.83

Thymol-Mal Green	RMSEV			rRMSEV(%)			
Meth. Blue	C1	C2	C3	C1	C2	C3	$s_{y/x}$
ICLS	8.187×10^{-7}	1.038×10^{-6}	3.795×10^{-7}	7.794	60.8	12.07	0.244
ICLS(W.I.)	9.43×10^{-7}	9.996×10^{-7}	5.014×10^{-7}	8.427	39.73	18.73	0.2141
DCLS(T.S.)							1.72
ILS(F.S.)	8.352×10^{-7}	1.419×10^{-6}	1.658×10^{-7}	7.365	124.7	7.208	
ILS(F.S.W.I.)	3.572×10^{-7}	9.5×10^{-7}	4.326×10^{-7}	3.575	59.53	23.07	
ILS(B.E.)	1.464×10^{-6}	3.458×10^{-7}	1.213×10^{-6}	16.55	35.65	68.56	
ILS(B.E.W.I.)	1.56×10^{-7}	9.221×10^{-7}	5.782×10^{-7}	1.192	50.29	24.14	
PCR	8.077×10^{-7}	1.136×10^{-6}	7.624	73.2	11.16	4.217	
PCR(M.C.)	1.454×10^{-6}	1.073×10^{-6}	7.046×10^{-7}	14.67	66.89	30.29	
PLS	8.036×10^{-7}	1.033×10^{-6}	3.496×10^{-7}	7.425	72.1	10.951	
PLS(M.C.)	3.789×10^{-7}	4.194×10^{-7}	1.991×10^{-7}	3.443	16.96	7.512	

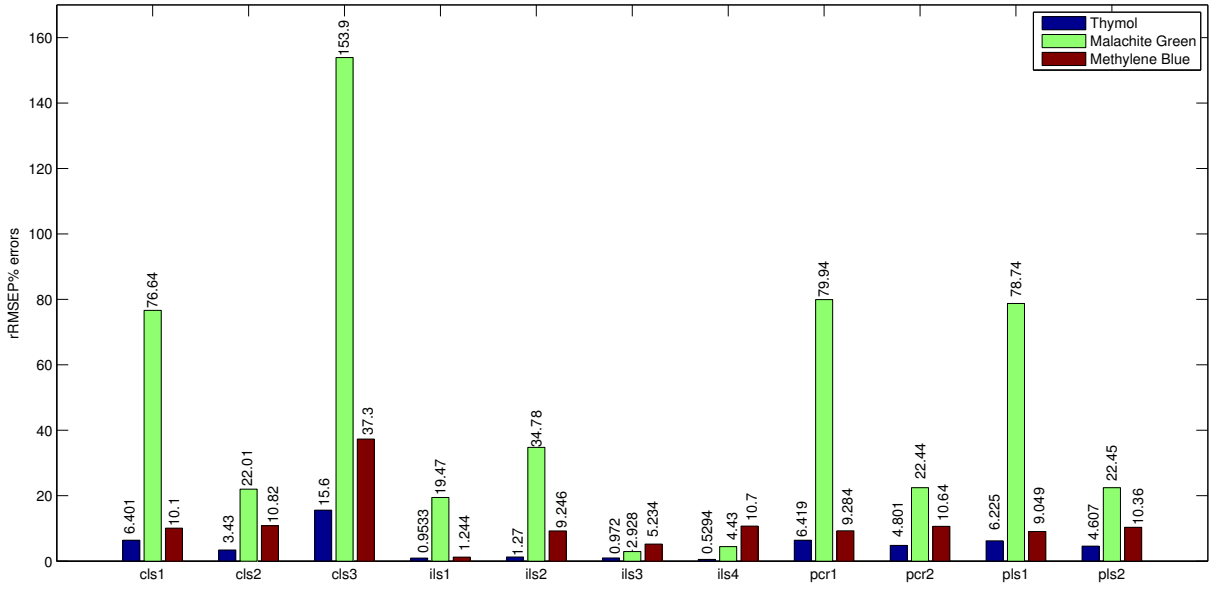


Figure 5.18: rRMSEP(%) error performance representation for Thymol-Malachite Green-Methylene Blue dataset

5.5 Simulated Data: Results

In this section the results for 2 and 3 components for the simulated datasets obtained, with the addition or without the addition of noise, are presented (tables 5.10, 5.11, 5.12, 5.12, 5.13, 5.14 and histograms 5.19, **??**). The results for 4 or 5 components are presented in the Appendix 6.1 (tables 1, 2), as wells as the corresponding histograms.

Table 5.10: Dataset 1, 2 factors - RMSE and standard deviation errors

	RMSEP		RMSEC		RMSEV	
Comp.1-Comp.2	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
DCLS	0.0001002	8×10^{-5}	0.0001137	9.072×10^{-5}		
ICLS	8.377×10^{-5}	7.98×10^{-5}	9.499×10^{-5}	9.048×10^{-5}	9.368×10^{-5}	7.901×10^{-5}
ICLS(W.I.)	7.665×10^{-5}	7.677×10^{-5}	9.387×10^{-5}	9.402×10^{-5}	9.21×10^{-5}	9.052×10^{-5}
DCLS(T.S.)	0.01281	0.01472	0.01569	0.01803		
ILS(F.S.)	5.53×10^{-14}	2.306×10^{-14}	1.659×10^{-13}	6.918×10^{-14}	1.678×10^{-13}	8.833×10^{-14}
ILS(F.S.W.I.)	1.768×10^{-16}	5.495×10^{-16}	5.303×10^{-16}	1.649×10^{-15}	2.375×10^{-16}	3.022×10^{-16}
ILS(B.E.)	1.256×10^{-13}	6.666×10^{-14}	3.769×10^{-13}	2×10^{-13}	5.193×10^{-5}	2.741×10^{-5}
ILS(B.E.W.I.)	4.595×10^{-13}	1.91×10^{-13}	1.379×10^{-12}	5.731×10^{-13}	0.0001587	4.441×10^{-5}
PCR	8.377×10^{-5}	7.98×10^{-5}	9.499×10^{-5}	9.048×10^{-5}	9.366×10^{-5}	7.9×10^{-5}
PCR(M.C.)	7.665×10^{-5}	7.677×10^{-5}	9.387×10^{-5}	9.402×10^{-5}	0.01316	0.01127
PLS	8.377×10^{-5}	7.98×10^{-5}	9.499×10^{-5}	9.048×10^{-5}	9.366×10^{-5}	7.9×10^{-5}
PLS(M.C.)	7.665×10^{-5}	7.677×10^{-5}	9.387×10^{-5}	9.402×10^{-5}	5.828×10^{-5}	6.186×10^{-5}

	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
Comp.1-Comp.2	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**	$\mathbf{s_{y/x}}$
DCLS	0.02824	0.02028	0.03202	0.023			0.004043
ICLS	0.02524	0.02046	0.02862	0.0232	0.02455	0.02005	0.002412
ICLS(W.I.)	0.02233	0.01813	0.02735	0.0222	0.02556	0.02226	0.002275
DCLS(T.S.)	4.402	5.03	5.391	6.16			0.2357
ILS(F.S.)	1.93×10^{-11}	6.525×10^{-12}	5.791×10^{-11}	1.957×10^{-11}	3.465×10^{-11}	1.969×10^{-11}	
ILS(F.S.W.I.)	4.788×10^{-14}	1.108×10^{-13}	1.436×10^{-13}	3.325×10^{-13}	4.416×10^{-14}	8.728×10^{-14}	
ILS(B.E.)	1.993×10^{-11}	2.549×10^{-11}	5.979×10^{-11}	7.647×10^{-11}	0.01672	0.004264	
ILS(B.E.W.I.)	8.677×10^{-11}	3.438×10^{-11}	2.603×10^{-10}	1.031×10^{-10}	0.0431	0.008883	
PCR	0.02524	0.02046	0.02862	0.02319	0.02456	0.02006	
PCR(M.C.)	0.02233	0.01812	0.02734	0.02219	4.01	3.516	
PLS	0.02524	0.02046	0.02862	0.02319	0.02456	0.02006	
PLS(M.C.)	0.02233	0.01812	0.02734	0.02219	0.01584	0.01453	

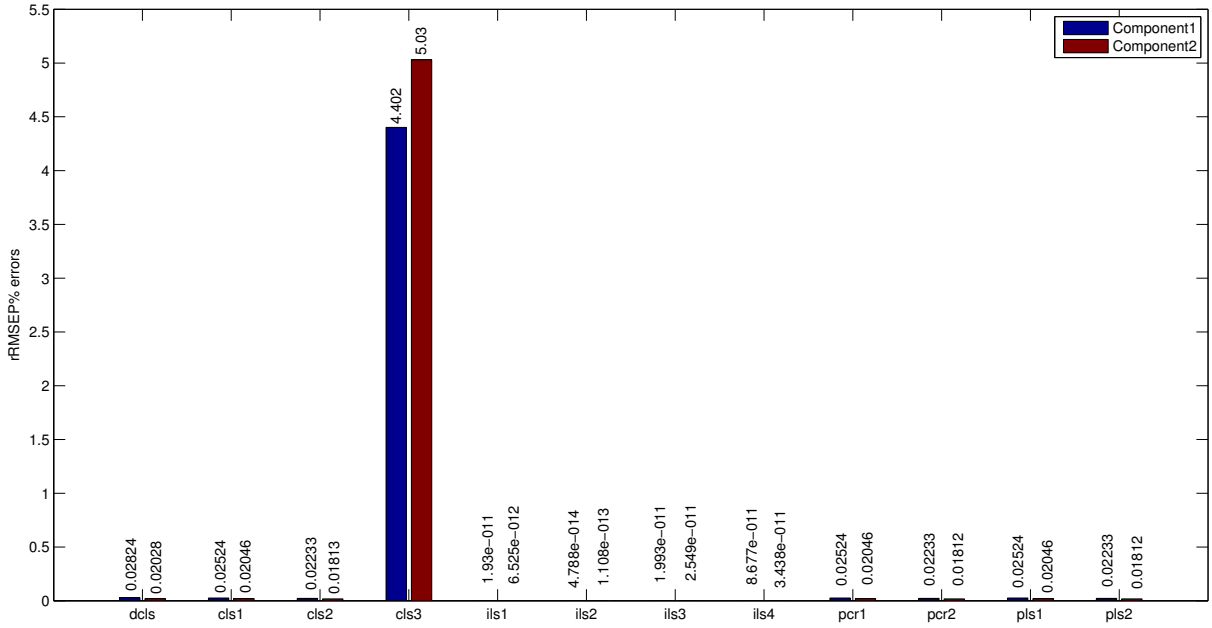


Figure 5.19: rRMSEP(%) error performance representation for 2 components of simulated spectra

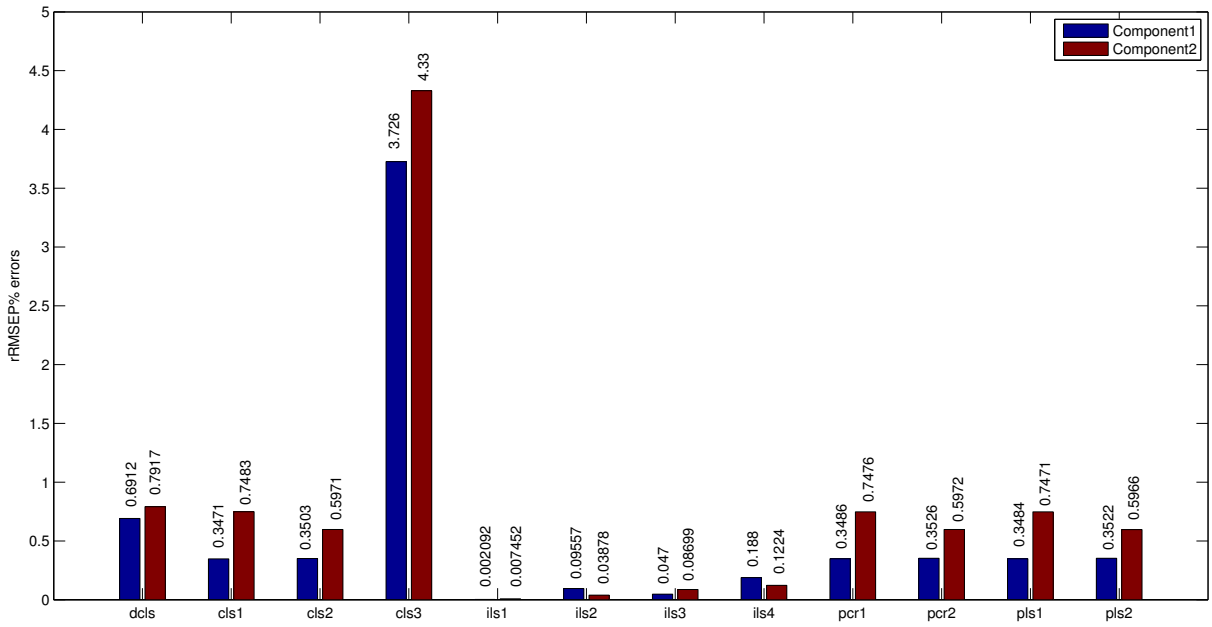


Figure 5.20: rRMSEP(%) error performance representation for 2 components of simulated spectra with 0.02 random noise

Table 5.11: Dataset 2, 2 factors (with noise=0.01) - RMSE and standard deviation errors

	RMSEP		RMSEC		RMSEV	
Comp.1-Comp.2	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
DCLS	0.001107	0.001061	0.001255	0.001203		
ICLS	0.0008994	0.001017	0.00102	0.001153	0.0009667	0.001174
ICLS(W.I.)	0.0008005	0.0009979	0.0009804	0.001222	0.001091	0.00124
DCLS(T.S.)	0.01276	0.01523	0.01562	0.01865		
ILS(F.S.)	7.96×10^{-6}	1.147×10^{-5}	2.388×10^{-5}	3.442×10^{-5}	0.000133	0.0001917
ILS(F.S.W.I.)	7.766×10^{-5}	0.0001067	0.000233	0.0003201	0.0003481	0.0004585
ILS(B.E.)	1.921×10^{-5}	7.985×10^{-5}	5.763×10^{-5}	0.0002395	0.0004073	0.001693
ILS(B.E.W.I.)	0.0006752	0.0001317	0.002026	0.000395	0.00553	0.001192
PCR	0.0008995	0.001017	0.00102	0.001153	0.0009838	0.001166
PCR(M.C.)	0.0008006	0.0009981	0.0009805	0.001222	0.01337	0.01159
PLS	0.0008993	0.001017	0.00102	0.001153	0.0009832	0.001166
PLS(M.C.)	0.0008003	0.0009978	0.0009802	0.001222	0.0006835	0.0008191

	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
Comp.1-Comp.2	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**	$\mathbf{s_{y/x}}$
DCLS	0.2685	0.29	0.3044	0.3288			0.04512
ICLS	0.1702	0.2867	0.1929	0.325	0.1892	0.2835	0.03963
ICLS(W.I.)	0.1939	0.2827	0.2375	0.3463	0.246	0.3114	0.03911
DCLS(T.S.)	4.248	5.256	5.202	6.438			0.2501
ILS(F.S.)	0.002388	0.003226	0.007164	0.009679	0.02935	0.05636	
ILS(F.S.W.I.)	0.01973	0.04056	0.05919	0.1217	0.08692	0.1308	
ILS(B.E.)	0.00912	0.01968	0.02736	0.05904	0.0806	0.4257	
ILS(B.E.W.I.)	0.2784	0.03712	0.8353	0.1114	1.239	0.3244	
PCR	0.1694	0.2865	0.1921	0.3248	0.1972	0.2805	
PCR(M.C.)	0.1937	0.2824	0.2373	0.3459	4.072	3.543	
PLS	0.1694	0.2864	0.1921	0.3248	0.1972	0.2802	
PLS(M.C.)	0.1937	0.2823	0.2372	0.3458	0.1527	0.205	

Table 5.12: Dataset 3, 2 factors (with noise=0.02) - RMSE and standard deviation errors

	RMSEP		RMSEC		RMSEV	
Comp.1-Comp.2	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
DCLS	0.002644	0.001905	0.002998	0.00216		
ICLS	0.002222	0.001806	0.002519	0.002048	0.002212	0.001916
ICLS(W.I.)	0.0022	0.001704	0.002695	0.002087	0.002729	0.002428
DCLS(T.S.)	0.009699	0.01317	0.01188	0.01613		
ILS(F.S.)	7.491×10^{-6}	2.094×10^{-5}	2.247×10^{-5}	6.281×10^{-5}	0.0001101	0.0003076
ILS(F.S.W.I.)	0.00026	0.0001777	0.00078	0.0005331	0.001364	0.0008357
ILS(B.E.)	0.0002439	0.0005027	0.0007318	0.001508	0.004788	0.009867
ILS(B.E.W.I.)	0.001025	0.0004204	0.003075	0.001261	0.006792	0.002723
PCR	0.002222	0.001807	0.00252	0.002049	0.002204	0.001921
PCR(M.C.)	0.002201	0.001705	0.002696	0.002089	0.01411	0.01164
PLS	0.002221	0.001805	0.002518	0.002047	0.002198	0.001919
PLS(M.C.)	0.002199	0.001704	0.002693	0.002087	0.001715	0.0015

	rRMSEP(%)		rRMSEC(%)		rRMSEV(%)		
Comp.1-Comp.2	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**	$\mathbf{s_{y/x}}$
DCLS	0.6912	0.7917	0.7838	0.8977			0.08703
ICLS	0.3471	0.7483	0.3936	0.8484	0.3607	0.5955	0.07436
ICLS(W.I.)	0.3503	0.5971	0.429	0.7313	0.4983	0.7211	0.07512
DCLS(T.S.)	3.726	4.33	4.564	5.303			0.2552
ILS(F.S.)	0.002092	0.007452	0.006277	0.02236	0.02852	0.08123	
ILS(F.S.W.I.)	0.09557	0.03878	0.2867	0.1164	0.3768	0.2068	
ILS(B.E.)	0.047	0.08699	0.141	0.261	1.449	3.297	
ILS(B.E.W.I.)	0.188	0.1224	0.564	0.3673	2.305	0.8589	
PCR	0.3486	0.7476	0.3953	0.8477	0.3414	0.6062	
PCR(M.C.)	0.3526	0.5972	0.4319	0.7314	4.244	3.673	
PLS	0.3484	0.7471	0.395	0.8471	0.3408	0.6049	
PLS(M.C.)	0.3522	0.5966	0.4314	0.7307	0.2962	0.4275	

Table 5.13: Dataset 4, 3 factors - RMSE and standard deviation errors

Comp.1-Comp.2	RMSEP			rRMSEP(%)		
Comp.3	C1	C2	C3	C1	C2	C3
DCLS	0.0001063	6.152×10^{-5}	7.901×10^{-5}	0.03249	0.02035	0.01959
ICLS	7.236×10^{-5}	5.223×10^{-5}	5.422×10^{-5}	0.02849	0.02062	0.01158
ICLS(W.I.)	3.705×10^{-5}	4.989×10^{-5}	1.085×10^{-5}	0.01381	0.01735	0.00162
DCLS(T.S.)	0.01263	0.02985	0.01253	5.012	11.84	2.504
ILS(F.S.)	2.453×10^{-13}	3.776×10^{-13}	1.169×10^{-13}	3.146×10^{-11}	1.159×10^{-10}	1.803×10^{-11}
ILS(F.S.W.I.)	6.59×10^{-15}	5.777×10^{-14}	1.436×10^{-13}	2.53×10^{-12}	1.513×10^{-11}	1.513×10^{-11}
ILS(B.E.)	3.452×10^{-5}	3.506×10^{-5}	3.408×10^{-5}	0.007621	0.01298	0.004144
ILS(B.E.W.I.)	2.742×10^{-14}	1.746×10^{-14}	5.017×10^{-14}	8.192×10^{-12}	4.251×10^{-12}	9.378×10^{-12}
PCR	7.236×10^{-5}	5.223×10^{-5}	5.422×10^{-5}	0.02849	0.02062	0.01158
PCR(M.C.)	3.705×10^{-5}	4.989×10^{-5}	1.085×10^{-5}	0.01381	0.01735	0.00162
PLS	7.236×10^{-5}	5.223×10^{-5}	5.422×10^{-5}	0.02849	0.02062	0.01158
PLS(M.C.)	3.705×10^{-5}	4.989×10^{-5}	1.085×10^{-5}	0.01381	0.01735	0.00162

Comp.1-Comp.2	RMSEC			rRMSEC(%)		
Comp.3	C1	C2	C3	C1	C2	C3
DCLS	0.0001345	7.782×10^{-5}	9.995×10^{-5}	0.0411	0.02574	0.02478
ICLS	9.153×10^{-5}	6.606×10^{-5}	6.858×10^{-5}	0.03604	0.02608	0.01465
ICLS(W.I.)	5.24×10^{-5}	7.055×10^{-5}	1.535×10^{-5}	0.01953	0.02454	0.002291
DCLS(T.S.)	0.01787	0.04222	0.01772	7.088	16.75	3.541
ILS(F.S.)	6.937×10^{-13}	1.068×10^{-12}	3.306×10^{-13}	8.899×10^{-11}	3.279×10^{-10}	5.099×10^{-11}
ILS(F.S.W.I.)	1.864×10^{-14}	1.634×10^{-13}	4.06×10^{-13}	7.156×10^{-12}	4.281×10^{-11}	5.159×10^{-11}
ILS(B.E.)	9.765×10^{-5}	9.915×10^{-5}	9.639×10^{-5}	0.02156	0.0367	0.01172
ILS(B.E.W.I.)	7.756×10^{-14}	4.937×10^{-14}	1.419×10^{-13}	2.317×10^{-11}	1.202×10^{-11}	2.653×10^{-11}
PCR	9.153×10^{-5}	6.606×10^{-5}	6.858×10^{-5}	0.03604	0.02608	0.01465
PCR(M.C.)	5.24×10^{-5}	7.055×10^{-5}	1.535×10^{-5}	0.01953	0.02454	0.002292
PLS	9.153×10^{-5}	6.606×10^{-5}	6.858×10^{-5}	0.03604	0.02608	0.01465
PLS(M.C.)	5.24×10^{-5}	7.055×10^{-5}	1.535×10^{-5}	0.01953	0.02454	0.002292

Comp.1-Comp.2	RMSEV			rRMSEV(%)			
Comp.3	C1	C2	C3	C1	C2	C3	$s_{y/x}$
DCLS							0.005218
ICLS	0.0001051	6.401×10^{-5}	5.664×10^{-5}	0.03171	0.0192	0.01026	0.002155
ICLS(W.I.)	6.056×10^{-5}	7.393×10^{-5}	1.782×10^{-5}	0.01929	0.02225	0.002632	0.001768
DCLS(T.S.)							0.3875
ILS(F.S.)	2.898×10^{-13}	1.746×10^{-13}	4.319×10^{-13}	6.038×10^{-11}	5.509×10^{-11}	7.967×10^{-11}	
ILS(F.S.W.I.)	6.182×10^{-14}	5.372×10^{-14}	9.195×10^{-14}	2.167×10^{-11}	1.985×10^{-11}	1.424×10^{-11}	
ILS(B.E.)	0.000614	0.0006235	0.0006061	0.2393	0.224	0.1234	
ILS(B.E.W.I.)	5.484×10^{-14}	4.855×10^{-14}	1.187×10^{-13}	1.504×10^{-11}	1.51×10^{-11}	2.409×10^{-11}	
PCR	0.0001051	6.405×10^{-5}	0.03171	0.01922	0.01026	0.003629	
PCR(M.C.)	0.03429	0.02939	0.03918	10.48	9.183	6.53	
PLS	0.0001051	6.405×10^{-5}	5.666×10^{-5}	0.03171	0.01922	0.01026	
PLS(M.C.)	3.028×10^{-5}	3.699×10^{-5}	8.899×10^{-6}	0.009656	0.01114	0.001313	

Table 5.14: Dataset 5, 3 factors (with noise=0.01) - RMSE and standard deviation errors

Comp.1-Comp.2	RMSEP			rRMSEP(%)		
Comp.3	C1	C2	C3	C1	C2	C3
DCLS	0.0009901	0.0009796	0.001101	0.4356	0.4314	0.2606
ICLS	0.0009056	0.0007398	0.0006459	0.3487	0.2522	0.1416
ICLS(W.I.)	0.0007168	0.0006621	0.0006454	0.321	0.2687	0.1426
DCLS(T.S.)	0.01514	0.02358	0.01142	5.919	9.416	2.245
ILS(F.S.)	9.556×10^{-5}	0.0001554	7.618×10^{-5}	0.03323	0.07217	0.01177
ILS(F.S.W.I.)	0.00056	0.0004385	0.0008081	0.1692	0.1693	0.1895
ILS(B.E.)	9.988×10^{-5}	0.0002064	0.000154	0.03235	0.08185	0.02688
ILS(B.E.W.I.)	0.001549	0.0009253	0.001443	0.5818	0.3404	0.254
PCR	0.0009057	0.0007399	0.0006459	0.3488	0.2521	0.1417
PCR(M.C.)	0.0007169	0.0006621	0.0006454	0.3209	0.2689	0.1427
PLS	0.0009054	0.0007397	0.0006457	0.3487	0.252	0.1416
PLS(M.C.)	0.0007166	0.000662	0.0006452	0.3208	0.2688	0.1427

Comp.1-Comp.2	RMSEC			rRMSEC(%)		
Comp.3	C1	C2	C3	C1	C2	C3
DCLS	0.001252	0.001239	0.001392	0.551	0.5457	0.3296
ICLS	0.001145	0.0009358	0.000817	0.4411	0.3191	0.1792
ICLS(W.I.)	0.001014	0.0009363	0.0009127	0.454	0.38	0.2016
DCLS(T.S.)	0.02142	0.03335	0.01615	8.371	13.32	3.175
ILS(F.S.)	0.0002703	0.0004395	0.0002155	0.094	0.2041	0.03328
ILS(F.S.W.I.)	0.001584	0.00124	0.002286	0.4787	0.4789	0.536
ILS(B.E.)	0.0002825	0.0005839	0.0004357	0.09151	0.2315	0.07603
ILS(B.E.W.I.)	0.004381	0.002617	0.004083	1.646	0.9628	0.7185
PCR	0.001146	0.0009359	0.000817	0.4412	0.3189	0.1792
PCR(M.C.)	0.001014	0.0009364	0.0009127	0.4539	0.3803	0.2018
PLS	0.001145	0.0009357	0.0008168	0.4411	0.3188	0.1792
PLS(M.C.)	0.001013	0.0009362	0.0009125	0.4537	0.3802	0.2018

Comp.1-Comp.2	RMSEV			rRMSEV(%)			
Comp.3	C1	C2	C3	C1	C2	C3	$s_{y/x}$
DCLS							0.05092
ICLS	0.001217	0.0009987	0.0008483	0.3732	0.2869	0.1458	0.03923
ICLS(W.I.)	0.001202	0.001036	0.001094	0.4568	0.3458	0.1993	0.03902
DCLS(T.S.)							0.3728
ILS(F.S.)	0.001359	0.002209	0.001083	0.4728	0.4852	0.1573	
ILS(F.S.W.I.)	0.003227	0.002087	0.004131	0.9983	0.4609	0.7927	
ILS(B.E.)	0.002146	0.004436	0.00331	0.7816	1.593	0.6935	
ILS(B.E.W.I.)	0.004628	0.002471	0.006254	1.403	0.747	1.021	
PCR	0.001206	0.001005	0.000826	0.3705	0.2871	0.1436	
PCR(M.C.)	0.03429	0.02939	0.03918	10.46	9.179	6.531	
PLS	0.001203	0.001003	0.000824	0.3704	0.287	0.1435	
PLS(M.C.)	0.0005945	0.0005427	0.0005403	0.2278	0.1828	0.09925	

Observations:

- It should be noted that the standard error of regression ($\mathbf{s_{y/x}}$) was not calculated for all of the algorithms, because it is based on the difference between the expected and predicted value that each time represents the vertical axis. In CLS algorithms this is associated with the Absorbance mixture values in the linear equation, whereas in all of the other algorithms it is associated with the difference between the predicted and the original **C** values, which is also estimated by the RMSE errors. Thus, there is no need to include this error in every algorithm.

- In some algorithms, such as DCLS with the pure spectra available and DCLS through slope the validation errors (RMSEP and rRMSEP(%)) are not calculated for the reason that the coefficients used for the estimation of the predicted concentrations are already available (the pure spectra are also known as regression coefficients) and there is no way to be re-estimated in order to perform leave-one-out cross-validation.

Conclusions on the results from experimental and simulated datasets:

- In general, the ILS algorithm seems to yield the smallest error compared with the other algorithms, although there may be big variances between the different versions of it.This may have to do with the small number of most "suitable" wavelengths selected which makes ILS more efficient. On the other side, this wavelength selection instead of using the whole spectrum may yield to larger errors due to the disability of the algorithm to detect unusual samples.

- The CLS algorithm for these datasets of not so many samples, generally, extracts the same results as PCR and PLS.

- As seen in the bibliography, as well as in practice, PLS indeed handles noisy data better than PCR. In other words, under the absence of noise, PCR yields the same results as PLS and PCR with mean-centered data the same results as PLS with mean-centered data. This is better represented though simulated data, especially when we add random noise on the already non-noisy data.

- The above fact may help reveal that some substances are more suitable for concentration quantification than others and some may be avoided, such as Methylene Blue and the green ones (Fast Green and Malachite Green), especially when there is overlapping of their spectra in the final mixed spectrum. This is best represented in the histograms where some substances appear to usually cause the errors' increase.

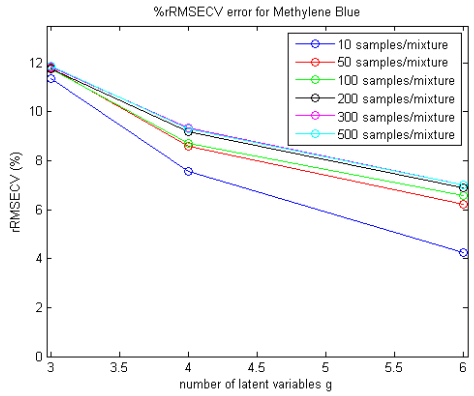
- Whether mean-centered data yield the best results or not compared to the non-mean-centered corresponding ones is not consistent. Although according to the literature mean-centered data are always preferred and forced to be used, this is not a rule in the case of experimental data, but counts for simulated data. The reason for the first category of data are the plethora of errors related to the experimental process as seen before (3.3), as well as the sediments the solutions may have developed.

- As for DCLS through slope, it seems to be the weakest and less efficient algorithm with the highest error in almost all the datasets and may not be preferable for further use (hyperspectral image processing) or in real world as the more the components the biggest the error it yields (according to the simulated datasets).

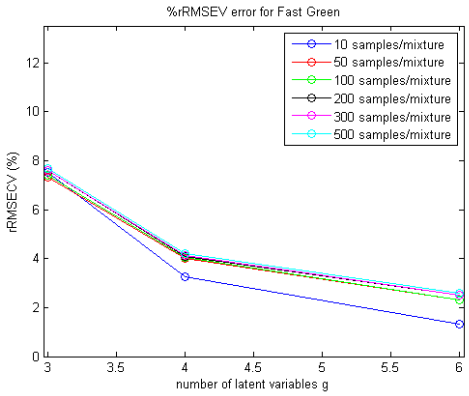
5.6 Hyperspectral Image Data: Results

After the application of the algorithmic methods on Simulated and Spectrophotometer data, we also applied the methods on the datasets acquired using the microscope configuration. More specifically, we obtained 2 datasets of hyperspectral images, one with mixtures of two components and the other with mixtures of three components. For every dataset we selected a training set for each of its samples, consisting of 10, 50, 100, 200, 300 or 500 mixture spectra. We subsequently applied the algorithms on indicative training sets, and found that all of them provided much worse results than PLS. The reason was that the PLS algorithm can deal with noisy data better than the other algorithms and also because of its complexity and the fact that it takes into account the interfering variables which may appear in the mixtures. Furthermore, with the increase of the latent variables the errors in PLS algorithm decrease even more.

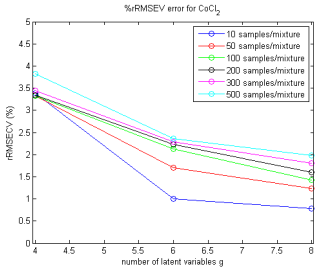
Further down we provide the results of PLS algorithm with mean-centered data for different training sets and number of latent variables g.



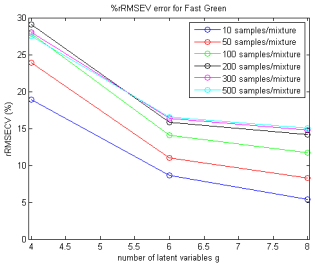
(a) Methylene Blue



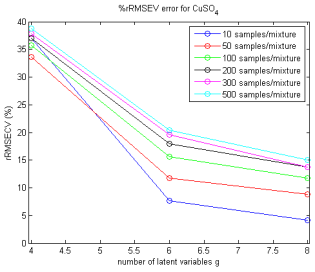
(b) Fast Green



(c) $CoCl_2$



(d) Fast Green

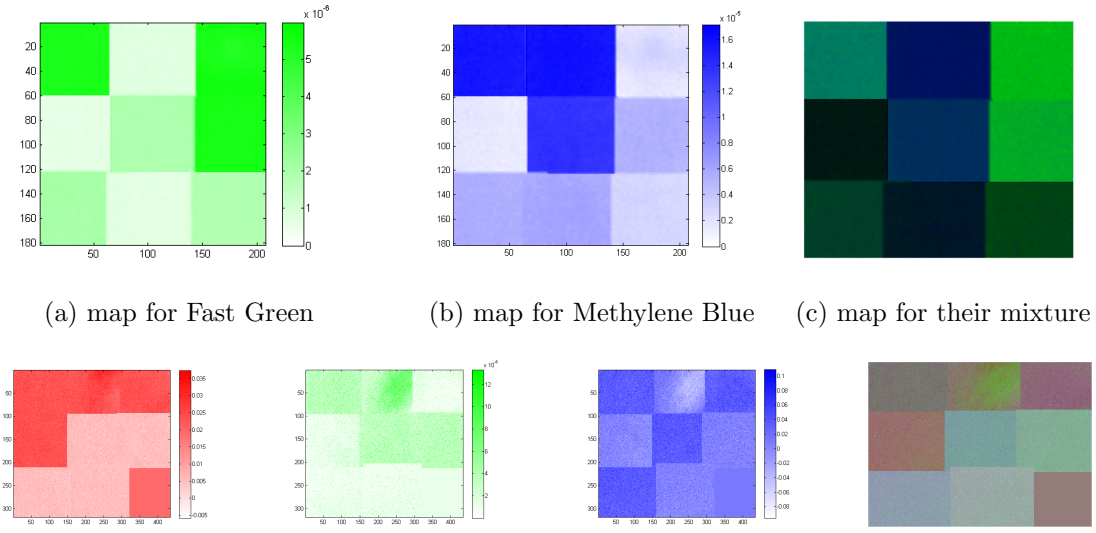


(e) $CuSO_4$

Figure 5.21: %rRMSEV error for datasets of 2(top) and 3(down) components

After the use of PLS in a training set and the concentrations estimation, the initial images were reconstructed and represented through thematic maps with pseudo-colors close to the colors of the individual stains participated in the mixture, as below:

It should be noted that the last sample in third row of the map for 3 components is blank because due to the full factorial design for 2 three components we have taken 8 mixtures to create this dataset.



(a) map for Fast Green (b) map for Methylene Blue (c) map for their mixture

(d) map for $CoCl_2$ (e) map for Fast Green (f) map for $CuSO_4$ (g) map for their mixture

Figure 5.22: Concentration maps for the individual components and their mixtures in the two datasets acquired by microscope. The brighter colors on the components maps correspond to higher concentrations in the mixture

Chapter 6

Conclusion

6.1 Future Work

The work in this thesis can be used as the basis for several future research directions, some of which are listed below:

1. Creation of experimental designs and application of algorithmic methods analyzed in chapter 4.2 for more than 5 constituents in a mixture.

2. Identification of individual spectra, based on spectra saved in spectral libraries, instead of concentration quantification only.

3. Implementation of more complex algorithms based on non-linear relationships, instead of only linear ones, like Support Vector Machines (SVM) and Analytic Neural Networks (ANN). In this case, a more physical analysis of the system could be made, considering the scattering effects of light into the samples.

4. Transferring the problem into brightfield microscopy and other types of spectroscopy, such as Raman (scattering) Spectroscopy and Fluorescence (emission) Spectroscopy.

5. Research in the infrared area of the electromagnetic spectrum in order to detect characteristics not visible with the naked eye.

6. Application in real tissue histochemical samples.

References

[1] Adamos, A.: Algorithmoi taksinomhshs dedomenwn yperfasmatikhs apeikonishs gia thn anixneush, tmhmatopoihsh kai tautopoihsh xarakthristikwn diagnwstikhs shmasias. Diploma thesis, Technical University of Crete (2011) available online: http://poseidon.library.tuc.gr/artemis/DT2006-0114/DT2006-0114.pdf. 3, 12

[2] Miramar College: Uv visible absorption spectroscopy. Available online: http://faculty.sdmiramar.edu/fgarces/LabMatters/Instruments/UV_Vis/Cary50.htm. 3

[3] Epitropou, G.: Hyper-spectral imaging and spectral segmentation algorithms for the non-destructive analysis of el greco'As paintings. Diploma thesis, Technical University of Crete, Greece (2008) 5, 73

[4] Ergasthrio Organikhs Xhmeias, G.P.A.: Askhsh 2: Fasmatofwtometria Available online: http://www.aua.gr/gr/dep/gen/ximia/ASKHSH-2ASf.pdf. 5, 7, 12

[5] Ocean Optics Inc.: Introduction to spectroscopy in the teaching lab using ocean optics spectrometers (2006) 6, 21

[6] Costas Balas , Christos Pappas and George Epitropou: Handbook of Biomedical Optics, Chapter 7. Multi/Hyper-Spectral Imaging. CRC Press (2011) 7

[7] Nicholas M. Short: Electromagnetic spectrum: Spectral signatures. Available online: https://www.fas.org/irp/imint/docs/rst/Intro/Part2_5.html. 9

[8] Hassan, K.W.: Novel regression methods for spectral data. Master thesis, Lappeenranta University of Technology, Lappeenranta (2012) 9

[9] Hibbert, D.B., Gooding, J.J.: Data analysis for chemistry. Oxford University Press New York (2006) 9, 44, 70

[10] Krastev, T.: Chemometrics. Available online: http://classifion.sicyon.com/References/Chemometrics.pdf. 9, 10, 11, 56

[11] James E. Burger: Hyperspectral NIR Image Analysis: Data Exploration, Correction, and Regression. PhD thesis, Swedish University of Agricultural Sciences (2006) 9

[12] Keshava, N., Mustard, J.F.: Spectral unmixing. Signal Processing Magazine, IEEE **19**(1) (2002) 44–57 12

[13] Grahn, H., Geladi, P.: Techniques and applications of hyperspectral image analysis. Wiley. com (2007) 14, 51

[14] Dissing, B.S., Nielsen, M.E., Ersbøll, B.K., Frosch, S.: Multispectral imaging for determination of astaxanthin concentration in salmonids. PloS one **6**(5) (2011) e19032 14

[15] Papadakis, A., Stathopoulos, E., Delides, G., Berberides, K., Nikiforidis, G., Balas, C.: A novel spectral microscope system: application in quantitative pathology. Biomedical Engineering, IEEE Transactions on **50**(2) (2003) 207–217 14

[16] Buszewski, B., Ulanowska, A., Kowalkowski, T., Cieśliński, K.: Investigation of lung cancer biomarkers by hyphenated separation techniques and chemometrics. Clinical Chemistry and Laboratory Medicine **50**(3) (2012) 573–581 15

[17] Justin A. Kerszulis: User guide and tutorial for using the Cary 500 in the Reynolds Research Group. Available online: https://ww2.chemistry.gatech.edu/reynolds/sites/ww2.chemistry.gatech.edu.reynolds/files/Cary.pdf. 21

[18] Zografos, O.: Hyper-spectral excitation-emission microscopy. Diploma thesis, Technical University of Crete, Greece (2011) 23

[19] News Medical: Biomarker - what is a biomarker? Available online: http://www.news-medical.net/health/Biomarker-What-is-a-Biomarker.aspx. 26

[20] Research Advocary Network: Biomarkers in Cancer Available online: http://researchadvocacy.org/images/uploads/downloads/BiomarkerinCancer_WebDownloadVersion.pdf. 26

[21] ScienceLab.com: Material safety data sheet, methyl orange, 1% MSDS Available online: http://www.sciencelab.com/msds.php?msdsId=9926081. 27

[22] Holtzman, N.A., Haslam, R.H.: Elevation of serum copper following copper sulfate as an emetic. Pediatrics **42**(1) (1968) 189–193 27

[23] Roger Hiorns: Seizure Available online: http://www.artangel.org.uk//projects/2008/seizure/about_the_project/seizure. 27

[24] Green, C.: Bordeaux etch - an electrochemical method Available online: http://www.greenart.info/galvetch/bordeaux.htm. 27

[25] Johnson, G.F.: The early history of copper fungicides. Agricultural History **9**(2) (1935) 67–79 27

[26] Srivastava, S., Sinha, R., Roy, D.: Toxicological effects of malachite green. Aquatic Toxicology **66**(3) (2004) 319–329 28

[27] International Program on Chemical Safety: Fast green fcf Available online: http://www.inchem.org/documents/jecfa/jecmono/v16je12.htm. 28

[28] Earnshaw, A., Greenwood, N.: Chemistry of the Elements. Elsevier (1997) 28

[29] Kiernan, J.A.: General oversight stains for histology and histopathology. Education guide. Special stains and H and E, pages (2010) 29–31 29

[30] Leardi, R.: Experimental design in chemistry: A tutorial. Analytica chimica acta **652**(1) (2009) 161–172 30, 31

[31] James N. Miller, Jane C. Miller: Statistics and Chemometrics for Analytical Chemistry. Prentice Hall, UK (2010) 30, 53, 70

[32] Brereton, R.G.: Chemometrics: data analysis for the laboratory and chemical plant. John Wiley & Sons (2003) 30

[33] Tom O'Haver, Department of Chemistry and Biochemistry, The University of Maryland: Curve fitting B: Multicomponent Spectroscopy Available online: http://terpconnect.umd.edu/~toh/spectrum/CurveFittingB.html. 46

[34] Kenneth R. Beebe, Randy J. Pell, Mary Beth Seasholtz: Chemometrics: A Practical Guide. John Wiley & Sons, Inc. (1998) 51, 55

[35] Gemperline, P.: Practical guide to chemometrics. CRC press (2010) 52, 53, 58, 69

[36] Jørgensen, B., Goegebeur, Y.: St02: Multivariate data analysis and chemometrics. University of Southern Denmark (2007) 52, 57, 59, 62, 75

[37] Kramer, R.: Chemometric techniques for quantitative analysis (1998) 54, 69

[38] Adams, M.J.: Chemometrics in analytical spectroscopy. Royal Society of Chemistry (2004) 55

[39] United Nations Educational, Scientific and Cultural Organization: Multiple Regression Model Available online: http://www.unesco.org/webworld/idams/advguide/Chapt5_2.htm. 58, 69

[40] Salt Okunur: Feature Selection Available online: http://www.cse.msu.edu/~cse802/Feature_selection.pdf. 58

[41] Mark, H., Workman Jr, J.: Chemometrics in spectroscopy. Academic Press (2010) 60, 75

[42] Buddenbaum, H., Steffens, M.: Mapping the distribution of chemical properties in soil profiles using laboratory imaging spectroscopy, svm and pls regression. EARSeL EProceedings **11**(1) (2012) 25–32 62

[43] Prof.S.Boyd: Least squares and least norm in Matlab Available online: http://see.stanford.edu/materials/lsoeldsee263/Additional4-ls_ln_matlab.pdf. 74

[44] Kallergi, G., Papadaki, M.A., Politaki, E., Mavroudis, D., Georgoulias, V., Agelaki, S.: Epithelial to mesenchymal transition markers expressed in circulating tumour cells of early and metastatic breast cancer patients. Breast Cancer Res **13**(3) (2011) R59 14

[45] Ruifrok, A.C., Johnston, D.A.: Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology **23**(4) (2001) 291–299 14, 15

[46] Abcam®: Rhodamine conjugation kit protocol. Available online: http://www.abcam.com/ps/products/102/ab102915/documents/ab102915%20Rhodamine%20Conjugation%20Kit%20%28website%29.pdf. 14

Appendix

Table 1: Dataset 1, 4 factors - RMSE and standard deviation errors

Comp.1-Comp.2 / Comp.3-Comp.4	RMSEP				rRMSEP(%)			
	C1	C2	C3	C4	C1	C2	C3	C4
DCLS	8.781×10^{-5}	8.287×10^{-5}	7.128×10^{-5}	7.845×10^{-5}	0.02142	0.03685	0.02478	0.01926
ICLS	2.874×10^{-5}	6.507×10^{-5}	5.38×10^{-5}	4.823×10^{-5}	0.009108	0.0244	0.02436	0.009893
ICLS(W.I.)	2.278×10^{-5}	6.449×10^{-5}	4.81×10^{-5}	1.14×10^{-5}	0.005605	0.02336	0.01753	0.002124
DCLS(T.S.)	0.01532	0.02326	0.02888	0.01274	3.977	9.488	11.49	2.546
ILS(F.S.)	8.798×10^{-14}	7.926×10^{-14}	1.144×10^{-13}	1.08×10^{-13}	2.187×10^{-11}	2.914×10^{-11}	3.813×10^{-11}	1.12×10^{-11}
ILS(F.S.W.I.)	1.913×10^{-16}	4.568×10^{-16}	3.243×10^{-16}	4.177×10^{-16}	4.158×10^{-14}	2.26×10^{-13}	1.153×10^{-13}	8.244×10^{-14}
ILS(B.E.)	2.653×10^{-14}	1.279×10^{-14}	4.339×10^{-16}	3.992×10^{-14}	3.737×10^{-12}	4.954×10^{-12}	1.426×10^{-13}	7.416×10^{-12}
ILS(B.E.W.I.)	8.161×10^{-16}	5.625×10^{-16}	1.928×10^{-16}	6.237×10^{-16}	1.962×10^{-13}	1.18×10^{-13}	6.509×10^{-14}	9.438×10^{-14}
PCR	2.874×10^{-5}	6.507×10^{-5}	5.38×10^{-5}	4.823×10^{-5}	0.009109	0.0244	0.02436	0.009893
PCR(M.C.)	2.278×10^{-5}	6.449×10^{-5}	4.81×10^{-5}	1.14×10^{-5}	0.005606	0.02336	0.01753	0.002123
PLS	2.874×10^{-5}	6.507×10^{-5}	5.38×10^{-5}	4.823×10^{-5}	0.009109	0.0244	0.02436	0.009893
PLS(M.C.)	2.278×10^{-5}	6.449×10^{-5}	4.81×10^{-5}	1.14×10^{-5}	0.005606	0.02336	0.01753	0.002123

Comp.1-Comp.2 / Comp.3-Comp.4	RMSEC				rRMSEC(%)			
	C1	C2	C3	C4	C1	C2	C3	C4
DCLS	0.0001242	0.0001172	0.0001008	0.0001109	0.03029	0.05211	0.03505	0.02723
ICLS	4.064×10^{-5}	9.203×10^{-5}	7.608×10^{-5}	6.821×10^{-5}	0.01288	0.03451	0.03445	0.01399
ICLS(W.I.)	3.72×10^{-5}	0.0001053	7.855×10^{-5}	1.861×10^{-5}	0.009153	0.03814	0.02862	0.003468
DCLS(T.S.)	0.02502	0.03798	0.04716	0.0208	6.495	15.49	18.76	4.158
ILS(F.S.)	2.488×10^{-13}	2.242×10^{-13}	3.235×10^{-13}	3.054×10^{-13}	6.185×10^{-11}	8.243×10^{-11}	1.079×10^{-10}	3.167×10^{-11}
ILS(F.S.W.I.)	5.411×10^{-16}	1.292×10^{-15}	9.172×10^{-16}	1.181×10^{-15}	1.176×10^{-13}	6.392×10^{-13}	3.261×10^{-13}	2.332×10^{-13}
ILS(B.E.)	7.504×10^{-14}	3.619×10^{-14}	1.227×10^{-15}	1.129×10^{-13}	1.057×10^{-11}	1.401×10^{-11}	4.034×10^{-13}	2.098×10^{-11}
ILS(B.E.W.I.)	2.308×10^{-15}	1.591×10^{-15}	5.453×10^{-16}	1.764×10^{-15}	5.548×10^{-13}	3.338×10^{-13}	1.841×10^{-13}	2.669×10^{-13}
PCR	4.064×10^{-5}	9.203×10^{-5}	7.608×10^{-5}	6.821×10^{-5}	0.01288	0.03451	0.03445	0.01399
PCR(M.C.)	3.72×10^{-5}	0.0001053	7.855×10^{-5}	1.861×10^{-5}	0.009154	0.03814	0.02862	0.003468
PLS	4.064×10^{-5}	9.203×10^{-5}	7.608×10^{-5}	6.821×10^{-5}	0.01288	0.03451	0.03445	0.01399
PLS(M.C.)	3.72×10^{-5}	0.0001053	7.855×10^{-5}	1.861×10^{-5}	0.009154	0.03814	0.02862	0.003468

Comp.1-Comp.2 / Comp.3-Comp.4	RMSEV				rRMSEV(%)				
	C1	C2	C3	C4	C1	C2	C3	C4	$s_{y/x}$
DCLS									0.006201
ICLS	3.63×10^{-5}	0.0001306	8.536×10^{-5}	5.281×10^{-5}	0.00855	0.04035	0.02909	0.009262	0.002527
ICLS(W.I.)	5.309×10^{-5}	0.0001476	0.0001159	2.476×10^{-5}	0.01152	0.04513	0.03623	0.004107	0.00238
DCLS(T.S.)									0.515
ILS(F.S.)	9.474×10^{-14}	2.061×10^{-13}	1.45×10^{-13}	1.301×10^{-13}	1.98×10^{-11}	5.736×10^{-11}	3.577×10^{-11}	2.309×10^{-11}	
ILS(F.S.W.I.)	1.332×10^{-15}	1.797×10^{-15}	7.633×10^{-16}	7.425×10^{-16}	2.077×10^{-13}	7.475×10^{-13}	2.151×10^{-13}	1.347×10^{-13}	
ILS(B.E.)	3.523×10^{-14}	3.173×10^{-14}	3.825×10^{-14}	4.508×10^{-14}	5.445×10^{-12}	1.049×10^{-11}	1.284×10^{-11}	7.806×10^{-12}	
ILS(B.E.W.I.)	4.094×10^{-16}	4.476×10^{-16}	2.741×10^{-16}	6.384×10^{-16}	9.113×10^{-14}	1.482×10^{-13}	8.5×10^{-14}	1.226×10^{-13}	
PCR	3.631×10^{-5}	0.0001306	8.536×10^{-5}	5.272×10^{-5}	0.008533	0.04034	0.02909	0.009237	
PCR(M.C.)	0.05714	0.05714	0.04898	0.06531	12.38	17.46	15.31	10.88	
PLS	3.631×10^{-5}	0.0001306	8.536×10^{-5}	5.272×10^{-5}	0.008533	0.04034	0.02909	0.009237	
PLS(M.C.)	1.991×10^{-5}	5.537×10^{-5}	4.347×10^{-5}	9.285×10^{-6}	0.004314	0.01692	0.01358	0.001548	

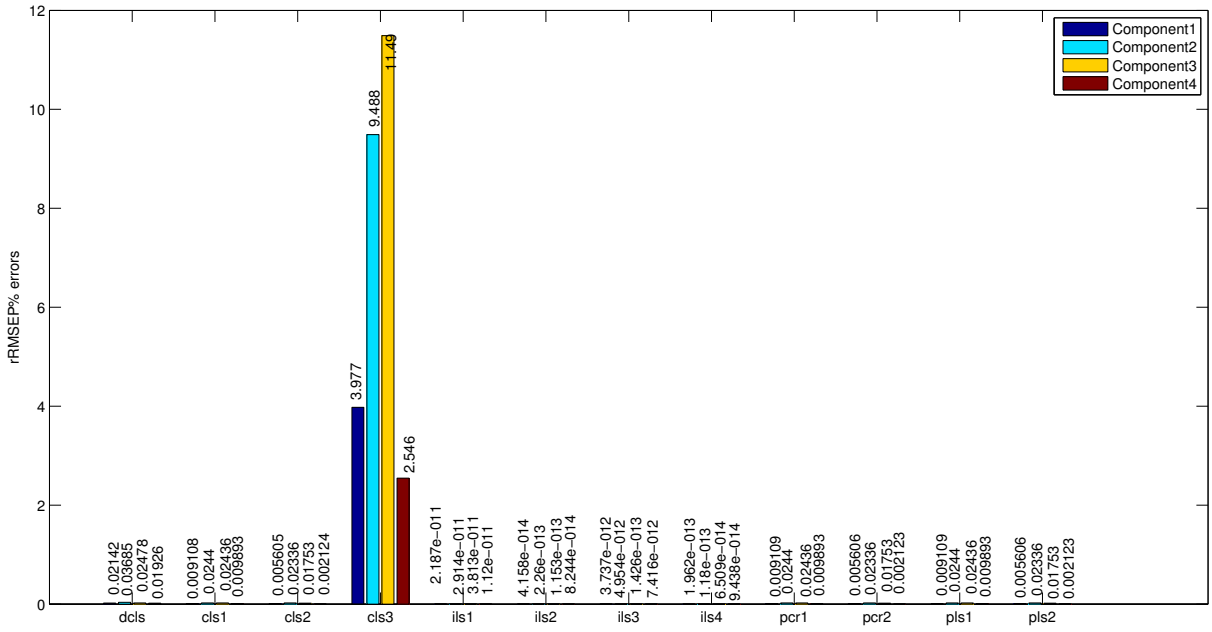


Figure 1: rRMSEP(%) error performance representation for 4 components of simulated spectra

Table 2: Dataset 2, 5 factors - RMSE and standard deviation errors

	rRMSEP(%)				
Comp.1-2-3-4-5	**C1**	**C2**	**C3**	**C4**	**C5**
DCLS	0.02704	0.02606	0.05389	0.04695	0.02935
ICLS	0.006329	0.01661	0.02875	0.0242	0.006055
ICLS(W.I.)	0.006542	0.009815	0.01917	0.005763	0.0001284
DCLS(T.S.)	2.655	11.06	3.737	13.99	1.842
ILS(F.S.)	4.814×10^{-11}	5.407×10^{-11}	5.007×10^{-11}	4.573×10^{-11}	7.342×10^{-11}
ILS(F.S.W.I.)	0.2073	0.0002117	0.0005209	0.0008121	1.173×10^{-13}
ILS(B.E.)	1.51×10^{-11}	9.466×10^{-12}	1.216×10^{-11}	6.023×10^{-12}	2.792×10^{-11}
ILS(B.E.W.I.)	1.398×10^{-13}	9.254×10^{-14}	1.661×10^{-13}	1.316×10^{-13}	7.438×10^{-14}
PCR	0.006329	0.01661	0.02875	0.0242	0.006055
PCR(M.C.)	0.006542	0.009815	0.01917	0.005763	0.0001284
PLS	0.006329	0.01661	0.02875	0.0242	0.006055
PLS(M.C.)	0.006542	0.009815	0.01917	0.005763	0.0001284

	rRMSEC(%)				
Comp.1-2-3-4-5	**C1**	**C2**	**C3**	**C4**	**C5**
DCLS	0.04416	0.04256	0.088	0.07666	0.04793
ICLS	0.01033	0.02712	0.04694	0.03951	0.009888
ICLS(W.I.)	0.01308	0.01963	0.03833	0.01153	0.0002568
DCLS(T.S.)	5.311	22.13	7.473	27.98	3.684
ILS(F.S.)	1.362×10^{-10}	1.529×10^{-10}	1.416×10^{-10}	1.294×10^{-10}	2.077×10^{-10}
ILS(F.S.W.I.)	0.5864	0.0005988	0.001473	0.002297	3.318×10^{-13}
ILS(B.E.)	4.27×10^{-11}	2.677×10^{-11}	3.44×10^{-11}	1.704×10^{-11}	7.898×10^{-11}
ILS(B.E.W.I.)	3.955×10^{-13}	2.617×10^{-13}	4.698×10^{-13}	3.721×10^{-13}	2.104×10^{-13}
PCR	0.01033	0.02712	0.04695	0.03951	0.009888
PCR(M.C.)	0.01308	0.01963	0.03834	0.01153	0.0002568
PLS	0.01033	0.02712	0.04695	0.03951	0.009888
PLS(M.C.)	0.01308	0.01963	0.03834	0.01153	0.0002568

	rRMSEV(%)					
Comp.1-2-3-4-5	**C1**	**C2**	**C3**	**C4**	**C5**	$\mathbf{s_{y/x}}$
DCLS						0.00727
ICLS	0.01959	0.0353	0.05401	0.04024	0.006495	0.002307
ICLS(W.I.)	0.02428	0.02606	0.06046	0.02073	0.0004589	0.001837
DCLS(T.S.)						0.6663
ILS(F.S.)	3.194×10^{-11}	4.388×10^{-11}	3.663×10^{-11}	3.764×10^{-11}	5.124×10^{-11}	
ILS(F.S.W.I.)	0.7699	0.0007456	0.001758	0.002785	9.217×10^{-13}	
ILS(B.E.)	1.17×10^{-11}	7.634×10^{-12}	2.238×10^{-11}	6.2×10^{-12}	8.646×10^{-12}	
ILS(B.E.W.I.)	1.153×10^{-13}	1.161×10^{-13}	1.708×10^{-13}	1.128×10^{-13}	1.434×10^{-13}	
PCR	0.01958	0.0353	0.05406	0.04024	0.006501	
PCR(M.C.)	12.96	20.63	29.1	25.51	18.14	
PLS	0.01958	0.0353	0.05406	0.04024	0.006501	
PLS(M.C.)	0.006069	0.006515	0.01512	0.005182	0.0001147	

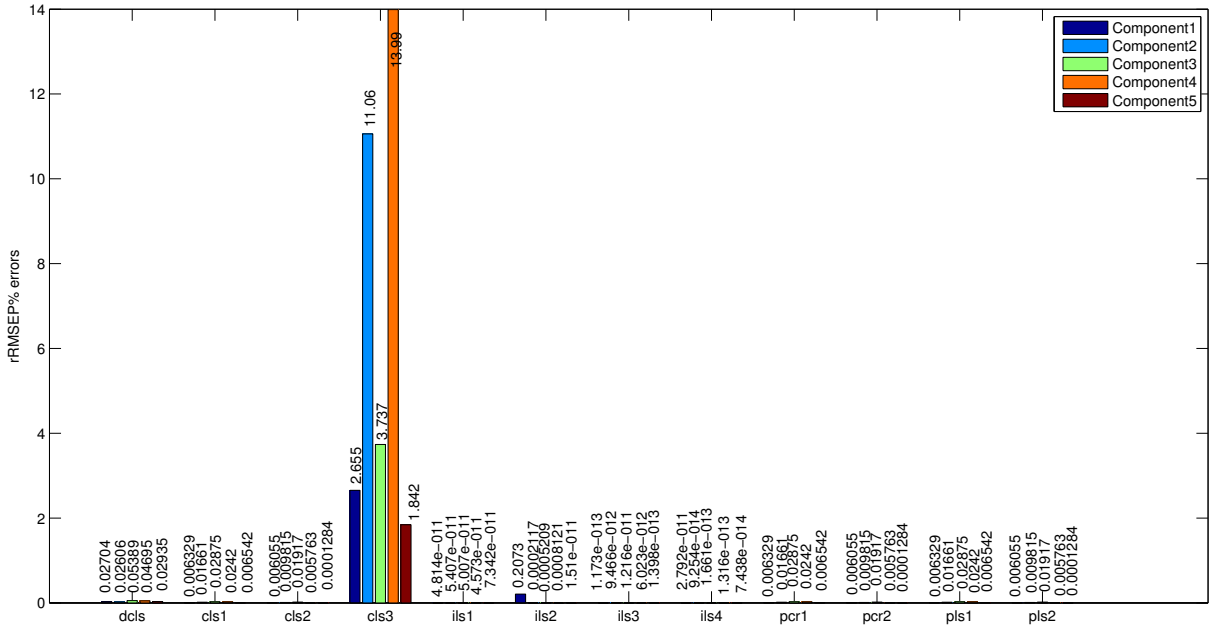


Figure 2: rRMSEP(%) error performance representation for 5 components of simulated spectra

In the following pages the experimental design for datasets of Cary and Simulated data is presented.

Table 3: Experimental Design for Methyl Orange - $CuSO_4$ dataset

Methyl Orange	**$CuSO_4$**
C_6	C_4
C_6	C_8
C_{11}	C_4
C_{11}	C_8
C_6	C_6
C_8	C_4
C_8	C_6
C_8	C_8
C_{11}	C_6

Table 4: Experimental Design for Thymol - $CoCl_2$ dataset

Thymol	**$CoCl_2$**
C_1	C_1
C_1	C_6
C_6	C_1
C_6	C_6
C_1	C_3
C_3	C_1
C_3	C_3
C_3	C_6
C_6	C_3

Table 5: Experimental Design for Thymol - Fast Green dataset

Thymol	**Fast Green**
C_1	C_4
C_1	C_8
C_6	C_4
C_6	C_8
C_1	C_6
C_3	C_4
C_3	C_6
C_3	C_8
C_6	C_6

Table 6: Experimental Design for $CoCl_2$ - Malachite Green dataset

$CoCl_2$	**Malachite Green**
C_1	C_5
C_1	C_{10}
C_6	C_5
C_6	C_{10}
C_1	C_7
C_3	C_5
C_3	C_7
C_3	C_{10}
C_6	C_7

Table 7: Experimental Design for Methylene Blue - $CuSO_4$ dataset

Methylene Blue	**$CuSO_4$**
C_1	C_4
C_1	C_8
C_5	C_4
C_5	C_8
C_1	C_6
C_3	C_4
C_3	C_6
C_3	C_8
C_5	C_6

Table 8: Experimental Design for Methylene Blue - Fast Green dataset

Methylene Blue	**Fast Green**
C_1	C_4
C_1	C_8
C_5	C_4
C_5	C_8
C_1	C_6
C_3	C_4
C_3	C_6
C_3	C_8
C_5	C_6

Table 9: Experimental Design for $CoCl_2$ - Methylene Blue dataset

$CoCl_2$	**Methylene Blue**
C_1	C_1
C_1	C_5
C_6	C_1
C_6	C_5
C_1	C_3
C_3	C_1
C_3	C_3
C_3	C_5
C_6	C_3

Table 10: Experimental Design for Methyl Orange - Fast Green - $CuSO_4$ dataset

Methyl Orange	**Fast Green**	**$CuSO_4$**
C_6	C_4	C_4
C_6	C_4	C_8
C_6	C_8	C_4
C_6	C_8	C_8
C_{11}	C_4	C_4
C_{11}	C_4	C_8
C_{11}	C_8	C_4
C_{11}	C_8	C_8

Table 11: Experimental Design for Thymol - Malachite Green - Methylene Blue dataset

Thymol	**Malachite Green**	**Methylene Blue**
C_1	C_5	C_1
C_1	C_5	C_5
C_1	C_{10}	C_1
C_1	C_{10}	C_5
C_6	C_5	C_1
C_6	C_5	C_5
C_6	C_{10}	C_1
C_6	C_{10}	C_5

Table 12: Experimental Design for 2 components, concentrations

C1 (M)	**C2** (M)
0.2	0.2
0.2	0.8
0.9	0.2
0.9	0.8
0.2	0.5
0.55	0.2
0.55	0.5
0.55	0.8
0.9	0.5

The pure components are obtained at C=1M. The individual spectra for the 1^{st} component are obtained at the following concentrations (in M): 0.2, 0.2875, 0.375, 0.4625, 0.55, 0.6375, 0.725, 0.8125, 0.9 and the individual spectra for the 2^{nd} component are obtained at the following concentrations (in M): 0.2, 0.275, 0.35, 0.425, 0.5, 0.575, 0.65, 0.725, 0.8

Table 13: Experimental Design for 3 components, concentrations

C1 (M)	**C2 (M)**	**C3 (M)**
0.9	0.8	1.2
0.9	0.8	0.8
0.9	0.2	1.2
0.9	0.2	0.4
0.2	0.8	1.2
0.2	0.8	0.4
0.2	0.2	1.2
0.2	0.2	0.4

The pure components are obtained at C=1M. The individual spectra for the 1st component are obtained at the following concentrations (in M): 0.2, 0.2875, 0.375, 0.4625, 0.55, 0.6375, 0.725, 0.8125, 0.9, the individual spectra for the 2nd component are obtained at the following concentrations (in M): 0.2, 0.275, 0.35, 0.425, 0.5, 0.575, 0.65, 0.725, 0.8 and the individual spectra for the 3rd component are obtained at the following concentrations (in M): 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2

Table 14: Experimental Design for 4 components, concentrations

C1 (M)	**C2 (M)**	**C3 (M)**	**C4 (M)**
1	0.9	0.8	1.2
1	0.9	0.2	0.4
1	0.2	0.8	0.4
1	0.2	0.2	1.2
0.3	0.9	0.8	0.4
0.3	0.9	0.2	1.2
0.3	0.2	0.8	1.2
0.3	0.2	0.2	0.4

The pure components are obtained at C=1M. The individual spectra for the 1^{st} component are obtained at the following concentrations (in M): 0.3, 0.3875, 0.475, 0.5625, 0.65, 0.7375, 0.825, 0.9125, 1, the individual spectra for the 2^{nd} component are obtained at the following concentrations (in M): 0.2, 0.2875, 0.375, 0.4625, 0.55, 0.6375, 0.725, 0.8125, 0.9, the individual spectra for the 3^{rd} component are obtained at the following concentrations (in M): 0.2, 0.275, 0.35, 0.425, 0.5, 0.575, 0.65, 0.725, 0.8 and the individual spectra for the 4^{th} component are obtained at the following concentrations (in M): 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2

Table 15: Experimental Design for 5 components, concentrations

C1 (M)	**C2 (M)**	**C3 (M)**	**C4 (M)**	**C5 (M)**
0.3	0.3	0.2	0.2	0.4
0.3	0.3	0.9	0.2	1.2
0.3	1	0.2	0.8	1.2
0.3	1	0.9	0.8	0.4
0.7	0.3	0.2	0.8	1.2
0.7	0.3	0.9	0.8	0.4
0.7	1	0.2	0.2	0.4
0.7	1	0.9	0.2	1.2

The pure components are obtained at C=1M. The individual spectra for the 1^{st} component are obtained at the following concentrations (in M): 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7 the individual spectra for the 2^{nd} component are obtained at the following concentrations (in M): 0.3, 0.3875, 0.475, 0.5625, 0.65, 0.7375, 0.825, 0.9125, 1, the individual spectra for the 3^{rd} component are obtained at the following concentrations (in M): 0.2, 0.2875, 0.375, 0.4625, 0.55, 0.6375, 0.725, 0.8125, 0.9, the individual spectra for the 4^{th} component are obtained at the following concentrations (in M): 0.2, 0.275, 0.35, 0.425, 0.5, 0.575, 0.65, 0.725, 0.8 and the individual spectra for the 5^{th} component are obtained at the following concentrations (in M): 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2