Prediction of Caffeine Content in Tea Using Near Infrared Spectroscopy and Multivariate Analysis Techniques

A Thesis submitted in partial fulfillment of the requirement for the Degree of

Master of Technology in
Instrumentation and Electronics Engineering

Submitted by

KUNDAN KUMAR MANDAL

Examination Roll No: M4IEE16-05

University Class Roll No: 001411103006

Registration no: 129479 of 2014-15

Under the Supervisions of

Dr. Bipan Tudu

Department of Instrumentation and Electronics Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata -700032

2016

Certificate of Recommendation

I hereby recommend that the thesis titled **"*Prediction of Caffeine in Tea Using Near Infrared Spectroscopy and Multivariate Analysis Techniques*"** carried out under my supervision by **Mr. Kundan Kumar Mandal** (Examination Roll No- M4IEE16-05, Registration No-**129479 of 2014-15**) may be accepted in partial fulfillment of the requirement for the degree of "Master of Technology in Instrumentation & Electronics Engineering" of Jadavpur University.

____

(Prof. Bipan Tudu)
Thesis Supervisor

Countersigned:

____

(Prof. Rajanikanta Mudi)
Head of the Department
Instrumentation and Electronics Engineering
Jadavpur University, Salt Lake Campus
Kolkata-700098

____

Prof. Sivaji Bandyopadhyay
(DEAN)
FET, Jadavpur University
Kolkata-700032

JADAVPUR UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY

<u>Certificate of Approval</u>

The foregoing thesis is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approved the thesis only for the purpose for which it is submitted.

....

Signature of Examiner ***Signature of Supervisor***

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who make it possible and whose constant guidance, suggestion and encouragement crown all the efforts with success. I therefore, acknowledge in this persons, as a token of my gratitude.

First of all, I am extremely grateful to my thesis guide, Dr. Bipan Tudu, Professor, Department of Instrumentation & Electronics Engineering, Jadavpur University for his valuable guidance, scholarly inputs and consistent encouragement I received throughout the thesis work. This feat was possible only because of the unconditional support provided by him. He has always made himself available to clarify my doubts despite his busy schedules and I am also highly indebted and grateful to him for pulling up my spirits in the most crucial and disheartening moments. His unflinching encouragement and support in various way provided the necessary impetus at crucial juncture of the project work without which the thesis would not been a reality. I am also thankful to him for shaping up the dimension of thesis work. I consider it as a great opportunity to do my thesis work under his guidance.

I gratefully acknowledge Dr. Rajib Bandyopadhyay, Professor, Department of Instrumentation and Electronics Engineering, Jadavpur University, for his constant motivation, suggestions, and giving right directions at decisive stages of the research work. I am also indebted to him for providing me the necessary platform for carruing forward to work.

I am grateful and beholden much to Dr. Rajanikanta Mudi, Professor and Head of Department of Instrumentation and Electronics Engineering, Jadavpur University, for his valuable advices and support.

I am grateful and beholden much to Mr. somdeb Chanda, Mr. Hemanta Naskar and Mr. Dilip Singh, Research Scholars, Department of Instrumentation and Electronics Engineering, Jadavpur University, for their valuable advices and support.

I would like to convey my gratitude to all the Faculty members, technical and non technical staff specially Mr. Goutam Majumder in the Instrumentation & Electronics

Engineering Department of Jadavpur University for their help whenever I need it to complete the entire thesis work smoothly.

Finally, my parents deserve special mention for their inspirable support and prayer that formed my psychological backbone. A word of thanks goes to my parents and friends for providing the necessary atmosphere of understanding, encouragement, moral support and constant help towards the successful execution of the project

Date:

(Kundan Kumar Mandal)

Contents

Page No.

Chapter 4 Instrument Review and Multivariate Analysis of Data 69-91

Synopsis of the Thesis

Near infrared spectroscopy is a fast, non-destructive and versatile technique with almost no sample preparation needed. Recent days it is being used in various field for quantitative and quality detection of a material like various chemical content in food products (tea, rice, milk, vegetables, wine etc...) along with few medical purposes. Medical applications, including those related to blood glucose measurements, tissue and major organ analysis, fatal analysis, and cancer research. Due to the versatility of NIR spectroscopy it become one of the emerging technology in recent few year and seeks the attention of researcher to work with it upgrade its application.

On the other hand, tea is the most consumable beverage throughout the world next to water. Due to the huge demand, tea industries become more and more concern regarding its quality. The main ingredient of tea is caffeine and polyphenol which affects human health, so industries need to be much conscious regarding these chemical content before providing it to consumers.

The prediction of chemical content could be done with various chemical methods but the downside of these conventional method is they are time consuming and a series of sample preparation is needed which makes the analysis complicated and cost effective.

NIR overcomes above demerits and provides a better option as it is a simple, quick (>30microsecond), non-destructive technique that provide multi-constituent analysis with good accuracy and precision. As NIR spectroscopy combined with multivariate analysis prediction model is a better and emerging technique for the qualitative and quantitative estimation of chemical compound in matters so it was tried to estimate quantity of caffeine in tea sample taken from the tea garden of Asam. A large number of NIR spectrums have been obtained through NIR spectrometer and principal component analysis (PCA) was applied on the observed data sample. Once principal components have been obtained, it can be used to construct a regression model. Multivariate analysis technique (such as principal component regression (PCR) and partial least square (PLS) regression analysis) has applied to the obtained NIR data sample. Finally optimized prediction model has been selected for the prediction of unknown caffeine containing sample. Since principal component regression uses principal component of independent variable only so its accuracy is not so good whereas in care partial least square regression method accuracy is obtained up to an acceptable limit.

Outline of Thesis

1. **Chapter one**, section 1.1 explains various varieties of tea and their benefits. Section 1.2 explains why NIR spectroscopy is used for prediction analysis whereas section 1.3 and 1.4 discussed chemometrics, pre-processing techniques respectively. Literature review has been discussed in section 1.5. Objective of thesis and conclusion of above topics is mentioned in 1.6 and 1.8
2. **Chapter two**, section 2.1 contains definition of NIR spectroscopy. Section 2.2 contains theory of vibrational spectroscopy under which range of electromagnetic spectrum, bond vibrations, harmonic and anharmonic oscillator, vibrational degree of freedom and overtones and recombination in NIR spectrum region have been discussed. Section 2.3 contains NIR spectral region and section 2.4 contains Measurement methods in NIR spectroscopy under which Dispersive Infrared Spectrometer, Fourier-Transform Infrared Spectrometer, Michelson Interferometer Source and Detector, Fourier Transform, Moving Mirror, Signal averaging, Computer and Spectra of FT-NIR have been discussed. Section 2.5 contain various transmittance mode (liquid, solid and gas). Section 2.6 and 2.7 contains comparison among NIR, MIR, Raman spectroscopy and conclusion of this chapter.

3. **Chapter three** discussed about various multivariate analytical technique where section 3.1 contain principal component analysis under which Statistics behind PCA; Standard deviation, Variance, Covariance, Covariance Matrix, Matrix algebra and Principal Component Analysis has been discussed. Section 3.2 contain Principal component regression under which Principal component regression model, Principal component regression model for unknown data prediction, Pre-processing of raw data, Reduction of non-linearity, Noise reduction and differentiation, Methods specific for NIR, Selection of pre-processing methods in NIR,PCA and building the model, Model optimisation and validation, Training, optimisation and validation, Measures of predictive ability, Optimization, Validation, External validation, Internal validation and Final model of PCR have been discussed. Section 3.3 talked about Partial least square regression technique, the subsections are Introduction, How Does PLS Work,

Goal, Simultaneous decomposition of predictors and dependent variables, PLS regression and covariance, A PLS regression algorithm, PLS regression and the singular value decomposition.

4. **Chapter four** contains data observation where section 4.1. Instrument review under which DWARF-Star NIR Detector, Spectroscopy lamps and light sources, VIS-NIR sources (SL1 Tungsten Halogen), Tungsten lamp have been discussed. Section 4.2 is NIR Experiment which contains: Data absorbing using Instrument, Principal component analysis of data, Principal component analysis of ten samples and Principal component analysis of all samples. Section 4.3 PCR on data under which PCR steps, Averaging the raw data and standardization, Principal component analysis, Leave One Out, Cross-Validation (LOOCV), Optimisation and Sample selection, Model formation, Conclusion have been discussed. Section is Partial Least Square Regression under which, Steps of PLS regression model, Averaging the raw data and standardization, Leave One Out Cross-Validation (LOOCV), Optimisation and Sample selection, Model formation and Conclusion of obtained PLS model has been discussed.
5. **Chapter five** contains Model accuracy and future scopes of NIR spectroscopy combined with Multivariate Regression Analysis Technique.

List of Figures

Page No.

List of Tables

Page No.

List of Abbreviations

AA	Antioxidant Activity
ACO-iPLS	Ant Colony Optimization Interval Partial Least Squares
BP-ANN	Back Propagation Artificial Neural Network
CA	Cluster Analysis
CBF	Cerebral Blood Flow
CBV	Cerebral Blood Volume
COV	Covariance
DTGS	Deuterium Tryglycine Sulfate
EC	Epicatechin
ECG	Epicathin-3-Gallate
EGC	Epigallocatechin
EGCG	Epigallocatechin-3-Gallate
EPMA	Electron Probe Micro-Analyzer
FTIRS	Fourier Transform Infrared Spectroscopy
FT	Fourier Transform
FFT	Fast Fourier-transformation
LA-ICP-MS	Laser Ablation Inductively Coupled Plasma Mass Spectrometry
LDA	Linear Discriminant Analysis
MVA	Multivariate data Analysis
MCR	Multivariate Curve Resolution
MCT	Mercury Cadmium Telluride
MSC	Multiplicative Scatter Correction
NIR	Near Infrared
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS-R	Partial Least Square Regression
RMSECV	Root Mean Square Error of Cross-Validation
RMSEP	Root Mean Square Error of Prediction
SNV	Standard Normal Variate
SD	Standard Deviation
SVMR	Support Vector Machine Regression

SIMCA	Soft Independent Modelling of Class Analogy
SEM	Scanning Electron Microscopy
TEAC	Trolox Equivalent Antioxidant Capacity
UV	Ultra Violet
VAR	Variance

Chapter 1

Preamble and Scope of the Thesis

1.0. Introduction

Drinking tea, punctuates our days with precious refreshing pauses, whether it is after a satisfying meal or when taking a much needed break in our busy schedule. Due to the huge demand of tea in today's world, it seems tea is the most consumable beverage next to water.

The food industries has to play an important role here by taking care of food safety and quality of beverages like tea, since it is directly related to people's health and social progress. Consumers are gradually looking for quality seals and trust marks on food products, and expect manufacturers and retailers to provide products of high quality. The quality of foods basically mean that the adequate quantity of chemical compounds which is beneficial for human health. All of these factors have influenced the need for reliable techniques to evaluate the food quality by analysing the quantity of chemical compounds. Considering the demands in practice, it is more necessary to develop fast, efficient and non-destructive method such as Near infrared (NIR) spectroscopy technology to accomplish the food quality detection.

NIR Specroscopy is a fast, non-destructive technique used for qualitative as well as quantitative detection of a sample. Since it uses infra-red ray to obtain spectrum of sample so it is non-destructive and almost no sample preparation is needed here. NIR spectroscopy has been used to quantify several compounds having a diverse antioxidant capacity such as carotenoids, polyphenols, fatty acids and glucosinolates in a wide range of food commodities (for example, wine, dairy products, tea, fruit, vegetables, herbs, spices and cereals).

The development and implementation of infrared (IR) methods for the analysis of antioxidants has been made possible by the development of multivariate data analysis (MVA) methods and techniques in the so called field of chemometrics.

Predictive methods (calibration) such as principal component regression (PCR) and partial least squares (PLS) regression are now widely used in the development of IR analytical methods for the prediction of chemical content in a wide range of products. Due to the versatile behaviour of NIR spectroscopy it has been tried to develop few prediction models which help us to predict the quantity of caffeine in unknown tea samples with an acceptable

accuracy. As we are concern about the tea sample so it is necessary to understand the possible chemical composition of tea samples along with what are the chemical content affect human life. Caffeine, one of the important ingredients of tea affects human body. So this thesis is concerned regarding its concentration in tea sample.

1.1. Tea

Several researches have been proven that tea is not only used as stimuli but has medicinal property too. The various constituent of tea are as follow:

- ❖ Antioxidants: Present in white, green and black tea.
  Example Polyphenols
  Tea polyphenols accounts for 30-42% of dry weight of green tea leaves. The main polyphenol in green tea are epicatechin(EC),epicathin-3-gallate(ECG), epigallocatechin(EGC), and epigallocatechin-3-gallate (EGCG).
- ❖ Caffeine: Tea contains caffeine in wide varying amount, the potency usually depends on the maturity of leaves and the region where they were grown and how they were processed, shipped and stored.
- ❖ Theine: Similar to caffeine.
- ❖ Tannin: The substance in tea which give bitter taste and most of its colour.
- ❖ Theophylline: It is similar to caffeine and present in minimal amount. Cocoa beans contains in larger amounts.
- ❖ Herbal teas: Herbal teas may not contain caffeine, tannin and theine. They are having various ingredient which beneficial for health.

1.2. Tea Variety and Benefits [2]

Most of the variety of tea comes from *Camellia Sinensis* plant. They differ from each other only in the sense how they are grown, processed, climate and geographical availabilities. *Camellia Sinensis* plant basically originated in Asian regions, bur due to the green revolution and globalisation it is found throughout the world. Recent days more than 3000 variety is cultivated and consumed throughout the world, so tea become most consumable beverage throughout the world next to water.

1.2.1. Types of Tea

Tea can be categorised basically in five types: white, black, green, oolong, and puer tea. All varieties are found and originate based on the geographical region and environment accordingly.

1.2.1.1. White Tea

It is one of the best category of tea, and famous for its natural sweetness, complexity and subtlety. They are hand processed, usually youngest shoots of tea plant is taken by escaping it from oxidation. It is brewed with very care at low temperature and a short interval of time. If steeping with hotter temperature is done, more caffeine would be extracted but it’s ok to have a little more caffeine than other tea.

White tea contains cetechins which may help fight cancer and cardiovascular disease. Drinking white tea might also reduce the risk of cancer recurrence for breast cancer survivors, according to the American Cancer Society.

1.2.1.2. Black Tea

Black tea has higher caffeine content then others (about 50-56%). Full oxidation takes place in leaves so its colour is dark brown or black. It is famous for its bitter and robust test as compare to other due to its high caffeine content.

Black tea is one of the most highly caffeinated varieties of tea, with about 40 milligrams of caffeine per cup. Black tea also contains thearubigins and theaflavins, two types of antioxidants that have been very useful to lower cholesterol levels.

1.2.1.3. Green Tea

After being picked only it allows to wither. The leaves are heated rapidly quickly so oxidation process stopped and it contain less amount of caffeine since it is brewed at less temperature and time. It provides more subtle flavours with many undertones and accents that connoisseurs treasure. It is helpful to decrease the risk of cardiovascular diseases.

1.2.1.4. Oolong Tea

It has caffeine content between green tea and black tea due to partial oxidation taken place here. The flavour of oolong teas is typically not as robust as blacks or as subtle as greens, but has its own extremely fragrant and intriguing tones.

It is useful to lower fat in body.

1.2.1.5. Puer Tea

It is perhaps the most mysterious of all tea. It is an aged black tea. It is very strong with an incredibly deep and rich flavour, and no bitterness.

Above mention tea are the basic varieties of tea, each variety contains a huge number of tones and flavours along with various medicinal benefits. Due to availability of different taste, fragments and benefits the consumers are seeking for better products. To fill their demands researcher in tea industries seeks one’s interest towards qualitative and quantitative analysis of tea products.

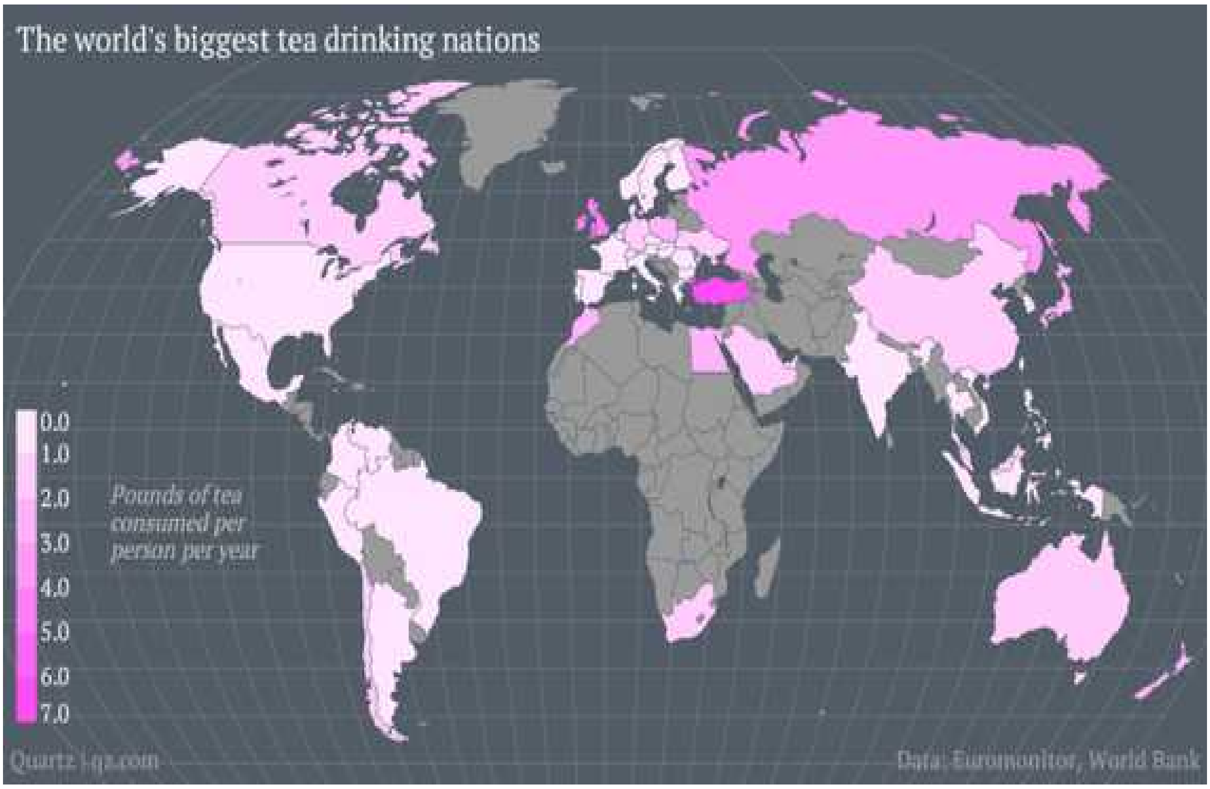

The world's biggest tea drinking nations
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Pounds of tea consumed per person per year
Quartz | qz.com
Data: Euromonitor, World Bank

Figure 1.1. Consumption of tea throughout the world [17].

1.3. Near Infrared Spectroscopy

NIR spectroscopy is based on the absorption of electromagnetic radiation at wavelength in the range 780–2500 nm. NIR spectra of foods comprise broad band arising from overlapping absorptions corresponding mainly to overtones and combinations of vibrational modes involving C–H, O–H, and N–H chemical bond. This makes it very feasible for measurements to be made in organic and biological systems. Radiation interacting with a sample may be absorbed, transmitted or reflected. Thus, there are different NIR spectroscopy measurement modes fitting different applications. In practice, the common modes are transmittance, interactance, transflectance, diffuse transmittance, and diffuse reflectance, with the last two being most frequently used.

In the wavelength range 1100–2500 nm, the amount of scattering makes the path length so high that transmittance through 1 cm thickness of most samples is negligible. This situation is called diffuse reflectance because most of the incident radiation is reflected. This measurement is suitable for thicker samples such as fruits and wheat power.

Near infrared (NIR) spectroscopy has proved to be a powerful analytical tool used in the agricultural, nutritional, petrochemical, textile and pharmaceutical industries. Since the 1990s, attempts have been made to simultaneously predict water, alkaloids and phenolic substance content in tea leaves using NIR spectroscopy.

1.4. Chemometrics

Chemometrics is the use of mathematical and statistical methods to improve the understanding of chemical information and to correlate quality parameters or physical properties to analytical instrument data. Patterns in the data are modelled, these models can then be routinely applied to future data in order to predict the same quality parameters. The result of the chemometrics approach is gaining efficiency in assessing product quality. It can lead to more efficient laboratory practices or automated quality control systems. The only requirements are an appropriate instrument and software to interpret the patterns in the data.

Spectroscopists need to use the following methods within a chemometrics software package to explore their data:

- Principal Component Analysis (PCA)

- Regression (PLS, PCR, MLR, 3-way PLS) and Prediction
- SIMCA and PLS-DA Classification
- Design of Experiments
- ANOVA and Response Surface Methodology
- Multivariate Curve Resolution (MCR)
- Clustering (K-Means)

Spectroscopy methods of chemical analysis are excellent for the application of chemometric methods, because the measurements at many different wavelengths provide inherently multivariate data. The chemist generally requires three categories of information from specimens under investigation: quantitative data, qualitative data, and fundamental information on the properties of the material. Spectroscopy has long been used for all three purposes; the recent application of chemometric algorithms has assisted greatly in these endeavors. Although there is some overlap, three chemometric methods correspond to the three types of information: multiple regression, discriminant analysis, and principal components analysis.

NIR spectroscopy with the predictive methods such as PCR and PLS algorithm was used to determine simultaneously some chemical compositions contents from tea such as caffeine and total polyphenols contents.

1.5. Pre-processing Methods

Proper application of spectroscopic data pre-processing, to reduce and correct interferences such as overlapped bands, baseline drifts, scattering, and path-length variation. Multiplicative Scatter Correction (MSC) pre-treatment is recommended to build reliable relationship between wheat protein content and spectral data, for scatter correction.

1.6. Literature Survey

In the past few year lots of paper has been presented on quantitative and qualitative analysis of tea sample using NIR spectroscopy. Different research used different approaches to reach their goal. Some of papers have been discussed below.

- H. owen-Reece, M.Smith, C. E. Wlwell abd J. C. Goldstone has work on the paper [1] Near Infrared Spectroscopy and given an overview on the physical prin-ciples involved in the measurement of tissue oxygenation with NIRS and describes how it may be used to continuously observe changes in cerebral haemodynamics and oxygen- ation. The clinical experience with NIRS is described in addition to the techniques for non-invasive measurement of cerebral blood flow (CBF) and cerebral blood volume (CBV). Absorption of light by coloured compounds, Biological absorbers, Differential path length factor, Equipment design of NIR spectroscopy for tissue has been described in this paper. Overall NIR spectroscopy is a better option for analysing tissue sample.
- Quansheng Chen, Jiewen Zhao, Xingyi Huang, Haidong Zhang , Muhua Liu has discussed in his paper [3] on Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy about the possibility to use near infrared (NIR) spectroscopy as a rapid method to predict quantitatively the content of caffeine and total polyphenols in green tea. A partial least squares (PLS) algorithm is used to perform the calibration. To decide upon the number of PLS factors included in the PLS model, the model is chosen according to the lowest root mean square error of cross-validation (RMSECV) in training. The correlation coefficient R between the NIR predicted and the reference results for the test set is used as an evaluation parameter for the models. The result showed that the correlation coefficients of the prediction models were R=0.9688 for the caffeine and R=0.9299 for total polyphenols. The study demonstrates that NIR spectroscopy technology with multivariate calibration analysis can be successfully applied as a rapid method to determine the valid ingredients of tea to control industrial processes.
- Chunmei Li and Bijun Xie (Department of Food Science and Technology, Huazhong Agriculture University, Wuhan, Hubei 430070, People's Republic of China) has explainer in their paper [4] about Evaluation of the Antioxidant and Pro-oxidant Effects of Tea Catechin Oxy-polymers. The paper discussed about how Tea catechin oxypolymers (TCOP) were prepared by oxidizing tea catechin with H_2O_2. It is also described that the scavenging effects of TCOP to both the hydroxyl radical and superoxide radical were stronger than that of TC, and also they had no pro-oxidant effect. Finally it was concluded that the antioxidant activity of TCOP was not less than or even more notable than that of TC.

- V.R. Sinija, H.N. Mishra (Agricultural & Food Engineering Department, Indian Institute of Technology, Kharagpur, West Bengal 721 302, India) has discussed in their paper [5] FT-NIR spectroscopy for caffeine estimation in instant green tea powder and granules. This article discussed the feasibility of measuring caffeine content in instant green tea and granules were investigated by Fourier Transform Near-Infrared (FT-NIR) spectroscopic technique. A calibration model was developed using pure caffeine standards of varying concentrations in the near-infrared region (4000–12000 cm_1). The developed model was validated using test validation technique. FT-NIR spectroscopy with chemometrics, using the PLS–first derivative plus straight line subtraction method could predict the caffeine content in tea samples accurately up to an R2 value greater than 0.98 and a standard error of prediction (SEP) value less than 2.0 with 6 factors in the prediction model. The developed model was applied to predict caffeine content in tea samples within 2–5 min.
- Antonio Moreda-Pineiro, Andrew Fisher, Steve J. Hill has discussed in paper [6] on the classification of tea according to region of origin using pattern recognition techniques and trace metal data Pattern recognition techniques were then used to classify the tea according to its geographical origin. Principal component analysis (PCA) and cluster analysis (CA), as exploratory techniques, and linear discriminant analysis (LDA) and soft independent modelling of class analogy (SIMCA), were used as classification procedures. In total, 17 elements (Al, Ba, Ca, Cd, Co, Cr, Cu, Cs, Mg, Mn, Ni, Pb, Rb, Sr, Ti, V, Zn) were determined in a range of 85 tea samples (36 samples from Asian countries, 18 samples from African countries, 24 commercial blends and seven samples of unknown origin). Natural groupings of the samples (Asian and African teas) were observed using PCA and CA. The application of LDA gave correct assignation percentages of 100.0% and 94.4% for the African and Asian teas, respectively, at a significance level of 5%. SIMCA offered percentages of 100.0% and 91.7% for African and Asian groups, respectively, at the same significance level. LDA, also at a significance level of 5%, allowed a 100% of correct case identification for the three classes China, India and Sri Lanka. the data obtained demonstrates that it is possible to assign unknown samples and tea blends according to their country of origin.
- Yong He, Xiaoli Li, Xunfei Deng has discussed in their paper [7] Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model about Visible/near-infrared spectroscopy (NIRS), with the characteristics of

high speed, non-destructiveness, high precision and reliable detection data, etc., is a pollution-free, rapid, quantitative and qualitative analysis method. A new approach for discrimination of varieties of tea by means of VIS/NIR spectroscopy (325–1075 nm) was developed in this work. The relationship between the reflectance spectra and tea varieties was established. The spectral data was compressed by the wavelet transform (WT). The features from WT can be visualized in principal component (PC) space, which can lead to discovery of structures correlative with the different class of spectra samples. It appeared to provide a reasonable clustering of the varieties of tea. The scores of the first eight principal components computed by PCA had been applied as inputs to a back propagation neural network with one hidden layer. The 200 samples of eight varieties were selected randomly to build BP-ANN model. This model was used to predict the varieties of 40 unknown samples. The recognition rate of 100% was achieved. This model comes to be reliable and practicable.

- ❖ Derya Kara (Department of Chemistry, Art and Science Faculty, Balikesir University, 10100 Balikesir, Turkey) has published in his paper [8] Evaluation of trace metal concentrations in some herbs and herbal teas by principal component analysis about Sixteen trace metallic analyte (Ba, Ca, Ce, Co, Cr, Cu, Fe, K, La, Mg, Mn, Na, Ni, P, Sr and Zn) in acid digests of herbal teas were determined and the data subjected to chemometric evaluation in an attempt to classify the herbal tea samples. Nettle, Senna, Camomile, Peppermint, Lemon Balm, Sage, Hollyhock, Linden, Lavender, Blackberry, Ginger, Galangal, Cinnamon, Green tea, Black tea, Rosehip, Thyme and Rose were used as plant materials in this study. Trace metals in these plants were determined by using inductively coupled plasma-atomic emission spectrometry and inductively coupled plasma-mass spectrometry. Principal component analysis (PCA), linear discriminant analysis (LDA) and cluster analysis (CA) were used as classification techniques. About 18 plants were classified into 5 groups by PCA and all group members determined by PCA are in the predicted group that 100.0% of original grouped cases correctly classified by LDA. Very similar grouping was obtained using CA.
- ❖ Sheida Makvandi, Massoud Ghasemzadeh-Barvarz, Georges Beaudoin , Eric C. Grunsky, M. Beth McClenaghan, Carl Duchesne hsa been bublished a peper [9] on Principal component analysis of magnetite composition from volcanogenic massive sulfide deposits: Case studies from the Izok Lake (Nunavut, Canada) and Halfmile Lake (New Brunswick, Canada) deposits about the analysis where Magnetite grains

from the Izok Lake (Nunavut, Canada) and the Halfmile Lake (New Brunswick, Canada) volcanogenic massive sulfide deposits, and from till covering the nearby areas were investigated using the scanning electron microscopy (SEM), electron probe micro-analyzer (EPMA), laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS), and optical microscopy. The method of robust estimation for compositional data (rob-composition) was applied to censored geochemical data, and the results were analyzed by principal component analysis (PCA). Textural relationships and mineral association of magnetite reveal the history of formation, and contribute to the explanation of characteristic compositional differences of magnetite from different geological settings. The integration of petrography and mineral chemistry allows discriminating magmatic, metamorphic and hydrothermal magnetite grains in the VMS deposits bedrock samples. Magmatic magnetite is found in Izok Lake gabbro, and Halfmile Lake syenite, felsic ash tuff and gossan samples, whereas magnetite in Izok Lake massive sulfides, gahnite-rich dacite and iron formations formed during the amphibolite facies metamorphism. In Halfmile Lake andesite, magnetite recrystallized during greenschist facies metamorphism. In the magnetite alteration zone associated to the Halfmile Lake deposit, hydrothermal magnetite has been overprinted by metamorphic magnetite. Halfmile Lake massive sulfides in chloritic argillite contain hydrothermal magnetite. PCA identifies discriminator elements and their contributions to magnetite composition fromdifferent Izok Lake Lake and Halfmile Lake bedrock samples. The results suggest that Si, Ca, Zr, Al, Ga, Mn, Mg, Ti, Zn, Co, Ni and Cr are discriminator elements for VMS deposits and their host bedrocks. The distinct chemical signatures for magnetite fromvarious bedrock lithologies demonstrate that magnetite grains of the same origin share more similarities in chemistry, as high Ti indicates magmatic sources for magnetite, whereas high Si, Ca and Mg are indicative of hydrothermal settings. Variable compositions of metamorphic magnetite suggest that the chemistry of this type of magnetite is controlled by the composition of host rocks, the grade of metamorphism and oxygen fugacity. PCA of EPMA and LA-ICP-MS data of magnetite from the Izok Lake and Halfmile Lake bedrock samples yield discrimination models for classification of magnetite grains from till. Decreases in the proportion of magnetite grains with the chemical signature of the Izok Lake massive sulfides and gahnite-rich dacite down-ice from the Izok Lake deposit show the use of magnetite chemistry in geochemical exploration. In the Halfmile Lake area, till magnetite grains with the

signature of VMS mineralization make a glacial dispersal train more than 2 km down-ice from the deposit.

- ❖ Quansheng Chen, Jiewen Zhao, Haidong Zhang, Xinyu Wang has been described in their paper [10] Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration about the feasibility to use near infrared (NIR) spectroscopy as a rapid analysis method to qualitative and quantitative assessment of the tea quality. NIR spectroscopy with soft independent modeling of class analogy (SIMCA) methodwas proposed to identify rapidly tea varieties in this paper. In the experiment, four tea varieties from Longjing, Biluochun, Qihong and Tieguanyin were studied. The better results were achieved following as: the identification rate equals to 90% only for Longjing in training set; 80% only for Biluochun in test set; while, the remaining equal to 100%. A partial least squares (PLS) algorithm is used to predict the content of caffeine and total polyphenols in tea. The models are calibrated by cross-validation and the best number of PLS factors was achieved according to the lowest root mean square error of cross-validation (RMSECV). The correlation coefficients and the root mean square error of prediction (RMSEP) in the test set were used as the evaluation parameters for the models as follows: R = 0.9688, RMSEP = 0.0836% for the caffeine; R = 0.9299, RMSEP = 1.1138% for total polyphenols. The overall results demonstrate that NIR spectroscopy with multivariate calibration could be successfully applied as a rapid method not only to identify the tea varieties but also to determine simultaneously some chemical compositions contents in tea.
- ❖ M.H. Zhang, J. Luypaert, J.A. Fernández Pierna, Q.S. Xu1, D.L. Massart has been analysed in the paper [11] Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration about A principal component regression (PCR) model is built for prediction of total antioxidant capacity in green tea using near-infrared (NIR) spectroscopy. The modelling procedures are systematically studied with the focus on outlier detection. Different outlier detection methods are used and compared. The root mean square error of prediction (RMSEP) of the final model is comparable to the precision of the reference method.
- ❖ H. Schulz, U. H. Engelhardt, A. Wegent, H. H. Drews, and S. Lapczynski has been published in their paper [12] Application of Near-Infrared Reflectance Spectroscopy to the Simultaneous Prediction of Alkaloids and Phenolic Substances in Green Tea Leaves about near-infrared reflectance spectroscopic (NIRS) method for the

prediction of polyphenol and alkaloid compounds in the leaves of green tea [Camellia sinensis (L.) O. Kuntze] was developed. Reference measurements of the individual catechins, gallic acid, caffeine, and theobromine were performed by reversed-phase HPLC. The total polyphenols were determined according to the colorimetric Folin-Ciocalteu assay. Using the partial least-squares algorithm, very good calibration statistics were obtained for the prediction of gallic acid, (-)-epicatechin, (-)-epigallocatechin, (-)-epicatechin gallate, (-)-epigallocatechin gallate, caffeine, and theobromine (R2 > 0.85) with standard deviation/standard error of cross-validation (SD/SECV) ratio ranging from 2.00 to 6.27. Simultaneously, the dry matter content of the tea leaves can be analyzed very precisely (R2) 0.94; SD/SECV) 4.12). Furthermore, it is possible to discriminate tea leaves of different age by principal component analysis on the basis of the received NIR spectra. Prediction of the total polyphenol content is performed with a lower accuracy, which might be due to the lack of specificity in the colorimetric reference method. The study demonstrates that NIRS technology can be successfully applied as a rapid method not only for breeding and cultivation purposes but also to estimate the quality and taste of green tea and to control industrial processes, for example, decaffeination.

- J. Luypaert, M.H. Zhang, D.L. Massart has been published in their paper [13] Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.) about the possibility to use near infrared spectroscopy (NIR) combined with PLS as a rapid method to estimate the quality of green tea. NIR is used to build calibration models to predict the content of caffeine, epigallocatechin gallate (EGCG) and epicatechin (EC) and for the prediction of the total antioxidant capacity of green tea. For the determination of the total antioxidant capacity, the trolox equivalent antioxidant capacity (TEAC) method is used. Until now, the prediction of the antioxidant capacity as such by use of NIR has not been reported. For caffeine and TEAC, models are build for the whole green tea leaves and also for the ground leaves. For the polyphenols (EGCG and EC), only models for the whole leaves are investigated. A partial least squares (PLS) algorithm is used to perform the calibration. To decide upon the number of PLS factors included in the PLS model, the model with the lowest root mean square error of cross-validation (RMSECV) for the training set is chosen. The correlation coefficient (r) between the predicted and the reference results for the test set is used as an evaluation parameter for the models: for the TEAC results r = 0.90 for the model with the whole leaves, r = 0.86 for the model with the powdered leaves are obtained. The caffeine prediction model has a correlation coefficient r = 0.96 for the whole leaves and r = 0.93 for the ground leaves. The correlation coefficient for the EGCG and the EC content models are, respectively 0.83 and 0.44.

- Huang Xiaowei, Zou Xiaobo, Zhao Jiewen, Shi Jiyong, Zhang Xiaolei, Mel Holmes has been described in their paper [14] Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models about how colony optimization models is used for the measurement of chemical compound in green tea it was described that Flowering tea has become a popular beverage consumed across the world. Anthocyanins content is considered as an important quality index of flowering tea. The feasibility of using near infrared (NIR) spectra at the wavelength range of 10,000–4000 cm^{-1} for rapid and non-destructive determination of total anthocyanins content in flowering tea was investigated. Ant colony optimization interval partial least squares (ACO-iPLS) and Genetic algorithm interval partial least squares (GA-iPLS) were used to develop calibration models for total anthocyanins content. The optimal ACO-iPLS model for total anthocyanins content (R = 0.9856, RMSECV = 0.1198 mg/g) had better performance than full-spectrum PLS, iPLS, and GA-iPLS models. It could be concluded that NIR spectroscopy has significant potential in the nondestructive determination of total anthocyanins content in flowering tea.
- Guangxin Ren, Shengpeng Wang, Jingming Ning, Rongrong Xu, Yuxia Wang, Zhiqiang Xing, Xiaochun Wan, Zhengzhu Zhang has described in their paper [15] Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS) about use of NIR spectroscopy with chemometric tools was utilized as a rapid analysis method to assess quality and to differentiate geographical origins of black tea. A partial least squares (PLS) algorithm was employed for the calibration of models predicting the levels of caffeine, water extract, total polyphenols, and free amino acids, while a factorization method was proposed to trace black tea from different geographical origins. In the calibration set, the root mean squared error of cross validation (%) and the correlation coefficient (R) for caffeine, water extracts, total polyphenols and free amino acids were 0.102%, 0.654%, 0.552%, and 0.248% and 0.983, 0.977, 0.975, and 0.943, respectively. In the prediction set, the root mean squared error of prediction and R for the corresponding

constituents were 0.160%, 0.685%, 0.594%, and 0.273% and 0.955, 0.962, 0.954, and 0.927, respectively. The identification accuracy for black tea from different geographical origins reached 94.3%. This study demonstrated that NIR spectroscopy can be successfully applied to rapidly determine the main chemical compositions and geographical origins of black tea.

- ❖ Quansheng Chena, Zhiming Guo, Jiewen Zhao, Qin Ouyang has been published in the paper [16] Comparisons of different regressions tools in measurement of antioxidant activity in green tea using near infrared spectroscopy that near infrared (NIR) spectroscopywas employed with the help of a regression tool for rapid and efficient measurement antioxidant activity (AA) in green tea. Three different linear and nonlinear regressions tools (i.e. partial least squares (PLS), back propagation artificial neural network (BP-ANN), and support vector machine regression (SVMR)), were systemically studied and compared in developing the model. The model was optimized by a leave-one-out cross-validation, and its performance was tested according to root mean square error of prediction (RMSEP) and correlation coefficient (Rp) in the prediction set. Experimental results showed that the performance of SVMR model was superior to the others, and the optimum results of the SVMR model were achieved as follow: RMSEP = 0.02161 and Rp = 0.9691 in the prediction set. The overall results sufficiently demonstrate that the spectroscopy coupled with the SVMR regression tool has the potential to measure AA in green tea.

1.7. Conclusion

NIR spectroscopy is a fast non-destructive technique where almost no sample preparation is required. NIR spectrometer is used to obtain spectrum of tea sample and with these samples first a calibration model would be developed with the known reference value of Caffeine then by prediction model, the unknown tea sample of Caffeine would be predicted. PCA, PCR and PLS regression models have been explored for calibration and prediction purpose. PCA is used for dimension reduction and segregation of sample in a plot. PCR model is constructed through the use of PC and then optimized model is selected and prediction is done.

Prediction of PCR model is not good enough so, next PLS model have been formed and it was found that its accuracy is good enough with low RMSEP value. The most noticeable thing during the analysis is once an efficient model is constructed then no need to go for other time consuming technique to quantify chemical contents in sample. Prediction of chemical contents can be done by obtaining spectrum from NIR instruments and converting them into a model compatible raw data. Finally the optimized model can be used to predict chemical content.

References

[1] H. Owen-Reece, M. Smith1, C. E. Elwell and J. C. Goldstone3 “Near infrared spectroscopy”, British Journal of Anaesthesia, 82(3) (1999) 418-26.

[2] http://www.teasource.com/pages/types-of-tea.

[3] Quansheng Chen, Jiewen Zhao, Xingyi Huang, Haidong Zhang , Muhua Liu “Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy”. Microchemical Journal 83 (2006) 42–47.

[4] Chunmei Li and Bijun Xie (Department of Food Science and Technology, Huazhong Agriculture University, Wuhan, Hubei 430070, People’s Republic of China) “Evaluation of the Antioxidant and Pro-oxidant Effects of Tea Catechin Oxypolymers”, J. Agric. Food Chem. 2000, 48, 6362-6366.

[5] V.R. Sinija, H.N. Mishra (Agricultural & Food Engineering Department, Indian Institute of Technology, Kharagpur, West Bengal 721 302, India) “FT-NIR spectroscopy for caffeine estimation in instant green tea powder and granules”, LWT - Food Science and Technology (2009), Vol. 42, No. 5 pp. 998-1002.

[6] Antonio Moreda-Pineiro, Andrew Fisher, Steve J. Hill “The classification of tea according to region of origin using pattern recognition techniques and trace metal data”, Journal of Food Composition and Analysis 16 (2003) 195–211.

[7] Yong He, Xiaoli Li, Xunfei Deng “Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model”, Journal of Food Engineering 79 (2007) 1238–1242.

[8] Derya Kara (Department of Chemistry, Art and Science Faculty, Balikesir University, 10100 Balikesir, Turkey) “Evaluation of trace metal concentrations in some herbs and herbal teas by principal component analysis”, Volume 114, Issue 1, 1 May 2009, Pages 347–354.

[9] Sheida Makvandi, Massoud Ghasemzadeh-Barvarz, Georges Beaudoin , Eric C. Grunsky,

M. Beth McClenaghan, Carl Duchesne “Principal component analysis of magnetite composition from volcanogenic massive sulfide deposits: Case studies from the Izok Lake (Nunavut, Canada) and Halfmile Lake (New Brunswick, Canada) deposits”, Volume 72, Part 1, January 2016, Pages 60–85.

[10] Quansheng Chen, Jiewen Zhao, Haidong Zhang, Xinyu Wang “Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration”, Volume 572, Issue 1, 14 July 2006, Pages 77–84.

[11] M.H. Zhang, J. Luypaert, J.A. Fernández Pierna, Q.S. Xu1, D.L. Massart “Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration”, Talanta 62 (2004) 25–35.

[12] H. Schulz, U. H. Engelhardt, A. Wegent, H. H. Drews, and S. Lapczynski “Application of Near-Infrared Reflectance Spectroscopy to the Simultaneous Prediction of Alkaloids and Phenolic Substances in Green Tea Leaves”, J. Agric. Food Chem. 1999, 47, 5064-5067.

[13] J. Luypaert, M.H. Zhang, D.L. Massart “Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.)”, Analytica Chimica Acta 478 (2003) 303–312.

[14] Huang Xiaowei, Zou Xiaobo, Zhao Jiewen, Shi Jiyong, Zhang Xiaolei, Mel Holmes “Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models”, Food Chemistry 164 (2014) 536–543.

[15] Guangxin Ren, Shengpeng Wang, Jingming Ning, Rongrong Xu, Yuxia Wang, Zhiqiang Xing, Xiaochun Wan, Zhengzhu Zhang “Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS)”, Volume 53, Issue 2, October 2013, Pages 822–826.

[16] Quansheng Chena, Zhiming Guo, Jiewen Zhao, Qin Ouyang “Comparisons of different regressions tools in measurement of antioxidant activity in green tea using near infrared spectroscopy”, Volume 60, 23 February 2012, Pages 92–97.

[17]https://www.google.co.in/search?q=consumption+of+tea+throughout+the+world+map+picture&biw.

Chapter 2

Theory and Measurement Modes of NIR

2.1. NIR

Near infrared energy was discovered by William Herchel in 19th century. First industrial application was possible on 1950 when NIR wore used as a add-on unit for other optical devices like UV, VIS or MIR spectrometer. In 1960, NIR was first used by USDA to detect internal quality of apple crops spoiled with a devastating condition.

Definition

NIRS is a simple, quick (>30microsecond), non-destructive technique that provide multi-constituent analysis with high accuracy and precision. It uses calibration method so it is called secondary device.

NIR spectra contains variety of chemical and physical information, it is also used to find maturity level and sugar content and provide indirect measure of taste and texture.

Material inspection is done by scanning the sample by near infrared ray and the identity and quality of material is confirmed by pattern recognition algorithm.

2.2. Theory of Vibrational Spectroscopy [1]

2.2.1. Electromagnetic Spectrum

Electromagnetic spectrum ranges from 10^{-4} to 10^{10} nano-meter. The visible part of the electromagnetic spectrum is, by definition, radiation visible to the human eye. Other detection systems reveal radiation beyond the visible regions of the spectrum and these are classified as radio wave, microwave, infrared, ultraviolet, X-ray and γ-ray[1].

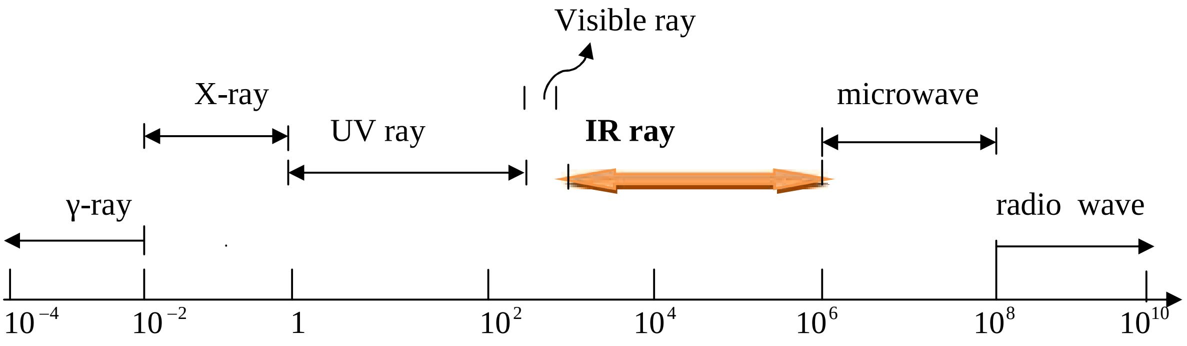


Figure 2.1. Electromagnetic Spectrum (nm) [1].

2.2.2. Energy States

Einstein, Planck and Bohr indicated that in many ways electromagnetic radiation could be regarded as a stream of particles (or quanta) for which the energy, E, is given by the Bohr equation, as follows:

$$E=h\nu=hc/\lambda \qquad \text{…………………………..2.1}$$

h (Planck constant) = 6.626 × 10–34 Js

ν is frequency of light wave

There are different levels of energy states used to define present energy state of an atom or molecule. When electromagnetic radiation exposed on the test sample, the bonds of molecules absorb the photon energy and transit to higher energy level. The transition of energy states depend upon the frequency of electromagnetic radiation.

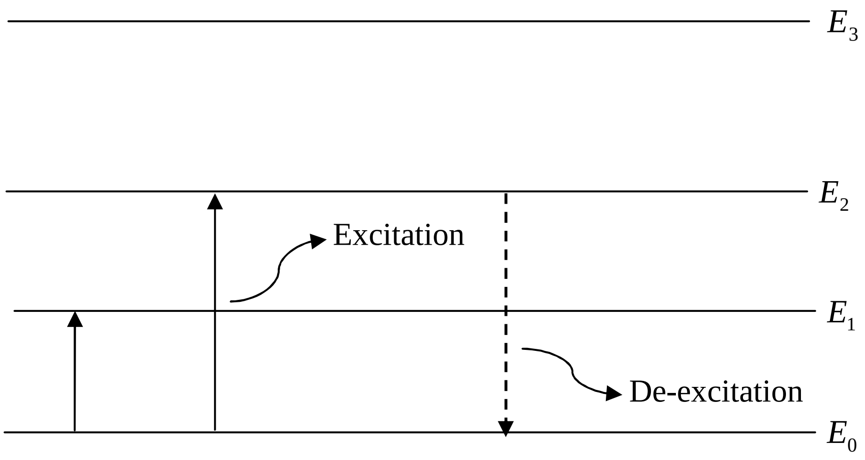


Figure 2. 2. Energy level diagram.

Electromagnetic waves ranges from gamma ray to radio wave (10^{-4} to 10^{10} nm). Since x-ray, γ-ray have small wavelength with very high frequency, they are very powerful that's why when it exposed to chemical bond, the bonds get break.

UV (190-350 nm) and VIS (350-780 nm) exposed to bonds the electrons goes to higher energy state and easily get back to ground state.

In case of NIR (780-2500 nm), molecule absorb IR ray without later re-emission by exciting certain vibration frequency.

The shown figure below represents the energy of photons at different frequencies of electromagnetic spectrum along with the various changes in molecule configuration

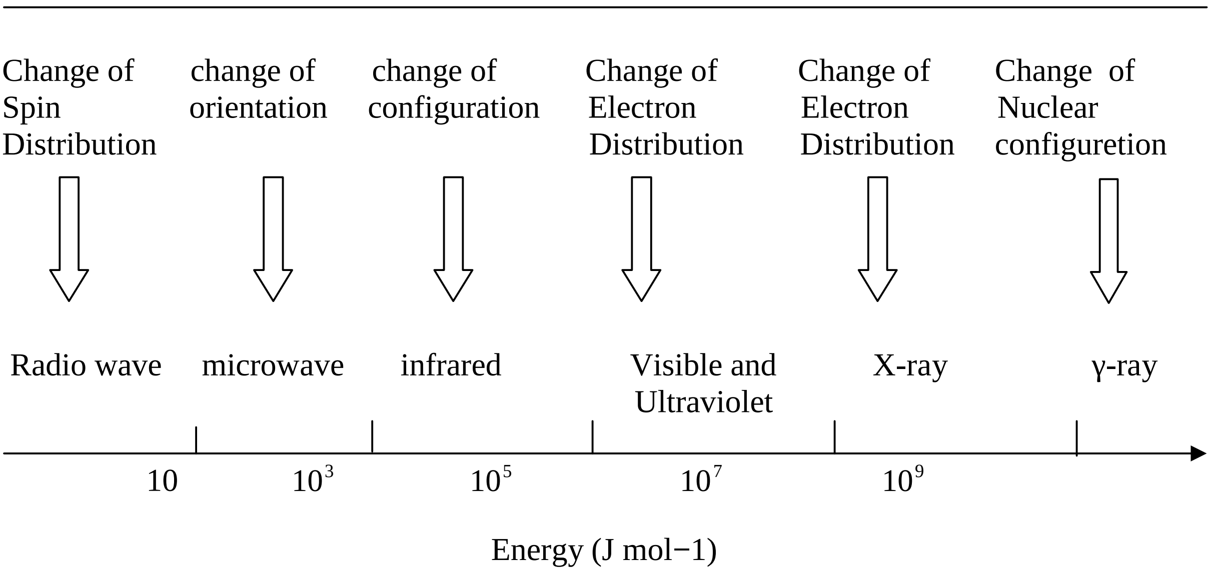


Figure 2.3. Regions of the Electromagnetic Spectrum and Photon Energy [1].

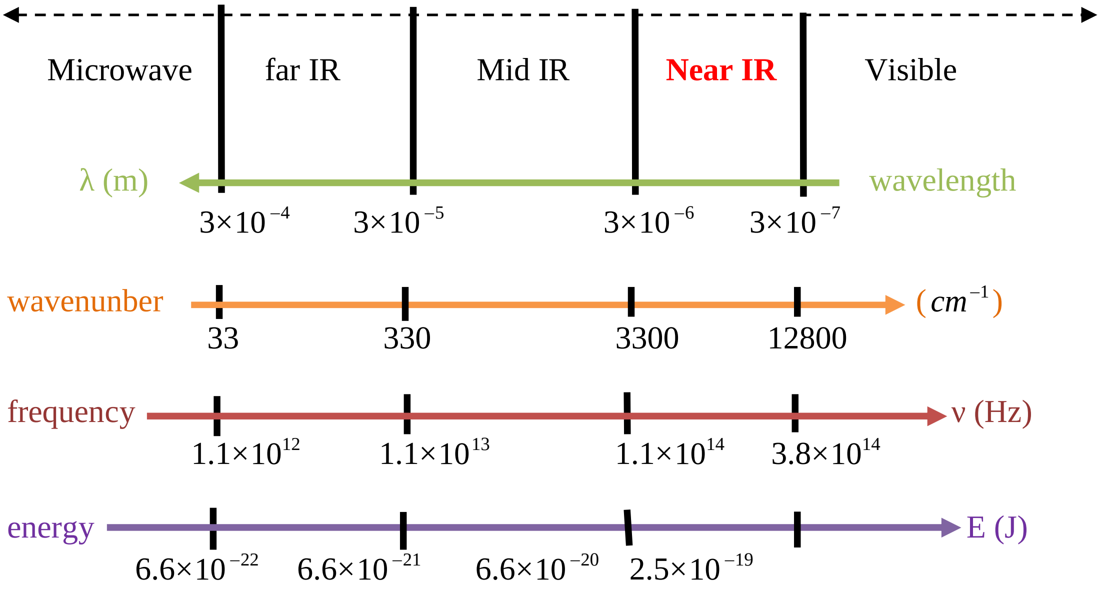


Figure 2.4. Infrared Spectrum.

2.2.3. Bond Vibration in Molecules

To understand infrared spectroscopy, we need to understand basic nature of bonds of molecule when it is exposed to external energy sources. Let us understand the diatomic molecule as a harmonic oscillator.

2.2.3.1. The Harmonic Oscillator

Quantized vibrational energy levels are specified when a molecules goes to vibrational state by interacting with external energy sources. Simplest example is vibration of a diatomic molecule. The atoms are located at an average internuclear distance r , the bond length. When atoms squeeze more closely it leads to rapid increase in repulsive forces and when atom moves apart attractive forces comes in action. Both displacements require an input of energy which can be described as a function of the distance between the two atoms.

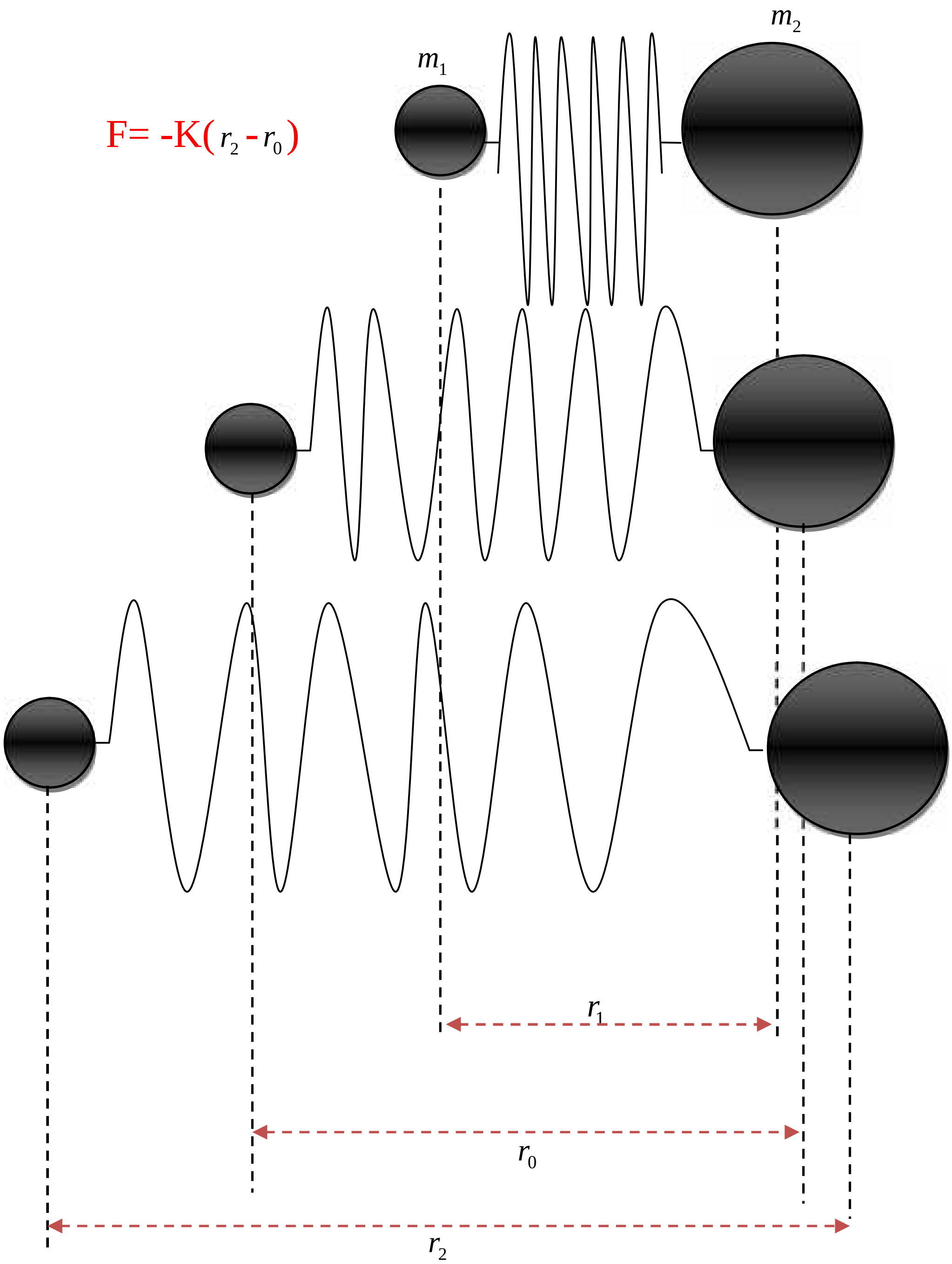


Figure 2.5. Inter-nuclear vibration.

The bond between atoms of mass m_1 and m_2 can be assumed as a spring as shown figure below, and the attraction and repulsion force can be explained by Hook's law

$$F= -K(r - r_0) \qquad \text{..............................} \ 2.2$$

K is constant, (r - r_0) is the distance difference to the equilibrium distance generated by the force(F). The force is directed against the displacement of the atoms and thus has a negative sign. The potential energy of the oscillating system

$$E = \frac{1}{2}K(r - r_0)^2 \qquad \text{..............................} \; 2.3$$

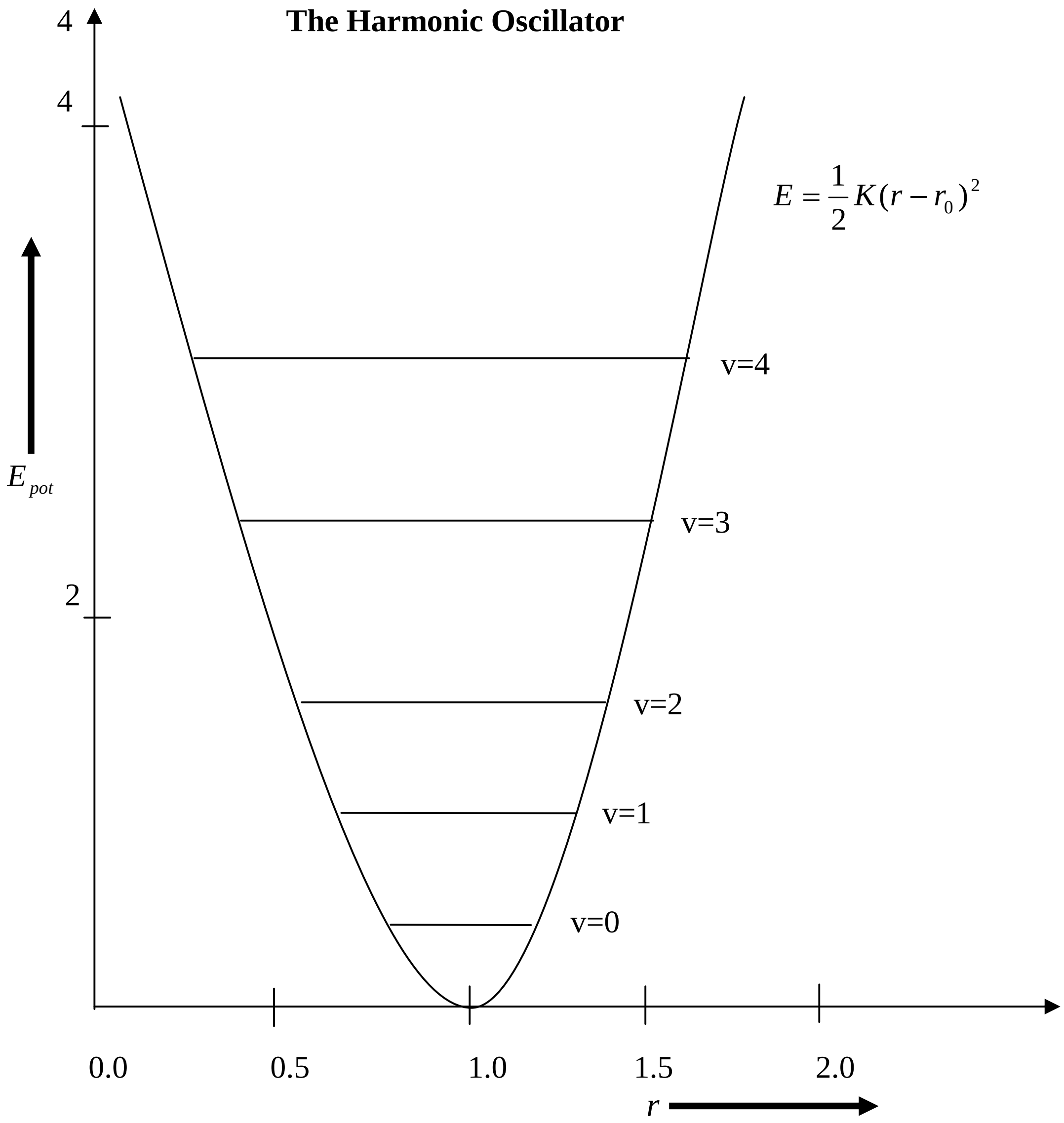


Figure 2.6. Potential energy, function of distance in a Harmonic Oscillator.

The vibration of such a diatomic molecule is characterized by an oscillation frequency that is given by classical mechanics:

$$\neg_{vib} = \frac{1}{6.28}\sqrt{\frac{K}{\text{«}}} \qquad \text{......................................} \; 2.4$$

where μ is the reduced mass of the system: $« = \frac{(m_1 m_2)}{(m_1 + m_2)}$ 2.5

Thus, the oscillation frequency is only dependent on the force constant (k) and on the masses of the oscillating atoms. For larger the mass of an atom, frequency (wave-number) of vibration is small. The amplitude of the vibration has no effect on the frequency.

According to classical mechanics, vibrational energies of molecules are quantized like other molecular energies. The allowed vibrational energies is:

$$E_v = (v + \frac{1}{2})h\neg_{vib} \qquad \text{.. 2.6}$$

$\neg$ is called the vibrational quantum number. The equation implies that the lowest vibrational energy (i.e. for v=0) $E = ({}^{1}\!/_{2})h\neg$. The implication is that a molecule can never have zero vibrational energy; the atoms can never be without a vibrational motion. The quantity $E = ({}^{1}\!/_{2})h\neg$ is known as the zero-point energy, that depends on the classical oscillation frequency and hence on the strength of the chemical bond and on the masses of atoms that participate in this bond.

Selection rules which explain the molecule undergoes vibrational changes:

- ***use of the quantum mechanics leads to a simple selection rule for the harmonic oscillator undergoing vibrational changes: Δv = ±1***

The energy of a transition between two vibrational states is then given by

$$E_{v+1} - E_v = (v + 1 + \frac{1}{2})h\neg - (v + \frac{1}{2})h\neg = h\neg \qquad \text{.................................2.7}$$

In addition to this selection rule the condition must be fulfilled that vibrational energy changes will only be observed in a spectrum if the vibration energy can interact with radiation, i.e. if the dipole moment of the molecule is changing with the vibration. Thus infrared spectra can be observed only in heteronuclear diatomic molecules, since homonuclear diatomic molecules have no dipole moment.

- ***To have resonance between the vibrating molecule and the exciting radiation, the frequency of the radiation must be identical to the frequency of the vibration.***
- ***A molecule can only absorb radiation when the incoming infrared radiation is of the same frequency as one of the fundamental modes of vibration of the molecule.***

2.2.3.2. Anharmonic Oscillator

Real bonds are elastic, but do not obey Hook's law. Bonds can break if they are stretch beyond a limiting distance and the molecule will dissociate. For amplitudes that exceed a certain value, the vibration must be described differently. A purely empiric description of such a behaviour was derived by P.M. Morse and is called the Morse function.

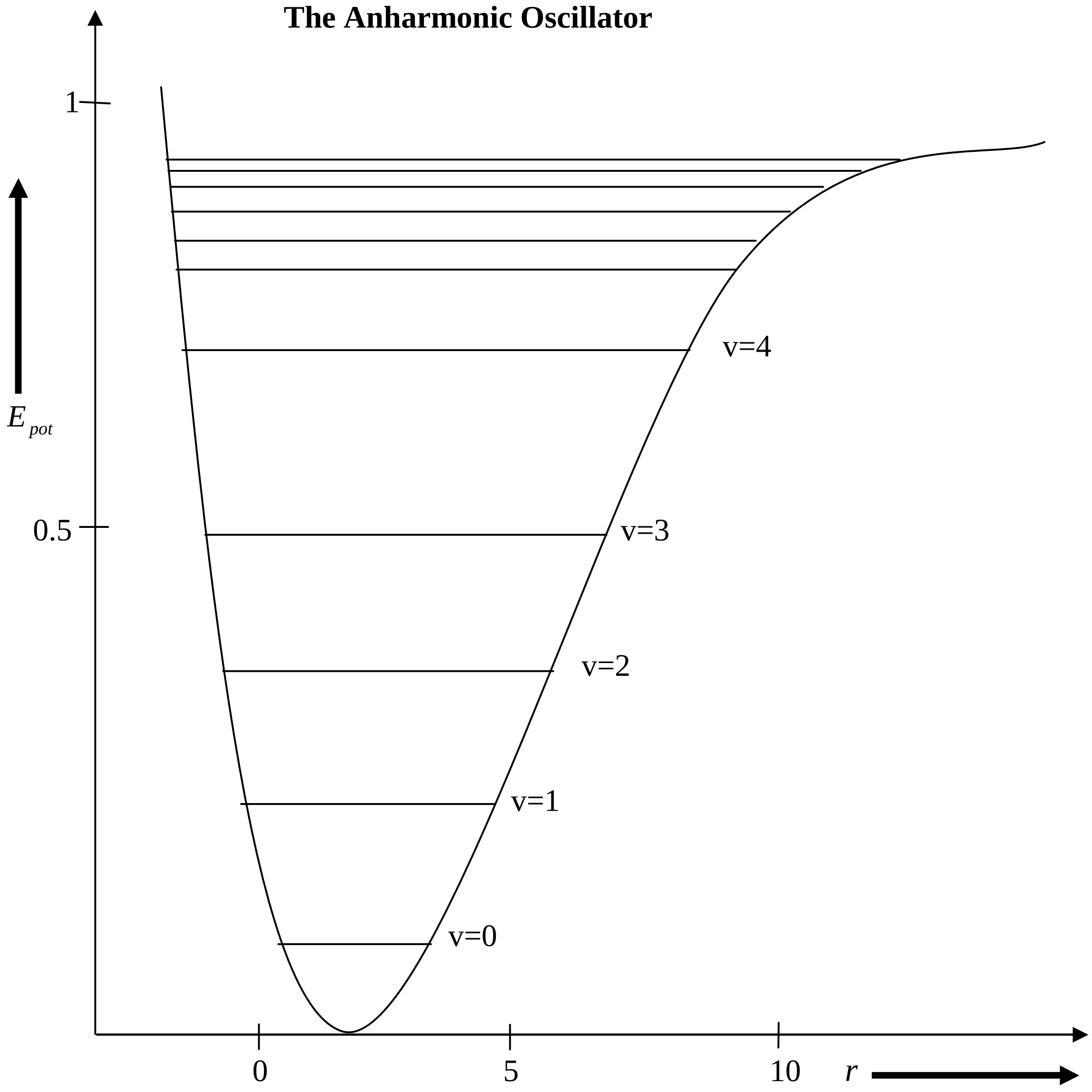


Figure 2.7. Anharmonic Oscillator.

The potential energy of the an-harmonic oscillator is described by a Morse function (Morse potential). If the Morse function is used instead of the energy of the harmonic oscillator to derive the discrete energy levels, that are given by

$$E_{v+1} - E_v = (v+1+\frac{1}{2})h\neg - (v+\frac{1}{2})h\neg = h\neg x \qquad \text{......................... 2.8}$$

where x is an an-harmonicity constant, that is small for bonding vibrations and always positive. The selection rules for the an-harmonic oscillator are

$$\Delta\, v = \pm 1, \pm\, 2, \pm\, 3, ...$$

2.2.3.3. Vibrational Degrees of Freedom

When molecule absorbs infrared energy it goes to vibration state and the fashion it can vibrate is depend upon the degree of vibration of particular bond. A molecule with N atoms can be described by the locations of each atom in space by three coordinates, x, y, z. The total number of such coordinates is therefore 3 N. The molecule has 3N degrees of freedom. In such a description, the position of the molecule and the bond-angles are fixed. The translation of the molecule in space is described by three degrees of translational freedom. For the rotational motion of a nonlinear molecule we need additional three degrees of freedom, for a linear molecule just two (the rotation around the bond axis of a linear molecule does not result in a change of the coordinates of the atoms). Therefore, a non-linear must have degrees of freedom of internal vibration.

3N-3-3 = 3N-6 (non-linear molecule) ***3N-3-2 = 3N-5 (linear molecule)***

CO_2 is a linear molecule so it has 4 degree of vibration whereas H_2O is non-linear so it has 3 degree of freedom. Some bonds can stretch in-phase (symmetrical stretching) or out-of-phase (asymmetric stretching), as shown in Figure below [1].

Symmetric stretching

Asymmetric stretching

Figure 2.8. Symmetric and asymmetric stretching vibrations [1].

For a vibration to give rise to the absorption of infrared radiation, it must cause a change in the dipole moment of the molecule. The larger this change, then the more intense will be the absorption band. For example CO_2

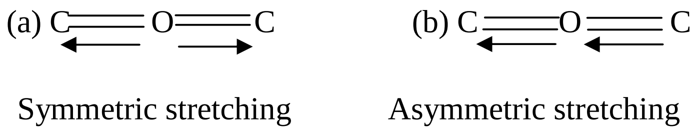


Figure 2.9. Vibration of carbon dioxide.

A dipole moment is a vector sum. If the two C=O bonds of CO_2 are stretched symmetrically, there is still no net dipole and so there is no infrared activity. However, in the asymmetric stretch, the two C=O bonds are of different length and, hence, the molecule has a dipole. Therefore, the vibration shown in Figure (b) is ‘infrared-active’.

Symmetrical molecules will have fewer ‘infrared-active’ vibrations than asymmetrical molecules. This leads to the conclusion that symmetric vibrations will generally be weaker than asymmetric vibrations, since the former will not lead to a change in dipole moment. It follows that the bending or stretching of bonds involving atoms in widely separated groups of the periodic table will lead to intense bands. Vibrations of bonds such as C−C or N=N will give weak bands. This again is because of the small change in dipole moment associated with their vibrations. Varieties of take place in a molecule due to infrared absorption the, these are listed below.

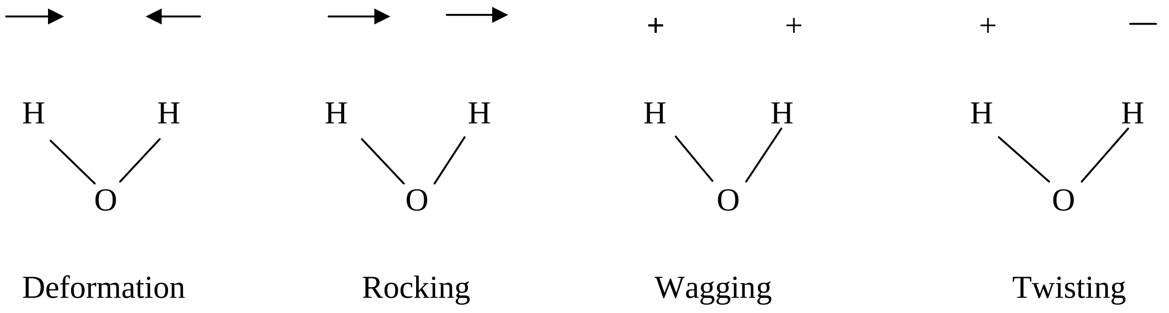


Figure 2.10. Different types of bending vibrations.

2.2.4. Overtones and Recombination

When radiation is exposed to a matter, molecular bond behaves as a anharmonic oscillator and the separation between two vibrational energy level is not equal as found in harmonic oscillator and vibrational level get closer and closer. Due to irregular distribution of energy levels the transition of molecule is not exactly ± 1 (i.e. Δv≠ ±1), these kind of situation arise overtones in molecular transition.

Combination bands arise when two fundamental bands absorbing at v1 and v2 absorb energy simultaneously. The resulting band will appear at (v1 + v2) wavenumbers. These two phenomena are found in NIR region which makes very useful NIR spectroscopy for quantitative analysis of a sample.

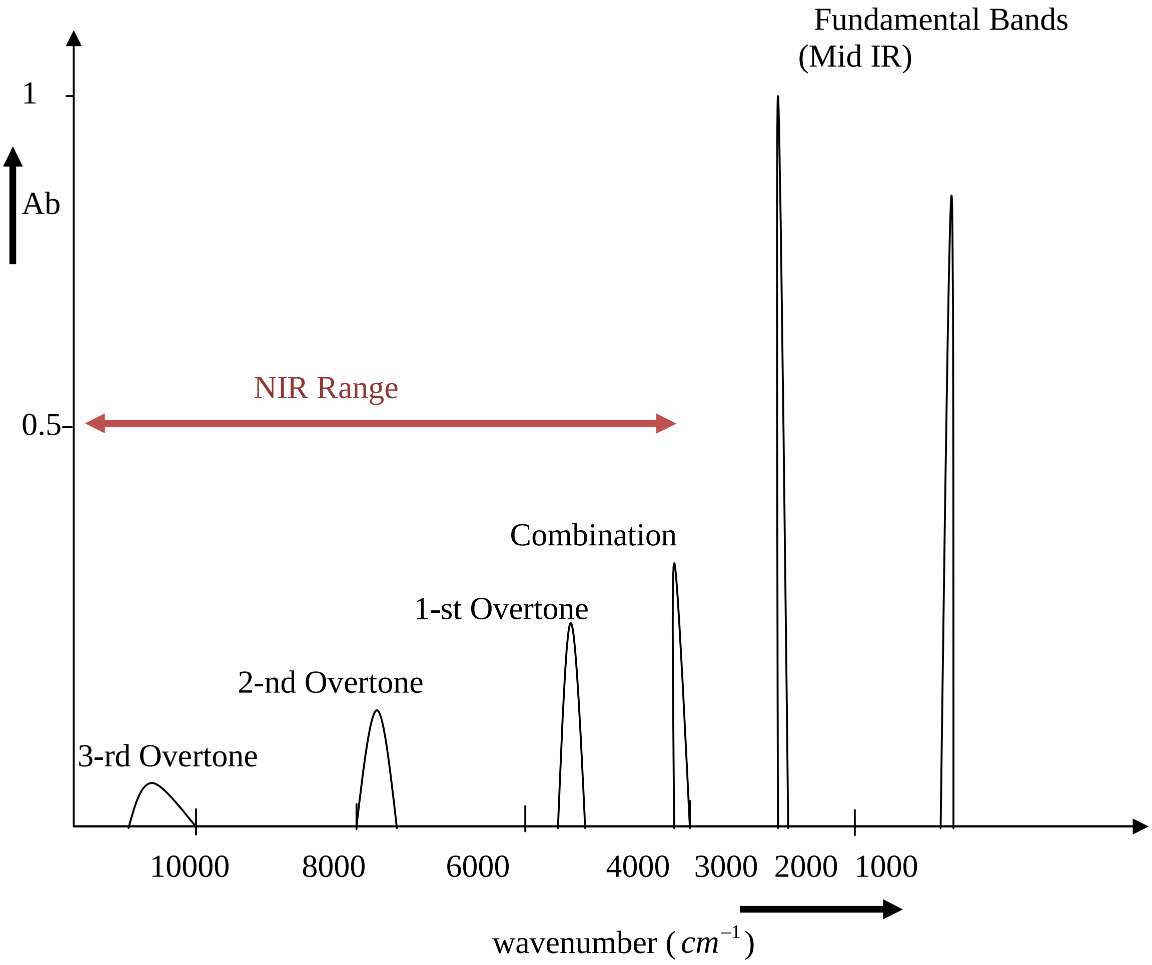


Figure 2.**10. Overtone and Combination in NIR Region [3].**

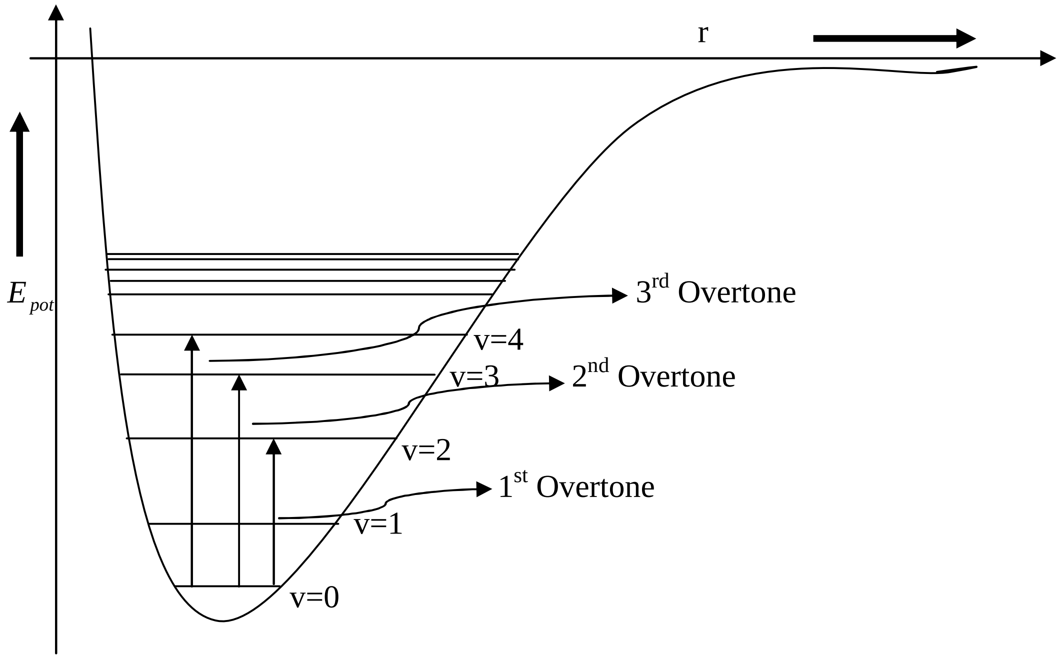


Figure 2.11. Overtone in Vibrational energy level

2.3. NIR Spectral Region

The NIR region of the electromagnetic spectrum extends from the end of the visible spectral region (700 nm or 14285 cm-1) to the beginning of the fundamental infrared (IR) spectral region (2500 nm or 4000 cm-1). The most prominent absorption bands occurring in the NIR region are related to the overtone and combination bands of the fundamental molecular vibrations of C–H, N–H, O–H, and S–H functional groups observed in the mid-IR spectral region (Figure 1, Table 2). Thus, most chemical and biochemical species exhibit unique absorption bands in the NIR spectral region that can be used for both qualitative and quantitative purposes.[4]

Spectroscopic techniques excel by their possibility to gain rapid and accurate information from the high-resolution spectra of solid and liquid samples without sample preparation.

All three spectroscopic techniques are economic and facilitate qualitative and quantitative as well as noninvasive and nondestructive analysis. Further advantages include that they are reagent- and waste-free and require no additional auxiliary chemicals. For all these reasons, spectroscopic techniques are ideally suited for industrial quality control and process monitoring. For nearly every application, there is the right technique.

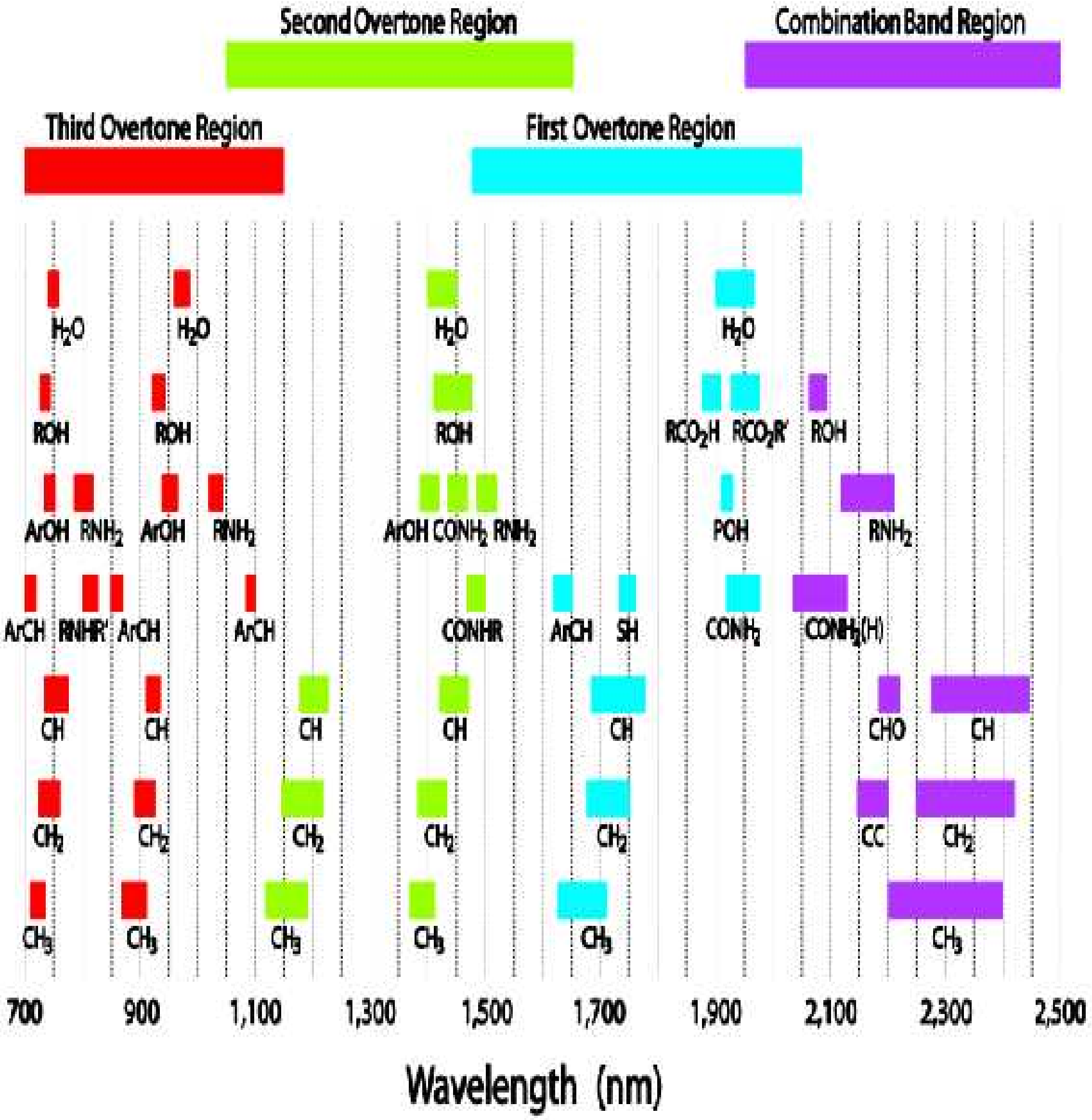


Figure 2.12. Molecular bond absorption in NIR range [4].

2.4. Measurement Methods in NIR Spectroscopy

In this section, how samples can be introduced into the instrument, the equipment required to obtain spectra and the pre-treatment of samples are examined. First, the various ways of investigating samples using the traditional transmission methods of infrared spectroscopy will be discussed. Reflectance methods, such as the attenuated total reflectance, diffuse reflectance and specular reflectance approaches, as well as photoacoustic spectroscopy, are also explained.

2.4.1. Dispersive Infrared Spectrometer

The dispersive element in dispersive instruments is contained within a monochromator. Figure below shows the optical path of an infrared spectrometer which uses a grating monochromator. Dispersion occurs when energy falling on the entrance slit is collimated onto the dispersive element and the dispersed radiation is then reflected back to the exit slit, beyond which lies the detector. The dispersed spectrum is scanned across the exit slit by rotating a suitable component within the monochromator. The widths of the entrance and exit slits may be varied and programmed to compensate for any variation of the source energy with wavenumber. In the absence of a sample, the detector then receives radiation of approximately constant energy as the spectrum is scanned.

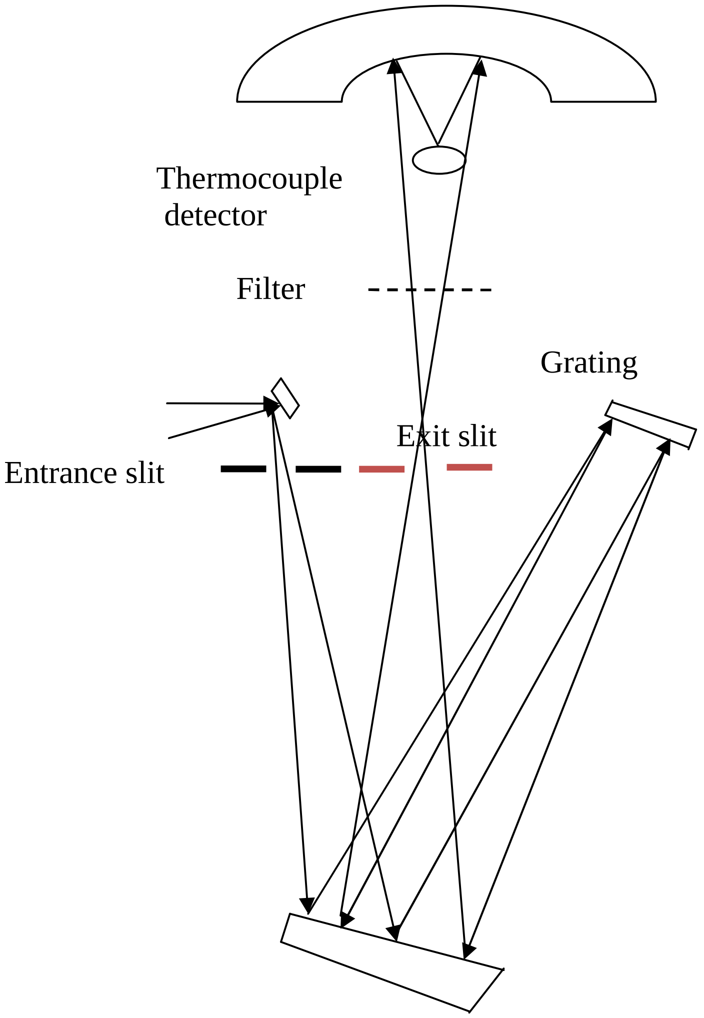


Figure 2.13. Schematic of the optical path of a double-beam infrared spectrometer [1].

Atmospheric absorption by CO2 and H2O in the instrument beam has to be considered in the design of infrared instruments. These contributions can be taken into account by using a double-beam arrangement in which radiation from a source is divided into two beams. These beams pass through a sample and a reference path of the sample compartment, respectively. The information from these beams is rationed to obtain the required sample spectrum.

The essential problem of the dispersive spectrometer lies with its monochromator. This contains narrow slits at the entrance and exit which limit the wavenumber range of the radiation reaching the detector to one resolution width. Samples for which a very quick measurement is needed, for example, in the eluant from a chromatography column, cannot be studied with instruments of low sensitivity because they cannot scan at speed. However, these limitations may be overcome through the use of a Fourier-transform infrared spectrometer.

2.4.2. Fourier-Transform Infrared Spectrometer

Fourier-transform infrared (FTIR) spectroscopy is based on the idea of the interference of radiation between two beams to yield an interferogram. The latter is a signal produced as a function of the change of pathlength between the two beams. The two domains of distance and frequency are interconvertible by the mathematical method of Fourier-transformation.

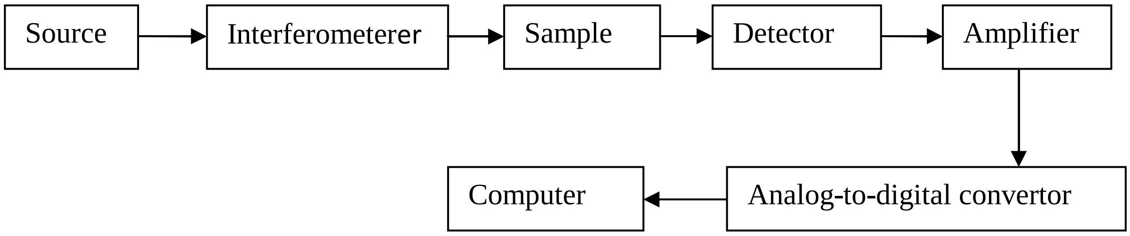


Figure 2.14. Block diagram of basic components of an FTIR spectrometer.

The radiation emerging from the source is passed through an interferometer to the sample before reaching a detector. Upon amplification of the signal, in which high-frequency contributions have been eliminated by a filter, the data are converted to digital form by an analog-to-digital converter and transferred to the computer for Fourier-transformation.

2.4.2.1. Michelson Interferometer

The most common interferometer used in FTIR spectrometry is a Michelsonv interferometer, which consists of two perpendicularly plane mirrors, one of which can travel in a direction perpendicular to the plane. A semi-reflecting film, the beamsplitter, bisects the planes of these two mirrors. The beamsplitter material has to be chosen according to the region to be examined. Materials such as germanium or iron oxide are coated onto an ‘infrared-transparent substrate such as potassium bromide or caesium iodide to produce beamsplitters for the mid- or near-infrared regions. Thin organic films, such as poly(ethylene terephthalate), are used in the far-infrared region.

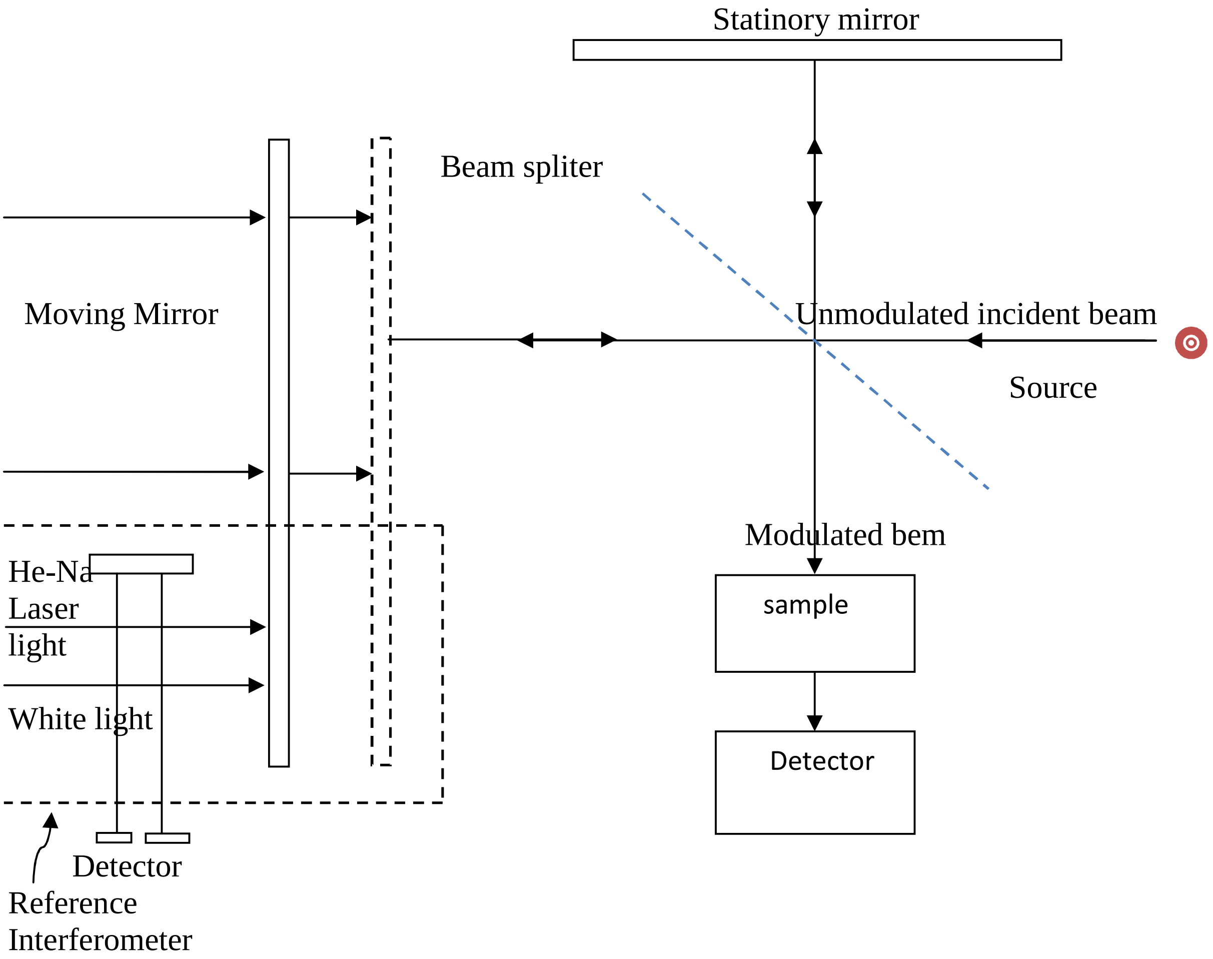
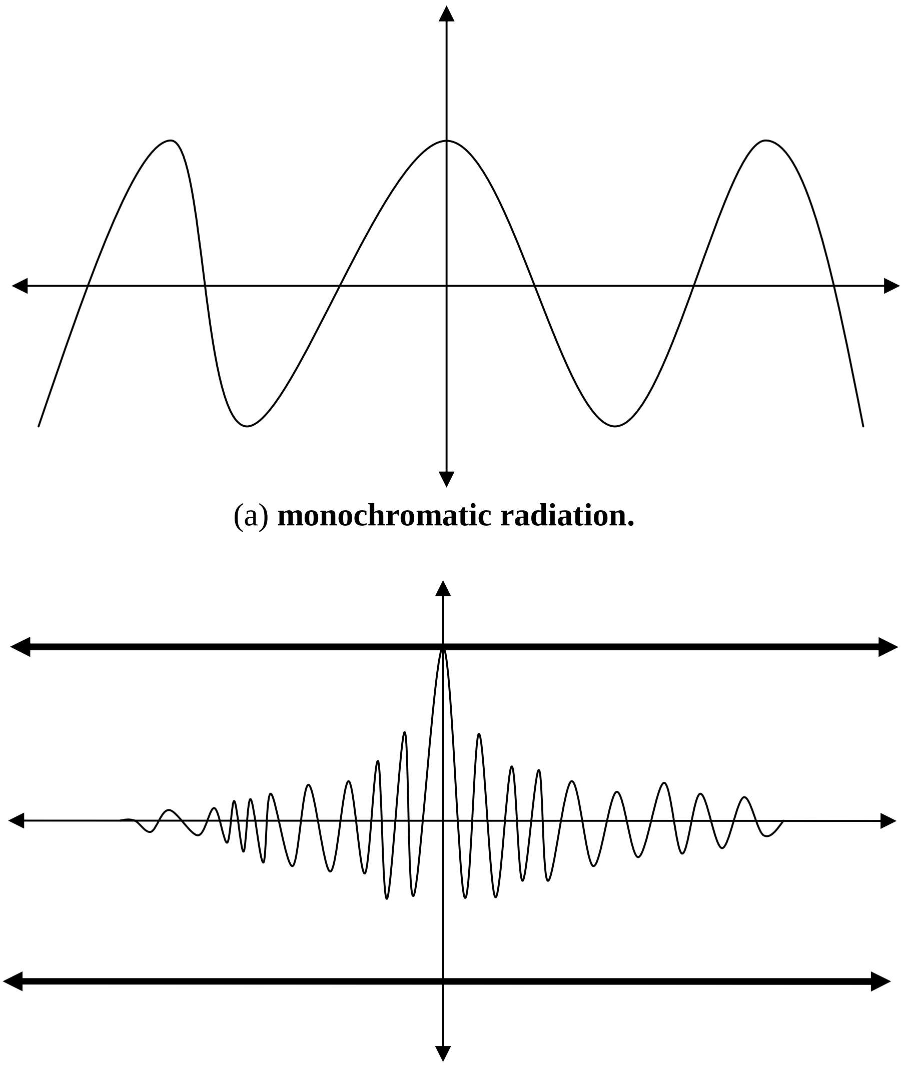


Figure 2.15. Schematic of a Michelson Interferometer [6].

If a collimated beam of monochromatic radiation of wavelength λ (cm) is passed into an ideal beam-splitter, 50% of the incident radiation will be reflected to one of the mirrors while 50% will be transmitted to the other mirror. The two beams are reflected from these mirrors, returning to the beam-splitter where they recombine and interfere. Fifty percent of the beam reflected from the fixed mirror is transmitted through the beamsplitter while 50% is reflected back in the direction of the source. The beam which emerges from the interferometer at 90◦ to the input beam is called the transmitted beam and this is the beam detected in FTIR spectrometry.

The moving mirror produces an optical path difference between the two arms of the interferometer. For path differences of (n + 1/2) λ, the two beams interfere destructively in the case of the transmitted beam and constructively in the case of the reflected beam. The resultant interference pattern is shown in



(a) **monochromatic radiation.**

(b) polychromatic radiation

Figure 2.16. Interferograms obtained for (a) monochromatic radiation and (b) polychromatic radiation [7].

Figure above for (a) a source of monochromatic radiation and (b) a source of polychromatic radiation (b). The former is a simple cosine function, but the latter is of a more complicated form because it contains all of the spectral information of the radiation falling on the detector.

2.4.2.2. Source and Detector

FTIR spectrometers use a Globar or Nernst source for the mid-infrared region. If the far-infrared region is to be examined, then a high-pressure mercury lamp can be used. For the near-infrared, tungsten−halogen lamps are used as sources. There are two commonly used detectors employed for the mid-infrared region. The normal detector for routine use is a pyroelectric device incorporating deuterium tryglycine sulfate (DTGS) in a temperature-resistant alkali halide window.

For more sensitive work, mercury cadmium telluride (MCT) can be used, but this has to be cooled to liquid nitrogen temperatures. In the far-infrared region, germanium or indium−antimony detectors are employed, operating at liquid helium temperatures. For the near-infrared region, the detectors used are generally lead sulfide photoconductors.

2.4.2.3. Fourier Transform

The essential equations for a Fourier-transformation relating the intensity falling on the detector, I(δ), to the spectral power density at a particular wavenumber, ν, given by B(ν), are as follows:

$$I(\delta) = \int_{0}^{-\infty} B(\nu)\cos(2\pi\nu\delta)d\,\nu \qquad \text{.......................................} \; 2.9$$

which is one half of a cosine Fourier-transform pair, with the other being:

$$\mathrm{B}(\nu) = \int_{0}^{-\infty} \mathrm{I}(\delta)\cos(2\pi\nu\delta)\mathrm{d}\,\delta \qquad \text{......................................} \; 2.10$$

These two equations are inter-convertible and are known as a Fourier-transform pair. The first shows the variation in power density as a function of the difference in pathlength, which is an interference pattern. The second shows the variation in intensity as a function of wavenumber. Each can be converted into the other by the mathematical method of Fourier-transformation. The essential experiment to obtain an FTIR spectrum is to produce an interferogram with and without a sample in the beam and transforming the interferograms into spectra of (a) the source with sample absorptions and (b) the source without sample absorptions. The ratio of the former and the latter corresponds to a double-beam dispersive spectrum.

The major advance toward routine use in the mid-infrared region came with a new mathematical method (or algorithm) devised for fast Fourier-transformation (FFT). This was combined with advances in computers which enabled these calculations to be carried out rapidly.

2.4.2.4. Moving Mirror

The moving mirror is a crucial component of the interferometer. It has to be accurately aligned and must be capable of scanning two distances so that the path difference corresponds to a known value.

A number of factors associated with the moving mirror need to be considered when evaluating an infrared spectrum. The interferogram is an analogue signal at the detector that has to be digitized in order that the Fourier-transformation into a conventional spectrum can be carried out. There are two particular sources of error in transforming the digitized information on the interferogram into a spectrum. First, the transformation carried out in practice involves an integration stage over a finite displacement rather than over an infinite displacement. The mathematical process of Fourier transformation assumes infinite boundaries.

The consequence of this necessary approximation is that the apparent line shape of a spectral line may be as shown in Figure below, where the main band area has a series of negative and positive side lobes (or pods) with diminishing amplitudes.

The process of apodization is the removal of the side lobes (or pods) by multiplying the interferogram by a suitable function before the Fourier-transformation is carried out. A suitable function must cause the intensity of the interferogram to fall smoothly to zero at its ends. Most FTIR spectrometers offer a choice of apodization options and a good general purpose apodization function is the cosine function, as follows:

$$F(D) = [1 + \cos(\pi D)]/2 \qquad \text{..................................} \; 2.11$$

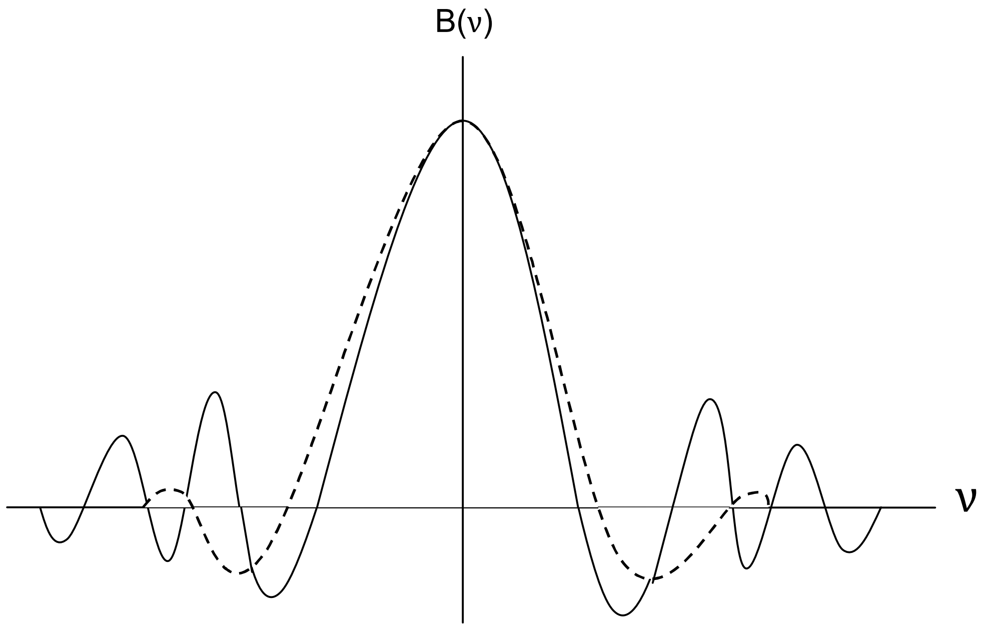


Figure 2.17. Instrument line shape without apodization[8].

D is the optical path difference. This cosine function provides a good compromise between reduction in oscillations and deterioration in spectral resolution. When accurate band shapes are required, more sophisticated mathematical functions may be needed. Another source of error arises if the sample intervals are not exactly the same on each side of the maxima corresponding to zero path differences. Phase correction is required and this correction procedure ensures that the sample intervals are the same on each side of the first interval and should correspond to a path difference of zero. The resolution for an FTIR instrument is limited by the maximum path difference between the two beams. The limiting resolution in wavenumbers (cm−1) is the reciprocal of the pathlength difference (cm). For example, a pathlength difference of 10 cm is required to achieve a limiting resolution of 0.1 cm−1. This simple calculation appears to show that it is easy to achieve high resolution. Unfortunately, this is not the case since the precision of the optics and mirror movement mechanism become more difficult to achieve at longer displacements of pathlengths.

2.4.2.5. Signal Averaging

The main advantage of rapid-scanning instruments is the ability to increase the signal-to-noise ratio (SNR) by signal-averaging, leading to an increase of signal-to-noise proportional to the square root of the time, as follows:

SNR α n1/2 .... 2.12

2.4.2.6. Computer

The computer forms a crucial component of modern infrared instruments and performs a number of functions. The computer controls the instrument, for example, it sets scan speeds and scanning limits, and starts and stops scanning. It reads spectra into the computer memory from the instrument as the spectrum is scanned; this means that the spectrum is digitized. Spectra may be manipulated using the computer, for example, by adding and subtracting spectra or expanding areas of the spectrum of interest. The computer is also used to scan the spectra continuously and average or add the result in the computer memory. Complex analyses may be automatically carried out by following a set of pre-programmed commands . The computer is also used to plot the spectra.

2.4.2.7. Spectra

Early infrared instruments recorded percentage transmittance over a linear wavelength range. It is now unusual to use wavelength for routine samples and the wavenumber scale is commonly used. The output from the instrument is referred to as a spectrum. Most commercial instruments present a spectrum with the wavenumber decreasing from left to right. The infrared spectrum can be divided into three main regions: the farinfrared (<400 cm−1), the mid-infrared (4000−400 cm−1) and the near-infrared (13 000−4000 cm−1). These regions will be described later in more detail in Chapter 3. Many infrared applications employ the mid-infrared region, but the near- and far-infrared regions also provide important information about certain materials. Generally, there are less infrared bands in the 4000−1800 cm−1 region with many bands between 1800 and 400 cm−1. Sometimes, the scale is changed so that the region between 4000 and 1800 cm−1 is contracted and the region between 1800 and 400 cm−1 is expanded to emphasize features of interest.

2.5. Transmittance Method

Transmission spectroscopy is the oldest and most straightforward infrared method. This technique is based upon the absorption of infrared radiation at specific wavelengths as it passes through a sample. It is possible to analyse samples in the liquid, solid or gaseous forms when using this approach.

2.5.1. Liquid and Solutions

There are several different types of transmission solution cells available. Fixed path-length sealed cells are useful for volatile liquids, but cannot be taken apart for cleaning. Semi-permanent cells are demountable so that the windows can be cleaned. A semi-permanent cell is illustrated in Figure below. The spacer is usually made of polytetrafluoroethylene (PTFE, known as 'Teflon') and is available in a variety of thicknesses, hence allowing one cell to be used for various path-lengths. Variable path-length cells incorporate a mechanism for continuously adjusting the path length, while a vernier scale allows accurate adjustment. All of these cell types are filled by using a syringe and the syringe ports are sealed with PTFE plugs before sampling.

Before producing an infrared sample in solution, a suitable solvent must be chosen. In selecting a solvent for a sample, the following factors need to be considered: it has to dissolve the compound, it should be as non-polar as possible to minimize solute−solvent interactions, and it should not strongly absorb infrared radiation. If quantitative analysis of a sample is required, it is necessary to use a cell of known path length.

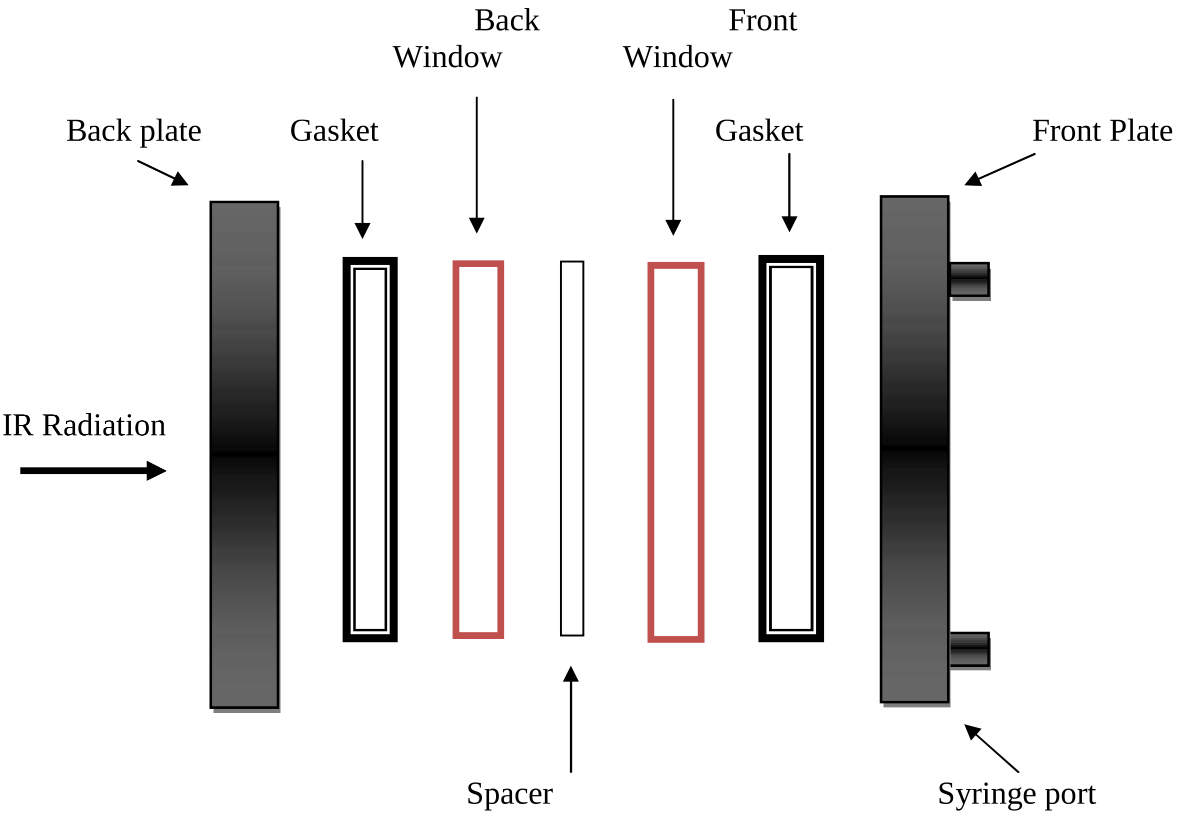


Figure 2.18. Schematic of a typical semi-permanent liquid cell [1].

2.5.2. Solid

There are three general methods used for examining solid samples in transmission infrared spectroscopy; i.e. alkali halide discs, mulls and films. The choice of method depends very much on the nature of the sample to be examined. The use of alkali halide discs involves mixing a solid sample with a dry alkali halide powder. The mixture is usually ground with an agate mortar and pestle and subjected to a pressure of about 10 ton in−2 (1.575 × 105 kgm−2) in an evacuated die. This sinters the mixture and produces a clear transparent disc. The most commonly used alkali halide is potassium bromide (KBr), which is completely transparent in the mid-infrared region.

2.5.3. Gases

Gases have densities which are several orders of magnitude less than liquids, and hence pathlengths must be correspondingly greater, usually 10 cm or longer. The walls are of glass or brass, with the usual choice of windows. The cells can be filled by flushing or from a gas

line. To analyse complex mixtures and trace impurities, longer path lengths are necessary. As the sample compartment size in the instrument is limited, a multi-reflection gas cell is necessary to produce higher path lengths. In such a cell, the infrared beam is deflected by a series of mirrors which reflect the beam back and forth many times until it exits the cell after having travelled the required.

2.6. Comparison among NIR MIR and Raman Spectroscopy

The most flexible technique, however, is NIRS. Especially the availability of efficient chemometric evaluation tools and software as well as light-fibre optics has made NIRS to an invaluable tool for academic research and industrial quality control. NIRS allows the determination of multiple values in a single determination. In the process environment, NIRS stands out by the possibility of on-site measurements and remote sampling – not least because the spectrometer can be placed far away from the sampling point [4].

Table 2.1. Comparison among NIR MIR and Raman Spectroscopy.

Character	Raman	Mid IR	Near IR
wavenumber	50-4000 cm^{-1}	200-4000 cm^{-1}	4000-12500 cm^{-1}
Bonds	Homonuclear bonds such as C-C,C=C,S=S	polar bonds such as C=O, C–O, C–F	H-containing bonds such as C–H, O–H, N–H, S–H
Absorption bands due to	scattered radiation	absorbed radiation (basic vibration)	absorbed radiation (overtones and combination)
Absorption	strong	weak	weak
Absorption band	well-resolved, assignable to specific chemical groups	well-resolved, assignable to specific chemical groups	series of overlapping bands
Signal intensity	poor	good	good
Qualification	intensity (I) ~ concentration	log I0 /I ~ concentration (Lambert-Beer law)	log I0/I ~ concentration (Lambert-Beer law)
Excitation condition	change of polarizability α	change of dipole moment μ	change of dipole moment μ
Selectivity	high	high	low, requires calibration and chemometrics
Interference	broad fluorescence baseline	water	Water, physical attributes (e.g., sample size, shape, and hardness)
Particle size	independent	dependent	dependent
Applicabilty for atline inline and online	good	poor	good
Radiation source	Monochromatic (laser VIS/NIR region)	Polychromatic by globar tungsten	Polychromatic by globar tungsten
Sample	none	reduced (except ATR*)	none

2.7. Conclusion

Earlier IR or UV/VIS spectroscopy was used to estimate chemical composition and quality of material because of the small spectrum range of NIR region. The spectrum and peaks were not clear and very complicated. Due to the developments of sensitive and advance detector

NIR spectroscopy comes in senerio. The benefits of NIR become more apparent and now peaks are found in this region become more useful.

The key benefits of a spectrometer include the ability to quickly and accurately measure samples while keeping the main analyzer away from any potentially dangerous processes. Fiber cables from the probes to the analyzer make this possible, but also introduce potential error sources for light measurement. The improved transmission causes the NIR peaks to be more accurate and thus more accurately analyzed.

Since NIR region shows overtones and recombination during the molecular vibration and these phenomena are very useful to for quantitative analysis purpose.

More advanced mathematics analysis techniques must be employed for NIR spectral data than for IR analysis. Multiple linear regressions (MLRs) and partial least squared regressions (PLRs) are used to analyze spectral peaks and glean useful data.

References

[1] Barbara Staurt, Willey Publication, “Infrared Spectroscopy: Fundamental and application”, introduction, pp3-5.

[2] E.R. Malinowski, Factor analysis in chemistry, 2nd. Ed., John Wiley, New York, 1991.

[3] Vysoka Skola, chemicko-technologicka V Prazw: NIR spectroscopy.

[4]Metrohm; A guide to near-infrared spectroscopic analysis of industrial manufacturing processes

[5] J. Braz. Chem. Soc. Saw Paulo; “Near Infrared Spectroscopy: Fundamentals, Practical aspects and Analytical Application” . Vol 14 no. 2, Mar./Apri. 2003.

[6] Stuart, B., “Modern Infrared Spectroscopy”, ACOL Series, Wiley, Chichester, UK, 1996. University of Greenwich, and reproduced by permission of the University of Greenwich.

[7] Barnes, A. J. and Orville-Thomas, W. J. (Eds), Vibrational Spectroscopy – Modern Trends, Elsevier, Amsterdam, Figure 2, p. 55 (1977).

[8]Barnes, A. J. and Orville-Thomas, W. J. (Eds), Vibrational Spectroscopy – Moder Trends,Elsevier, Amsterdam, Figure 3, p. 55 (1977).

Chapter 3

Multivariate Analysis Technique

3.1. Principal Component Analysis

Once NIR data (i.e. the graph of absorbance versus frequency) is obtained, it is needed to analyse these data to full fill necessary requirement and needs. For qualitative detection Mid IR region is useful whereas for quantitative detection NIR region appropriate.

Some multivariate analysis technique like PCR and PLS is very use full to analyse the quantity of chemical compound.

Principal component analysis is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of large dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data.

While analysing a experimental data basically we are calculating values of number of variables of different samples, and plotting graph of huge number of variables are very complicated unless we do not have a multi-dimensional coordinate system.

PCA reduces the number of variable with the help of statistical process and construct few principle components which contain most of the variance of given variables without loss of much information. PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.

3.1.1. Statistics behind PCA

Principal component is calculated by the use of statistical tools. Let us sea the mathematical and statistical used in PCA. The entire subject of statistics is based around the idea that you have this big set of data, and you want to analyse that set in terms of the relationships between the individual points in that data set.

3.1.1.1. Standard Deviation

The Standard Deviation (SD) of a data set is a measure of how spread out the data is. “The average distance from the mean of the data set to a point”. The way to calculate it is to compute the squares of the distance from each data point to the mean of the set, add them all up, divide by n- 1, and take the positive square root. As a formula:

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)}} \qquad \text{.....................................} \ 3.1$$

To understand standard deviation, we need a data set. There are a number of things that we can calculate about a data set. For example, we can calculate the mean of the sample. I assume that the reader understands what the mean of a sample is, and will only give the formula. There are a number of things that need to calculate about a data set. Mean of a sample is:

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \qquad \text{.......................................} \ 3.2$$

$\bar{x}$ indicates mean of x. (said “X bar”) to indicate the mean of the set . All this formula says is “Add up all the numbers and then divide by how many there are”. Unfortunately, the mean doesn’t tell us a lot about the data except for a sort of middle point. For example, these two data sets have exactly the same mean (10), but are obviously quite different:

[0 8 12 20] and [8 9 11 12]

The mean of above two set is same but the spread of data is different and it can be calculated by the use of standard deviation formulae.

3.1.1.2. Variance

Variance is another measure of the spread of data in a data set. In fact it is almost identical to the standard deviation. The formula is this:

$$\pm^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{(n-1)}$$ 3.3

$\pm^2$ is the usual symbol for variance of a sample. Both these measurements are measures of the spread of the data. Standard deviation is the most common measure, but variance is also used.

3.1.1.3. Covariance

Standard deviation and variance only operate on 1 dimension, so that you could only calculate the standard deviation for each dimension of the data set independently of the other dimensions. However, it is useful to have a similar measure to find out how much the dimensions vary from the mean with respect to each other.

Covariance is such a measure. Covariance is always measured between 2 dimensions. If you calculate the covariance between one dimension and itself, you get the variance. So, if you had a 3-dimensional data set (x, y , z), then you could measure the covariance between the x and y dimensions, the y and z dimensions, and the z and x dimensions. Measuring the covariance between x and x, or y and y, or z and z would give you the variance of the x, y and z dimensions respectively.

The formula for covariance is very similar to the formula for variance. The formula for variance could also be written like this:

$$\mathrm{VAR(X)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n}$$ 3.4

$$\mathrm{COV(X,Y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$ 3.5

3.1.1.4. Covariance Matrix

Covariance is always measured between 2 dimensions. If we have a data setwith more than 2 dimensions, there is more than one covariance measurement that can be calculated. For example, from a 3 dimensional data set (dimensions X, Y, Z) we can calculate COV(X,Y), COV(Y,Z) and COV(Z,X) In fact, for an n -dimensional data set, you can calculate $\frac{n!}{(n-2)!*2}$ different covariance values.

An example of covariance matrix for an imaginary 3 dimensional data set, using the usual dimensions X,Y and Z . Then, the covariance matrix has 3 rows and 3 columns, and the values are this:

$$C = \begin{bmatrix} C & (X,X) & C & (X,Y) & C & (X,Z) \\ C & (Y,X) & C & (Y,Y) & C & (Z,Y) \\ C & (Z,X) & C & (Z,Y) & C & (Z,Z) \end{bmatrix} \qquad \text{..................................} \ 3.6$$

3.1.1.5. Matrix Algebra

(a) Eigen vector: If a square matrix is multiplied with a dimensional compatible vector matrix and the product is transformed vector matrix with some multiple of it then the vector matrix is called the eigenvector of given square matrix. For example:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \overset{\text{eigenvalue}}{4} \times \underset{\text{Eigen vector}}{\begin{bmatrix} 3 \\ 2 \end{bmatrix}} \qquad \text{.................................} \ 3.7$$

(b) Eigen value: Eigen values are closely related to eigenvectors. It is the the amount by which the original vector was scaled after multiplication by the square matrix.

3.1.2. Principal Component Analysis

Finally we come to Principal Components Analysis (PCA). What is it? It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data.

The other main advantage of PCA is that once we have found these patterns in the data, and we compress the data, by reducing the number of dimensions, without much loss of information. This technique used in image compression.

3.1.2.1. Method of PCA

Initially we have absorbed data with large number of variable of different sample, our aim is to analyse the data of different sample by reducing the dimension. Finally we will plot the graph of few principal components having highest variance capture.

Step 1: Standardization of raw data by subtracting their mean either row wise or column wise. This produces a data set whose mean is zero.

Step 2: calculate the covariance matrix. The size of matrix depends on the no of variable, if n number of variable for each sample then n×n size covariance matrix will be formed.

Step 3: calculate the eigenvalues and eigenvector of covariance matrix. It is important to notice that these eigenvectors are both unit eigenvectors i.e. their lengths are both 1. This is very important for PCA. For n×n matrix we can have n number of eigenvector and n number of eigenvalues. The eigenvector corresponds to highest value of eigenvalue is contains maximum variability of variable.

Step 4: plot graph between PC1 and PC2.

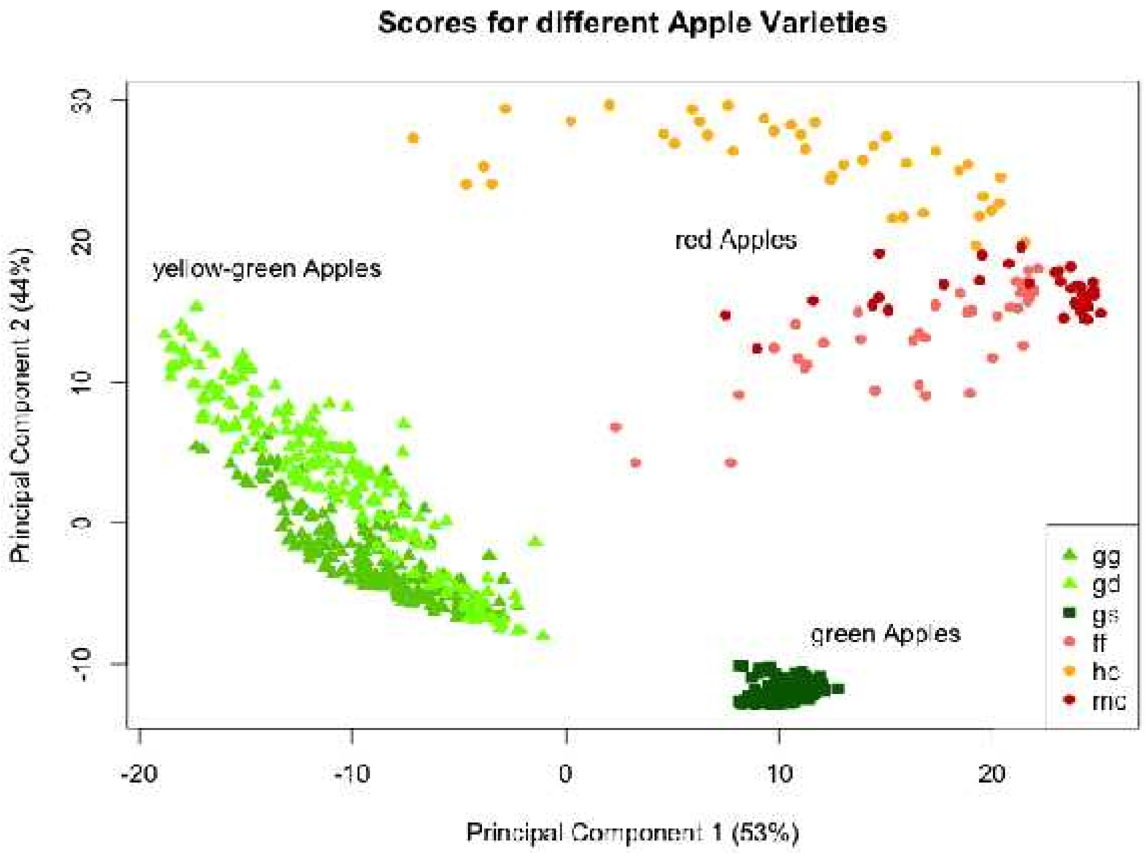


Figure 3.1. PCA plot of different variety of apple [8].

3.2. Principle Component Regression

PCR is a two-step multivariate calibration method: in the first step, a Principal Component Analysis, PCA, of the data matrix X is performed. The measured variables (e.g., absorbance at different wavelengths) are converted into new ones (scores on latent variables). This is followed by a multiple linear regression step, MLR, between the scores obtained in the PCA step and the characteristic y to be modelled.

$$X=\begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad \text{.....................................} \ 3.8$$

where $X_1 = [X_{11}\ X_{12}\ \ldots\ X_{1P}]$ is the row vector containing the absorbance measured at p wavelengths (the spectrum) for the first sample, x2 is the row vector containing the spectrum for the second sample and so on. By manually setting the projection onto the principal component directions with small eigenvalues set to 0 (i.e., only keeping the large ones), dimension reduction is achieved.

3.2.1. Principal Component Regression Model

Following the usual notation, suppose our regression equation may be written in matrix form as

$$Y = XB + e \qquad \text{.. } 3.9$$

where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors or residuals.

3.2.1.1. PC Regression Basics

In ordinary least squares, the regression coefficients are estimated using the formula

$$\hat{B} = (X^{'}X)X^{'}Y^{-1} \qquad \text{.. } 3.10$$

Note that since the variables are standardized, $X^{'}X$ =R, where R is the correlation matrix of independent variables. To perform principal components (PC) regression, we transform the independent variables to their principal components.

Mathematically, we write X'X = PDP'= Z'Z where D is a diagonal matrix of the eigenvalues of $X^{'}X$, P is the eigenvector matrix of $X^{'}X$, and Z is a data matrix (similar in structure to X) made up of the principal components. P is orthogonal so that P’P = I.

When we regress Y on Z_1 and Z_2, multicollinearity is no longer a problem. We can then transform our results back to the X scale to obtain estimates of B. These estimates will be biased, but we hope that the size of this bias is more than compensated for by the decrease in variance. That is, we hope that the mean squared error of these estimates is less than that for least squares.

Mathematically, the estimation formula becomes

$$\hat{A} = (Z^{'}Z)Z^{'}Y = DZ^{'}Y \qquad \text{................................ } 3.11$$

because of the special nature of principal components. Notice that this is ordinary least squares regression applied to a different set of independent variables. The two sets of regression coefficients, A and B, are related using the formulas:

$$A = P'B \qquad \text{............................ 3.12}$$

and

$$B = PA \qquad \text{............................ 3.13}$$

Omitting a principal component may be accomplished by setting the corresponding element of A equal to zero.

Hence, the principal components regression may be outlined as follows:

1. Complete a principal components analysis of the X matrix and save the principal components in Z.

2. Fit the regression of Y on Z obtaining least squares estimates of A.

3. Set the last element of A equal to zero.

4. Transform back to the original coefficients using B = PA.

3.2.2. Principal Component Regression Model for Unknown Data Prediction

PCR model is used for quantitative analysis, we construct a model with some known data and use that model to predict unknown data sample. Let us see the various step we need to follow to make a optimize model which can able to predict unknown sample with a high accuracy without over-fitting.

3.2.2.1. Pre-processing of Raw Data

Pre-processing of raw data is needed before data analysis. Proper application of spectroscopic data pre-processing, to reduce and correct interferences such as overlapped bands, baseline drifts, scattering, and pathlength variation.

3.2.2.1.1. Reduction of Non-linearity

A very different type of pre-processing is applied to correct for the non-linearity due to measuring transmittance or reflectance. To decrease non-linearity problems, reflectance (R) or transmittance (T) is transformed into absorbance (A):

$$A = \log_{1}\ (1/R) \qquad \text{........................... 3.14}$$

The equipment normally provides these values directly. For solid samples another approach is the Kubelka-Munk transformation [1]. In this case, the reflectance values are transformed into Kubelka-Munk units (K/S), using the equation:

$$\frac{K}{S} = \frac{(1-R)^2}{2R} \qquad \text{.............................. 3.15}$$

where K is the absorption coefficient and S the scatter coefficient of the sample at a given wavelength.

3.2.2.1.2. Noise Reduction and Differentiation

When applying signal processing, the main aim is to remove part of the noise present in the signal or to eliminate some sources of variation (e.g. background) not related to the measured y-variable. It is also possible to try and increase the differences in the contribution of each component to the total signal and in this way makes certain wavelengths more selective. The type of pre-processing depends on the nature of the signal.

General purpose methodologies are smoothing and differentiation. By smoothing one tries to reduce the random noise in the instrumental signal. The most used chemometric methodology is the one proposed by Savitzky and Golay. It is a moving window averaging method. The principle of the method is that, for small wavelength intervals, data can be fitted by a polynomial of adequate degree, and that the fitted values are a better estimate than those measured, because some noise has been removed. For the initial window the method takes the first 2m+1 points and fits, by least squares, the corresponding polynomial of order o. The fitted value for the point in position m replaces the measured value. After this operation, the window is shifted one point and the process is repeated until the last window is reached. Instead of calculating the corresponding polynomial each time, if data have been obtained at equally spaced intervals, the method uses tabulated coefficients in such a way that the fitted value for the centre point in the window is computed as:

$$x_{ij}^{*} = \frac{\sum_{k=-m}^{m} c_k x_{i,j+k}}{Norm} \qquad \text{.............................. 3.16}$$

Where x_{ij}^{*} represents the fitted value for the center point in the window, $x_{i,j+k}$ represents the 2m+1 original values in the window, ck is the appropriate coefficient value for each point and Norm is a normalising constant. Because the values of ck are the same for all windows, provided the window size and the polynomial degree are kept constant, the use of the tabulated coefficients simplifies and accelerates the computations.

For computational use, the coefficients for every window size and polynomial degree can be obtained in [5]. The user must decide the size of the window, 2m+1, and the order of the polynomial to be used. Errors in the original tables were corrected later [6]. These coefficients allow the smoothing of extreme points, which in the original method of Savitzky-Golay had to be removed. Recently, a methodology based on the same technique has been proposed [7], where the degree of the polynomial used is optimised in each window. This methodology has been called Adaptive-Degree Polynomial Filter (ADPF).

3.2.2.1.3. Methods Specific for NIR

The following methods are applied specifically to NIR data of solid samples. Variation between individual NIR diffuse reflectance spectra is the result of three main sources:

- Non-specific scatter of radiation at the surface of particles.
- Variable spectral path length through the sample.
- chemical composition of the sample.

In calibration we are interested only in the last source of variance. One of the major reasons for carrying out pre-processing of such data is to eliminate or minimise the effects of the other two sources. For this purpose, several approaches are possible.

Multiplicative Scatter (or Signal) Correction (MSC) has been proposed . The light scattering or change in path length for each sample is estimated relative to that of an ideal sample. In principle this estimation should be done on a part of the spectrum which does not contain chemical information, i.e. influenced only by the light scattering. However the areas in the spectrum that hold no chemical information often contain the spectral background where the SNR may be poor. In practice the whole spectrum is sometimes used. This can be done provided that chemical differences between the samples are small. Each spectrum is then corrected so that all samples appear to have the same scatter level as the ideal. As an estimate of the ideal sample, we can use for instance the average of the calibration set. MSC performs best if an offset correction is carried out first. For each sample:

$$x_i = a + b\bar{x}_j + e \qquad \text{.......................} \ 3.17$$

where x_i is the NIR spectrum of the sample, and $\bar{x}_j$ symbolises the spectrum of the ideal sample (the mean spectrum of the calibration set). For each sample, a and b are estimated by ordinary least-squares regression of spectrum xi vs. spectrum $\bar{x}_j$ over the available wavelengths. Each value x_{ij} of the corrected spectrum xi (MSC) is calculated as:

$$x_{ij}(MSC) = \frac{x_{ij} - a}{b}; j = 1,2,.., p \qquad \text{...........................} \ 3.18$$

The mean spectra must be stored in order to transform in the same way future spectra. Standard Normal Variate (SNV) transformation has also been proposed for removing the multiplicative interference of scatter and particle size. An example is given in figure 3a, where several samples of wheat are measured. SNV is designed to operate on individual sample spectra. The SNV transformation centres each spectrum and then scales it by its own standard deviation:

$$x_{ij}(SNV) = \frac{x_{ij} - \bar{x}_i}{SD}; j = 1,2,.., p \qquad \text{.................................} \ 3.19$$

where x_{ij} is the absorbance value of spectrum i measured at wavelength j, x_i is the absorbance mean value of the uncorrected ith spectrum and SD is the standard deviation of

$$\sqrt{\frac{\sum_{j=1}^{p}(x_{ij} - \bar{x}_i)^2}{p-1}} \qquad \text{......................................} \ 3.20$$

the p absorbance values. Spectra treated in this manner have always zero mean and variance equal to one, and are thus independent of original absorbance values.

The global absorbance of NIR spectra is generally increasing linearly with respect to the wavelength , but it increases curvilinearly for the spectra of densely packed samples. A second-degree polynomial can be used to standardise the variation in curvilinearity:

$$x_i = a\text{a}^{*2} + b\text{a}^{*} + c + e_i \qquad \text{......................................} \ 3.21$$

where xi symbolises the individual NIR spectrum and l* the wavelength. For each sample, a, b and c are estimated by ordinary least-squares regression of spectrum xi vs. wavelength over the range of wavelengths. The corrected spectrum xi (DTR) is calculated by:

$$x_i(DRT) = x_i - a\text{a}^{*2} - b\text{a}^{*} - c = e_i \qquad \text{...........................}3.22$$

Normally de-trending is used after SNV transformation. Second derivatives can also be employed to decrease baseline shifts and curvilinearity, but in this case noise and complexity of the spectra increases.

It has been demonstrated that MSC and SNV transformed spectra are closely related and that the difference in prediction ability between these methods seems to be fairly small.

3.2.2.1.4. Selection of Pre-processing Methods in NIR

The best pre-processing method will be the one that finally produces a robust model with the best predictive ability. Unfortunately there seem to be no hard rules to decide which pre-processing to use and often the only approach is trial and error. The development of a methodology that would allow a systematic approach would be very useful. It is possible to obtain some indication during pre-processing. For instance, if replicate spectra have been measured, a good pre-processing methodology will produce minimum differences between replicates though this does not necessarily lead to optimal predictive value. If only one measure per sample is given, it can be useful to compute the correlation between each of the original variables and the property of interest and do the same for the transformed variable. It is likely that good correlations will lead to a good prediction. However, this approach is uni-variate and therefore does not give a complete picture of predictive ability. Depending on the physical state of the samples and the trend of the spectra, a background and/or a scatter correction can be applied. If only background correction is required, offset correction is usually preferable over differentiation, because with the former the SNR is not degraded and because differentiation may lead to less robust models over time. If additionally scatter correction is required, SNV and MSC yield very similar results.

SNV equation

$$x_{ij} = \frac{(x_{ij} - \bar{x}_i)}{SD}; j = 1,2,...p \qquad \text{.............................3.23}$$

MSC equation:

$x_{ij} = \frac{(x_{ij} - a)}{b}; j = 1,2,...p$ For each sample, a and b are estimated by ordinary least-squares regression of spectrum x_i vs. Spectrum $\bar{x}_{ij}$ over the available wavelengths. An advantage of SNV is that spectra are treated individually, while in MSC one needs to refer to other spectra. When a change is made in the model, e.g. if, because of clustering, it is decided to make two local models instead of one global one, it may be necessary to repeat the MSC pre-processing. Non-linear behaviour between X and y appears (or increases) after some of the pre-processing methods. This is the case for instance for SNV. However this does not cause problems provided the differences between spectra are relatively small.

3.2.2.2. PCA and Building the Model

Next step is to do PCA on pre-processed data. The score obtained by PCA is used for making model by using multiple linear regression method. The first step (PCA) does not involve y (concentration) values but only the X (absorbance) data. In the second, the following model is built:

$$y = f(t) = b_0 + \sum_{i=1}^{a} b_i t_i \qquad \text{........................ 3.24}$$

When developing the mathematical model we must answer two questions, namely, how many variables must be entered (r) and which ones. The most common way, which we will call Top-Down selection (TD), is to add the PCs in the order of explained variance until validation (explained in next section) shows that there is no significant improvement in the prediction.

The explained variance is not necessarily related with the property of interest, because when the PCs are obtained without considering the values of y. This means that in TD large PCs, irrelevant for y, may still be included in the model. y can be taken into account by entering the PCs in order of the importance for the model. In building the model, it seems more logical to select the PCs by order of correlation or prediction ability for the y variable.

When the selection of the PCs is performed by correlation, first the correlation coefficients between the scores on the different PCs and the y-values are obtained. Then these values are sorted by absolute value (irrespective of sign) and the PCs are entered in this order until no further significant improvement is obtained in the validation step.

In the selection of PCs by prediction ability, the prediction errors for all the models with one PC are calculated, and the one that provides the minimum value is selected. Two approaches have been proposed to calculate the prediction error: In [98] a test set was used and in [99] leverage-corrected residuals (section 9) were applied. When the first PC to be entered has been selected, all the possible models with this PC and a second one are tested, and the model with minimum prediction error is selected. The procedure continues until no further significant improvement is obtained. The method based on correlation is the simplest and is therefore to be preferred.

3.2.2.3. Model Optimisation and Validation

3.2.2.3.1. Training, Optimisation and Validation

The determination of the optimal complexity of the model (the number of PCs that should be included in the model) requires the estimation of the prediction error that can be reached. Ideally, a distinction should be made between training, optimisation and validation. Training is the step in which the regression coefficients are determined for a given model. In PCR, this means that the b-coefficients are determined for a model that includes a given set of PCs. Optimisation consists of comparing different models and deciding which one gives best prediction. In PCR, the usual procedure is to determine the predictive power of models with 1, 2, 3, … PCs and to retain the best one. Validation is the step in which the prediction with the chosen model is tested independently. In practice, as we will describe later, because of practical constraints in the number of samples and/or time, less than three steps are often included. In particular, analysts rarely make a distinction between optimisation and validation and the term validation is then sometimes used for what is essentially an optimisation. While this is acceptable to some extent, in no case should the three steps be reduced to one. In other words, it is not acceptable to draw conclusions about optimal models and/or quality of prediction using only a training step. The same data should never be used for training, optimising and validating the model. If we do, it is possible and even probable that we will overfit the model and prediction error obtained in this way may be over-optimistic. Overfitting is the result of using a too complex model. Consider a univariate situation in which three samples are measured. The y = f(x) model really is linear (first order), but the experimenter decides to use a quadratic model instead. The training step will yield a perfect result: all points are exactly on the line. If, however, new samples are predicted, then the performance of the quadratic model will be worse than the performance of the linear one.

3.2.2.3.2. Measures of Predictive Ability

Several statistics are used for measuring the predictive ability of a model. The prediction error sum of squares, PRESS, is computed as:

$$PRESS = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad \text{.................... 3.25}$$

where y_i is the actual value of y for object i and $\hat{y}_i$ the y-value for object i predicted with the model under evaluation, ei is the residual for object i (the difference between the predicted and the actual y-value) and n is the number of objects for which y_i is obtained by prediction.

The mean squared error of prediction (MSEP) is defined as the mean value of PRESS:

$$MSEP = \frac{PRESS}{n} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n} \qquad \text{.......................... 3.26}$$

It’s square root is called root mean squared error of prediction, RMSEP:

$$RMSEP = \sqrt{MSEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \qquad \text{.............................. 3.27}$$

All these quantities give the same information. In the chemometrics literature it seems that RMSEP values are preferred, partly because they are given in the same units as the y-variable.

3.2.2.3.3. Optimisation

The RMSEP is determined for models with increasing complexity. PCs are included according to the Top-Down or Best Subset Selection procedure .Usually the result is presented as a plot showing RMSEP as a function of the number of components and is called the RMSEP curve. This curve often shows an intermediate minimum and the number of PCs for which this occurs is then considered to be the optimal complexity of the model. A problem which is sometimes encountered is that the global minimum is reached for a model with a very high complexity. A more parsimonious model is often more robust (the parsimonity principle). Therefore, it has been proposed to use the first local minimum or a deflection point is used instead of the global minimum. If there is only a small difference between the RMSEP of the minimum and a model with less complexity, the latter is often chosen. The decision on whether the difference is considered to be small is often based on the experience of the analyst. We can also use statistical tests that have been developed to decide whether a more parsimonious model can be considered statistically equivalent. In that case the more parsimonious model should be preferred. An F-test or a randomisation t-test have been proposed for this purpose. The latter requires less statistical assumptions about data and model properties, and is to be preferred. However in practice it does not always seem to yield reliable results.

3.2.2.3.4. Validation

The model selected in the optimisation step is applied to an independent set of samples and the y-values (i.e. the results obtained with the reference method) and y_i -values (the results obtained with multivariate calibration) are compared. The interpretation is usually done visually: does the line with slope 1 and intercept 0 represent the points in the graph sufficiently well? It is necessary to check whether this is true over the whole range of concentrations (non-linearity) and for all meaningful groups of samples, e.g. for different clusters. If a situation is obtained when most samples of a cluster are found at one side of the line, a more complex modelling method or a model for each separate cluster of samples may yield better results.

Sometimes a least squares regression line between y and $\hat{y}$ is obtained and a test is carried out to verify that the joint confidence interval contains slope = 1 and intercept = 0. Similarly a paired t-test between y and $\hat{y}$ values can be carried out. This does not obviate, however, the need for checking non-linearity or looking at individual clusters.

An important question is what RMSEP to expect? If the final model is correct, i.e. there is no bias then the predictions will often be more precise than those obtained with the reference method, due to the averaging effect of the regression. However, this cannot be proved from measurements on validation samples, the reference values of which were obtained with the reference method. The RMSEP value is limited by the precision (and accuracy) of the reference method. For that reason, RMSEP can be applied at the optimisation stage as a kind

of target value. An alternative way of deciding on model complexity therefore is to select the lowest complexity which leads to an RMSEP value comparable to the precision of the reference method.

3.2.2.3.5. External Validation

In principle, the same data should not be used for developing, optimising and validating the model. If we do this, it is possible and even probable that we will overfit the model and prediction errors obtained in this way may be over-optimistic. Terminology in this field is not standardised. We suggest that the samples used in the training step should be called the training set, those that are used in optimisation the evaluation set and those for the validation the validation set. Some multivariate calibration methods require three data sets. This is the case when neural nets are applied (the evaluation set is then usually called the monitoring set). In PCR and related methods, often only two data sets are used (external validation) or, even only one (internal validation). In the latter case, the existence of a second data set is simulated. We suggest that the sum of all sets should be called the calibration set. Thus the calibration set can consist of the sum of training, evaluation and validation sets, or it can be split into training and test set, or it can serve as the single set applied in internal validation. Applied with care, external and internal validation methods will warn against overfitting.

External validation uses a completely different group of samples for prediction (sometimes called the test set) from the one used for building the model (the training set). Care should be taken that both sample sets are obtained in such a way that they are representative for the data being investigated. One should be aware that with an external test set the prediction error obtained may depend to a large extent on how exactly the objects are situated in space in relationship to each other.

It is important to repeat that, in the presence of measurement replicates, all of them must be kept together either in the test set or in the training set when data splitting is performed. Otherwise, there is no perturbation, nor independence, of the statistical sample.

The preceding paragraphs apply when the model is developed from samples taken from a process or a natural population. If a model was created with artificial samples with y-values outside the expected range of y-values to be determined, for the reasons explained in section 10, then the test set should contain only samples with y-values in the expected range.

3.2.2.3.6. Internal Validation

One can also apply what is called internal validation. Internal validation uses the same data for developing the model and validating it, but in such a way that external validation is simulated. Four different methodologies were employed:

a. Random splitting of the calibration set into a training and a test set. The splitting can then have a large influence on the obtained RMSEP value.

b. Cross-validation (CV), where the data are randomly divided into d so-called cancellation groups. A large number of cancellation groups correspond to validation with small perturbation of the statistical sample, whereas a small number of cancellation groups correspond to a heavy perturbation. The term perturbation is used to indicate that the data set used for developing the model in this stage is not the same as the one developed with all calibration objects. Too small a perturbation means that over-fitting is still possible. The validation procedure is repeated as many times as there are cancellation groups. At the end of the validation procedure each object has been once in the test set and d-1 times in the training set. Suppose there are 15 objects and 3 cancellation groups, consisting of objects 1-5, 6-10 and 11-15. We mentioned earlier that the objects should be assigned randomly to the cancellation groups, but for ease of explanation we have used the numbering above. The b-coefficients in the model that is being evaluated are determined first for the training set consisting of objects 6-15 and objects 1-5 function as test set, i.e. they are predicted with this model. The PRESS is determined for these 5 objects. Then a model is made with objects 1-5 and 11-15 as training and 6-10 as test set and, finally, a model is made with objects 1-10 in the training set and 11-15 in the test set. Each time the PRESS value is determined and eventually the three PRESS values are added, to give a value representative for the whole data set (PRESS values are more indicated here to RMSEP values, because PRESS values are variances and therefore additive).

c. leave-one-out cross-validation (LOO-CV), in which the test sets contain only one object (d = n). Because the perturbation of the model at each step is small (only one object is set aside), this procedure tends to over-fit the model. For this reason the leave-more-out methods described above may be preferable. The main drawback of LOO-CV is that the computation is slow because PCA must be performed on each matrix after object deletion. Fast algorithms are described where the speed of calculation is greatly improved.

Another way to improve the speed is based on the use of leverage-corrected residuals, where the leave-one-out cross-validated values are replaced by the fitted values from the least squares model corrected by the leverage value, using the equation:

$$\hat{e}_i^{1c}(r) = \frac{\hat{e}_i^{1c}(r)}{1-h_i(r)} \qquad \text{......................................} \quad 3.28$$

where $\hat{e}_i^{1s}(r)$ is the obtained residual for object i after fitting r factors by using least-squares model, hi(r) is the leverage value for object i after fitting r factors and $\hat{e}_i^{1c}(r)$ is the corresponding predicted residual when r factors are used in the leave-one-out cross-validation of the model. It is fast to perform, compared to complete cross-validation, because only one singular value decomposition of the data matrix is needed. In the absence of outliers, results from leave-one-out cross-validation and leverage-correction are similar. However, leverage-correction must be employed as a quick-and-dirty method and the results must be confirmed later with another method.

d. Repeated random splitting (repeated evaluation set method) (RES). The procedure described in a, is repeated many times. In this way, at the end of the validation procedure, we hope that an object has been in the test set several times with different companions. Stable results are obtained after repetition of the procedure several times (even hundreds of times). To have a good picture of the prediction error we have to use both low and high percentages of objects in the evaluation set [2].

3.2.3. Final Model

The model which gives best prediction without overfitting is selected and its reliability and future use for other sample prediction will depend upon the accuracy of the model. If the accuracy of model is greater than 80% for most of the sample then the model is good.

3.3. Partial Least Square Regression

3.3.1. Introduction

Partial least square regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression technique. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). It originated in the social sciences (specifically economy, Herman Wold 1966) but became popular first in chemometrics (i.e., computational chemistry) due in part to Herman's son Svante, and in sensory evaluation (Martens & Naes, 1989). But Partial least square regression is also becoming a tool of choice in the social sciences as a multivariate technique for non-experimental and experimental data alike (e.g., neuroimaging). It was first presented as an algorithm akin to the power method (used for computing eigenvectors) but was rapidly interpreted in a statistical framework [3].

Research in science and engineering often involves using controllable and/or easy-to-measure variables (factors) to explain, regulate, or predict the behavior of other variables (responses). When the factors are few in number, are not significantly redundant (collinear), and have a well-understood relationship to the responses, then multiple linear regression (MLR) can be a good way to turn data into information. However, if any of these three conditions breaks down, MLR can be inefficient or inappropriate. In such so-called soft science applications, the researcher is faced with many variables and ill-understood relationships, and the object is merely to construct a good predictive model. For example, spectrographs are often used to estimate the amount of different compounds in a chemical sample. In this case, the factors are the measurements that comprise the spectrum; they can number in the hundreds but are likely to be highly collinear. The responses are component amounts that the researcher wants to predict in future samples.

Partial least squares (PLS) is a method for constructing predictive models when the factors are many and highly collinear. Note that the emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables. For example, PLS is not usually appropriate for screening out factors that have a negligible effect on the response. However, when prediction is the goal and there is no practical need to limit the number of measured factors, PLS can be a useful tool.

3.3.1.1. How Does PLS Work?

In principle, MLR can be used with very many factors. However, if the number of factors gets too large (for example, greater than the number of observations), you are likely to get a model that fits the sampled data perfectly but that will fail to predict new data well. This phenomenon is called over-fitting. In such cases, although there are many manifest factors, there may be only a few underlying or latent factors that account for most of the variation in the response. The general idea of PLS is to try to extract these latent factors, accounting for as much of the manifest factor variation as possible while modelling the responses well. For this reason, the acronym PLS has also been taken to mean ‘‘projection to latent structure.’’ It should be noted, however, that the term ‘‘latent’’ does not have the same technical meaning in the context of PLS as it does for other multivariate techniques.

3.3.2. Goal

The goal of Partial least square regression is to predict dependant variable (say Y) from independent variable (say X) and to describe their common structure. When Y is a vector and X is full rank, this goal could be accomplished using ordinary multiple regression. When the number of predictors is large compared to the number of observations, X is likely to be singular and the regression approach is no longer feasible (i.e., because of multicollinearity). Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using stepwise methods) another one, called principal component regression, is to perform a principal component analysis (pca) of the X matrix and then use the principal components of X as regressors on Y. The orthogonality of the principal components eliminates the multicolinearity problem. But, the problem of choosing an optimum subset of predictors remains. A possible strategy is to keep only a few of the first components. But they are chosen to explain X rather than Y, and so, nothing guarantees that the principal components, which “explain” X, are relevant for Y. By contrast, PLS regression finds components from X that are also relevant for Y. Specifically, PLS regression searches for a set of components (called latent vectors) that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the covariance between X and Y. This step generalizes PCA. It is followed by a regression step where the decomposition of X is used to predict Y.

The overall goal (shown in the lower box) is to use the factors to predict the responses in the population. This is achieved indirectly by extracting latent variables T and U from sampled factors and responses, respectively. The extracted factors T (also referred to as X-scores) are used to predict the Y-scores U, and then the predicted Y-scores are used to construct predictions for the responses. This procedure actually covers various techniques, depending on which source of variation is considered most crucial.

- Principal Components Regression (PCR): The X-scores are chosen to explain as much of the factor variation as possible. This approach yields informative directions in the factor space, but they may not be associated with the shape of the predicted surface.
- Maximum Redundancy Analysis (MRA) (van den Wollenberg 1977): The Y-scores are chosen to explain as much of the predicted Y variation as possible. This approach seeks directions in the factor space that are associated with the most variation in the responses, but the predictions may not be very accurate.
- Partial Least Squares: The X- and Y-scores are chosen so that the relationship between successive pairs of scores is as strong as possible. In principle, this is like a robust form of redundancy analysis, seeking directions in the factor space that are associated with high variation in the responses but biasing them toward directions that are accurately predicted.

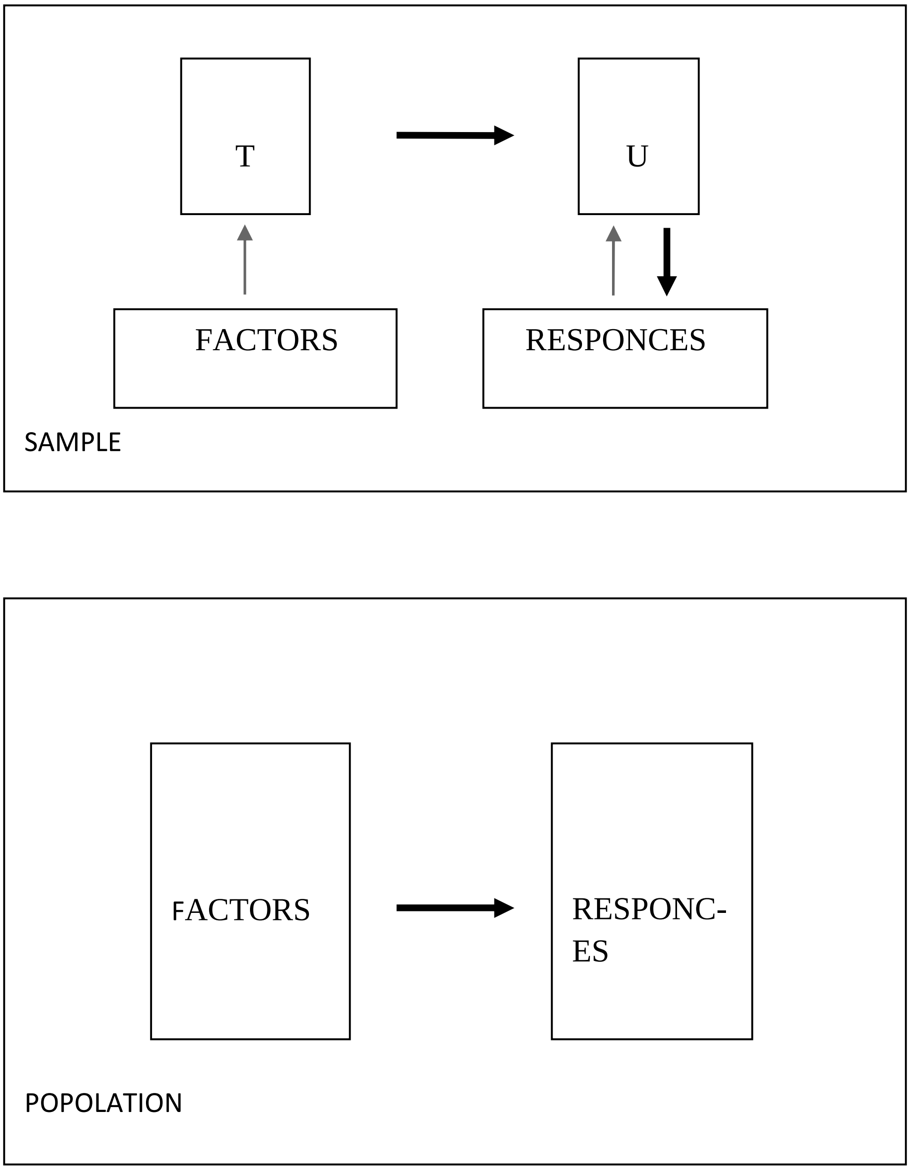


Figure 3.2. Indirect modelling [4].

3.3.3. Simultaneous Decomposition of Predictors and Dependent Variables

Pls regression decomposes both X and Y as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are decomposed as X = TPT with TTT = I with I being the identity matrix (some variations of the technique do not require T to have unit norms). By analogy with PCA T is called the score matrix, and P the loading matrix (in PLS regression the loadings are not orthogonal). Likewise, Y is estimated as $\hat{Y}$ = TBCT where B is a diagonal matrix with the “regression weights” as diagonal elements (see below for more details on these weights). The columns of T are the latent vectors. When their number is equal to the rank of X, they perform an exact decomposition of X. Note, however, that they only estimate Y. (i.e., in general $\hat{Y}$ is not equal to Y).

3.3.4. PLS Regression and Covariance

The latent vectors could be chosen in a lot of different ways. In fact in the previous formulation, any set of orthogonal vectors spanning the column space of X could be used to play the role of T. In order to specify T, additional conditions are required. For PLS regression this amounts to finding two sets of weights w and c in order to create (respectively) a linear combination of the columns of X and Y such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors t = Xw and u = Yc with the constraints that $w^T w$ = 1, $t^T t$ = 1 and $t^T u$ be maximal. When the first latent vector is found, it is subtracted from both X and Y and the procedure is re-iterated until X becomes a null matrix [3].

3.3.5. A PLS Regression Algorithm

The properties of PLS regression can be analysed from a sketch of the original algorithm. The first step is to create two matrices: E = X and F = Y. These matrices are then column centred and normalized (i.e., transformed into Z-scores). Sum of square obtained by these matrices are denoted SSX and SSY. Before starting the iteration process, the vector u is initialized with random values. (in what follows the symbol ∞ means “to normalize the result of the operation”).

Step 1. w ∞ $E^T u$ (estimate X weights).

Step 2. t ∞ Ew (estimate X factor scores).

Step 3. c ∞ $F^T t$ (estimate Y weights).

Step 4. u = Fc (estimate Y scores).

If t has not converged, then go to Step 1, if t has converged, then compute the value of b which is used to predict Y from t as b = $t^T u$, and compute the factor loadings for X as p = $E^T t$. Now subtract (i.e., partial out) the effect of t from both E and F as follows E = E−tp^T and F = F−btc^T. The vectors t, u, w, c, and p are then stored in the corresponding matrices, and the scalar b is stored as a diagonal element of B. The sum of squares of X (respectively Y) explained by the latent vector is computed as $p^T p$ (respectively b^2), and the proportion of variance explained is obtained by dividing the explained sum of squares by the corresponding total sum of squares (i.e., SSX and SSY). If E is a null matrix, then the whole set of latent vectors has been found, otherwise the procedure can be re-iterated from Step 1 on.

3.3.6. PLS Regression and the Singular Value Decomposition

The iterative algorithm presented above is similar to the power method which finds eigenvectors. So PLS regression is likely to be closely related to the eigen and singular value decompositions, and this is indeed the case. For example, if we start from Step 1 which computes: w $\infty E^T u$, and substitute the rightmost term iteratively, we find the following series of equations: w ∞ $E^T u$ $\infty E^T Fc$ $\infty E^T FFTt$ ∞ $E^T FFTtEw$. This shows that the first weight vector w is the first right singular vector of the matrix $X^T Y$. Similarly, the first weight vector c is the left singular vector of $X^T Y$. The same argument shows that the first vectors t and u are the first eigenvectors of $XX^T YYT$ and $YY^T XXT$.

3.3.7. Prediction of the Dependent Variables

The dependent variables are predicted using the multivariate regression formula as $\hat{Y}$ = TBC^T = XB_{PLS} with B_{PLS}= $(P^{T+})BC^T$ (where (P^{T+}) is the Moore-Penrose pseudo-inverse of PT). If all the latent variables of X are used, this regression is equivalent to principal component regression. When only a subset of the latent variables is used, the prediction of Y is optimal for this number of predictors. An obvious question is to find the number of latent variables needed to obtain the best generalization for the prediction of new observations. This is, in general, achieved by cross-validation techniques such as bootstrapping. The interpretation of the latent variables is often helped by examining graphs akin to PCA graphs.

3.4. Conclusion

While indentifying quality and quantity of chemical composition in any mater we need and specific instruments and experimental setups with required compatible environment which leads to cost and time consumption.

To overcome above needs statistics play a vital role in pattern recognition analysis field. On the basis of statistical data a hypothesis can be made, whether these data can be useful or not for further unknown sample prediction. Multivariate analysis is a technique that is used to analyse data and also to make a regression model which will be helpful to estimate and analyse new unknown sample. Multivariate regression model includes principal component analysis (PCA), principal component regression (PCR), partial least square (PLS) regression model. In above model, on the basis of sample data collected using instrument can be used to construct a prediction model with help of software like MATLAB, LAB-VIEW, MINI-TAB etc.

Above analysis (chapter 4) describes how to perform PCA where the principal component (PC) explains variability in variable. The first few PC contains most of the variance. PCA is basically used to obtain principal component for further analysis along with segregate each sample clustering in a plot through reducing the dimension. As raw data had taken from instrument with a huge number of observation and variable and it is very difficult to visualize these data in a normal plot until large dimension graphics devices is not available. PCA reduces these dimension with a minute lose of information and confirm the repeatability of instrument on the basis of clustering formed.

Next is to construct a model which has capability to provide acceptable accuracy during quantitative prediction. Principal component of the data sample is used to construct a model through the regression technique. This is called principal component regression analysis. As this model only uses the PC’s of independent variable so its accuracy is not as good as PLS regression model.

Partial least square regression analysis uses latent vector to predict sample, here PCA of both independent and dependent variable data is done and the correlation between scores of both PCA is optimised with certain accuracy and with the help of scores value of dependant variable new sample prediction is performed. The latent vector is nothing but PLS component which provides the number of dimension through which the model can observe unknown sample. It is needed to achive the goal is to identify those PLS component which provide better accuracy without over-fitting. During the analysis,a PLS model is constructed which gives a satisfactory result.

References

[1] P. Kubelka, New contributions to the optics of intensely light-scattering materials part 1, Journal of the optical Society of America 38(5) (1948) 448-457.

[2] http://www.vub.ac.be/fabi/multi/pcr/pages/menu/m1.html

[3] Partial Least Squares (PLS) Regression. Herv´e Abdi1 The University of Texas at Dallas

[4] An Introduction to Partial Least Squares Regression Randall D. Tobias, SAS Institute Inc., Cary, NC

[5] P.A. Gorry, General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method, Anal. Chem. 62 (1990) 570-573.

[6] J. Steinier, Y. Termonia and J. Deltour, Comments on smoothing and differentiation of data by simplified least square procedure, Anal. Chem. 44 (1972) 1906-1909.

[7] P. Barak, Smoothing and differentiation by an adaptive-degree polynomial filter, Anal. Chem. 67 (1995) 2758-2762.

[8] http://oceanoptics.com/chemometric-analysis-of-food-quality/

Chapter 4

Instrument review and Multivariate Analysis of Data

4.1. Instrument Review

4.1.1. DWARF-Star NIR Detector

StellarNet’s newest NIR spectrometer, the DWARF-Star, is small, robust, and equipped with high performance InGaAs detector array for the 900-1700nm wavelength range and achieves resolving resolutions to 1.25nm. The DWARF-Star features no moving parts and is packaged in a small rugged metal enclosure (5”x3”x2”) for portable, process, and OEM applications. Advancements in electronic and optical design have allowed for size reduction never before achieved in a NIR spectrometer. The InGaAs detector is a Sensors Unlimited linear photo diode array with 512 pixels (1024 optional) 25μm by 500μm tall to provide maximum sensitivity. The detector has an integrated thermo electric cooler (TEC) maintained at –10 °C, stabilized within +/-0.1 °C. The NIR spectrometers accept a single strand SMA-905 terminated, low OH, fiber optic cable as input. Several models provide a variety of operational ranges and resolutions suitable for both spectroscopy and optical spectrum analysis. Each DWARF-Star includes free SpectraWiz® Software and a developer’s toolbox of source codes, customizable demo programs, and full spectroscopy applications in LAB-VIEW, Visual Basic, Delphi Pascal, and MS Visual C. High speed spectral data acquisition with advanced features, such as time series analysis and episodic data capture with rapid sample logging are standard features. Post processing techniques such as baseline correction, data smoothing, and spectral derivatives are included. Additionally, add-on chemometrics packages are available for complete multivariate calibration, analysis, and runtime with the DWARF-Star [1].

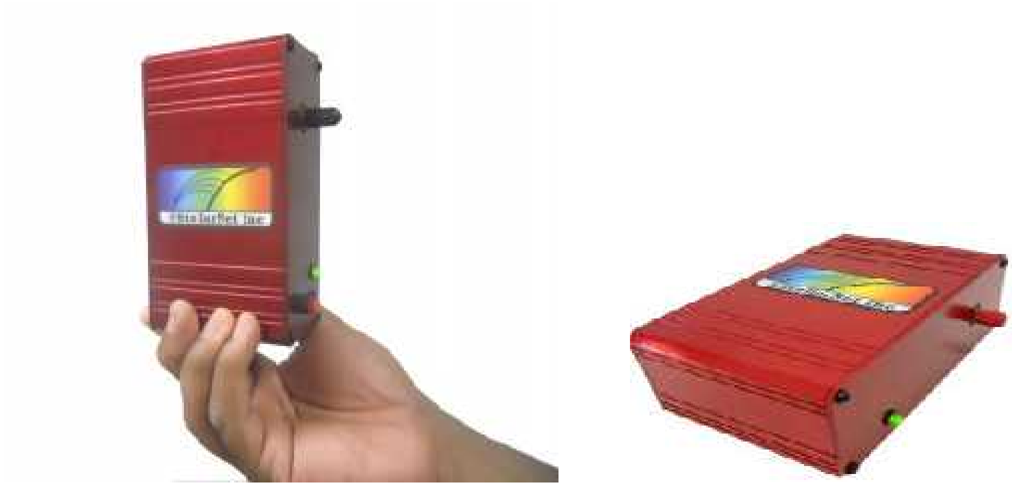


Figure 4.1. DWARF-Star NIR Detector [2].

The StellarNet DWARF-Star fiber optic spectrometers are available in several models, to provide optimal ranges and resolutions for various NIR applications in the 900-1700nm range. The standard detector is a 512 element photo diode array with 25 x 500μm tall pixels and has zero defects. The units interface to a PC via USB-2 and can be operated simultaneously with StellarNet UV-VIS spectrometers to provide a Dual-Detector Super Range (Dual-DSR) spectroscopy system. StellarNet also offers light sources, probes, and sampling accessories to facilitate virtually any NIR application. The miniature DWARF-Star NIR spectrometer is ideal for process analytical technology for industries such as food and drug, chemical, oil and gas, and plastics. The DWARF-Star's miniature size, low cost, and rugged design also make it ideal for the field, enabling on-site product analysis and quality control never before attainable [1].

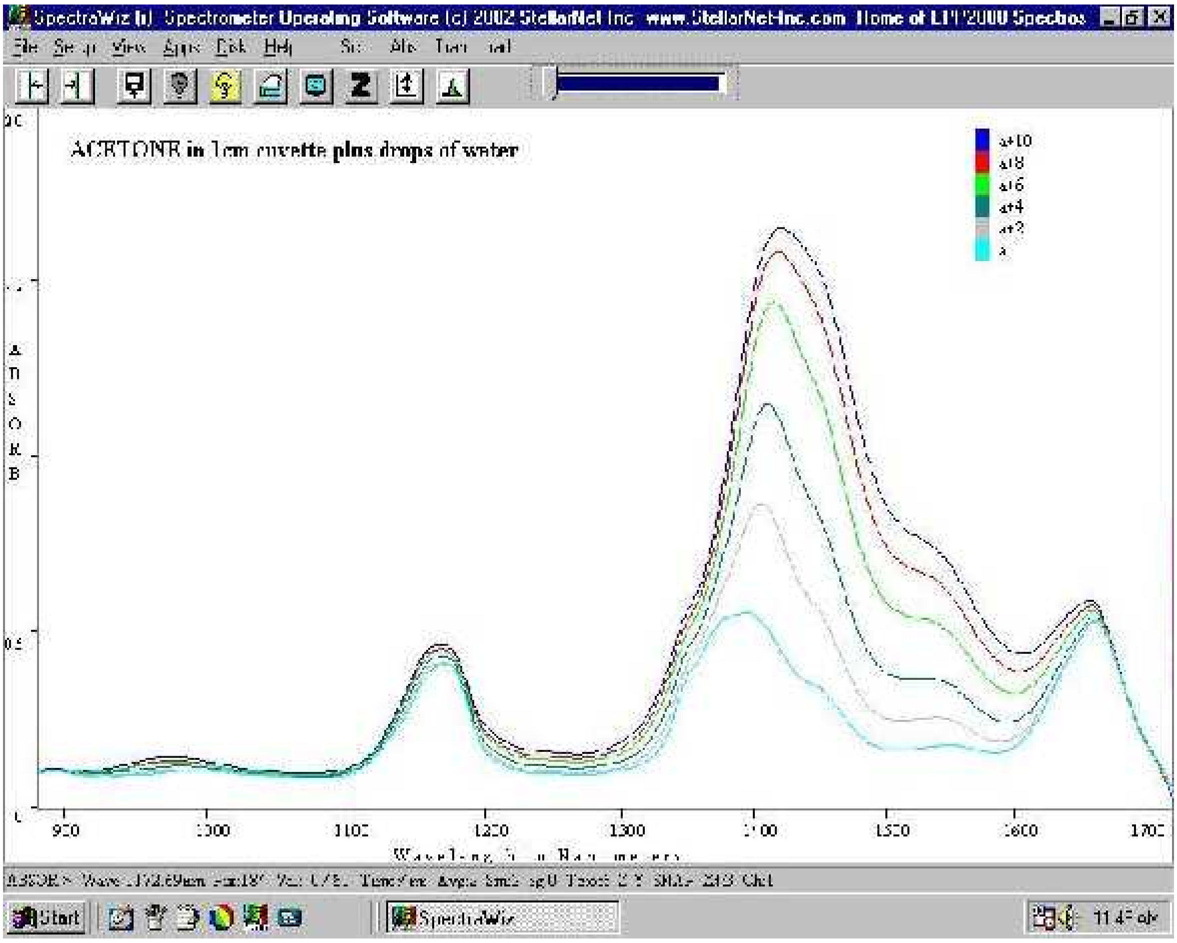

SpectraWiz (r) Spectrometer Operating Software (c) 2002 StellarNet Inc www.StellarNet-Inc.com Home of EPP2000 Spectros
File Setup View Apps Disk Help
ACETONE in 1cm cuvette plus drops of water
a+10
a+8
a+6
a+4
a+2
a
ABSORB
0.5
0
900
1000
1100
1200
1300
1400
1500
1600
1700
Wavelength in Nanometers
Start
SpectraWiz

Figure 4.2. DWARF-Star Sample Spectra for Acetone [3].

4.1.2. Spectroscopy Lamps and Light Sources.

he StellarNet portable Light Sources facilitate measurements in the UV, VIS, and NIR for various samples and application types using standard SMA-905 optical connectors. These low cost modular components are robust, reliable, and designed to last in portable and industrial environments.

4.1.2.1. VIS-NIR Sources (SL1 Tungsten Halogen) [4]

350 nm-2200nm Spectral Range- effective for colour, reflectance, transmittance, and absorbance measurements, Extensive Life- The SL1 has a 10,000 hour Tungsten Halogen lamp filled with Krypton gas, Great Compatibility/Versatility- manipulate output with colour-enhancing or signal-attenuating filters.Small Footprint- measures only 1.5” x 3” x 3.5”, Maximum Flexibility- several models and options to choose from to meet all your application needs [4].

Figure 4.3 SL1 Tungsten Halogen[4]

4.1.2.2. Tungsten Lamp

Tungsten lamp with diffused absorbance spectroscopy is used in the frequency range of 900 1700 nm. The overtones and recombination is formed in NIR region so it is usefull for the quantitative analysis purpose.

The NIR radiation is exposed on the sample at three emitting sources in radial path of a circle by making 120^{0} between the radial path, so that the radiation can cover almost all the sample powder.

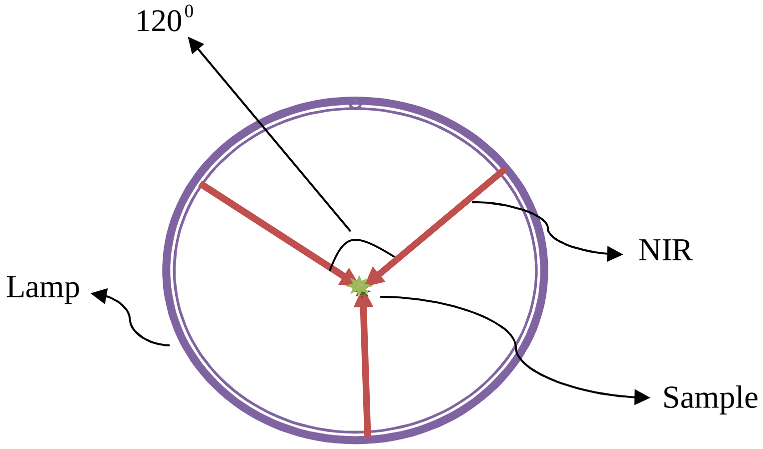


Figure 4.4. Exposure of NIR radiation on source using tungsten lamp.

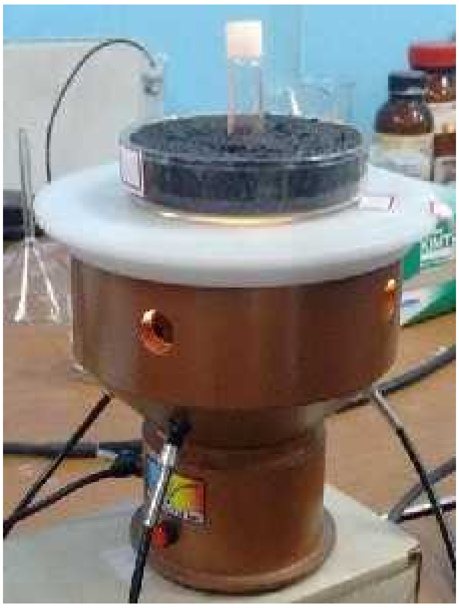


Figure 4.5. Tungsten lamp [5].

4.2. NIR Spectrum Measurement

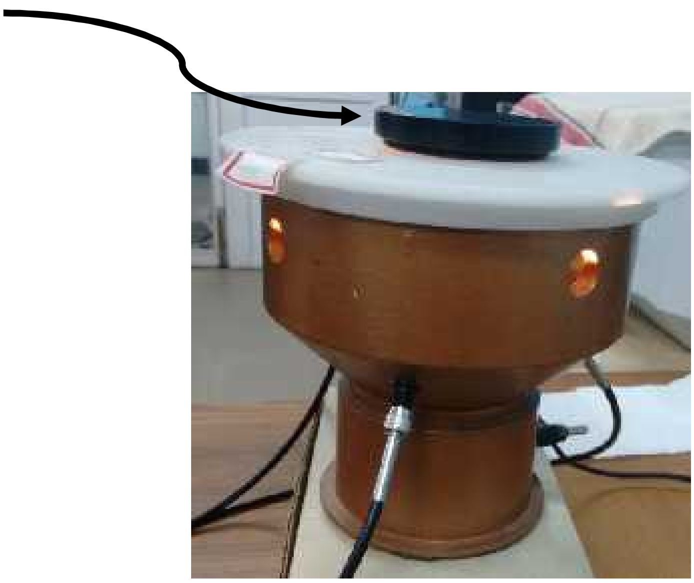
Due to the huge consumption of tea in the world, the tea companies need to take care of quality and taste of tea. Now a day peoples are more concern regarding quality and price so they are seeking for better quality of tea at reasonable price. The food industries has to play an important role here by taking care of food safety and quality of beverages like tea, since it is directly related to people's health and social progress. All these requirement and condition drags attention of researcher towards the tea beverage to make this product good as much as possible.

NIRS is a fast, non-destructive technique used for qualitative as well as quantitative detection of a sample. Since it uses infra-red ray to obtain spectrum of sample so it is non-destructive and almost no sample preparation is needed here. Near infrared (NIR) spectroscopy has been used to quantify several compounds having a diverse antioxidant capacity such as carotenoids, polyphenols, fatty acids and glucosinolates in a wide range of food commodities. for example, wine, dairy products, tea, fruit, vegetables, herbs, spices and cereals.

NIR spectroscopy is used to analyse data and with the use of multivariate prediction technique we tried to make prediction model for unknown data prediction.

4.2.1. Data Observation

Numbers of tea samples were taken from different tea garden of Asam in the middle of year 2015. All these variety of tea sample have different caffeine content, NIR spectroscopy have been used to obtain various peak between the frequency range 900 to 1700 nm. StellerNet NIR instrument compatible with SpectraWiz® software has been used which provide diffused absorbance versus wavenumber. Instrument was calibrated at dark and light condition first then the data was observed.



Black plate

Figure 4.6. Dark calibration

(a) Tea Sample

(b) Pure Caffeine

Figure 4.7. (a) tea sample (b) pure caffeine

Total 21 tea samples and one pure caffeine sample has been used to observe data. For each sample four observations at one position was taken then the sample tube is shaken well and rotated by 120^0 and data absorbed. For each sample twelve observations were taken. All the observation is kept save in ABS format.

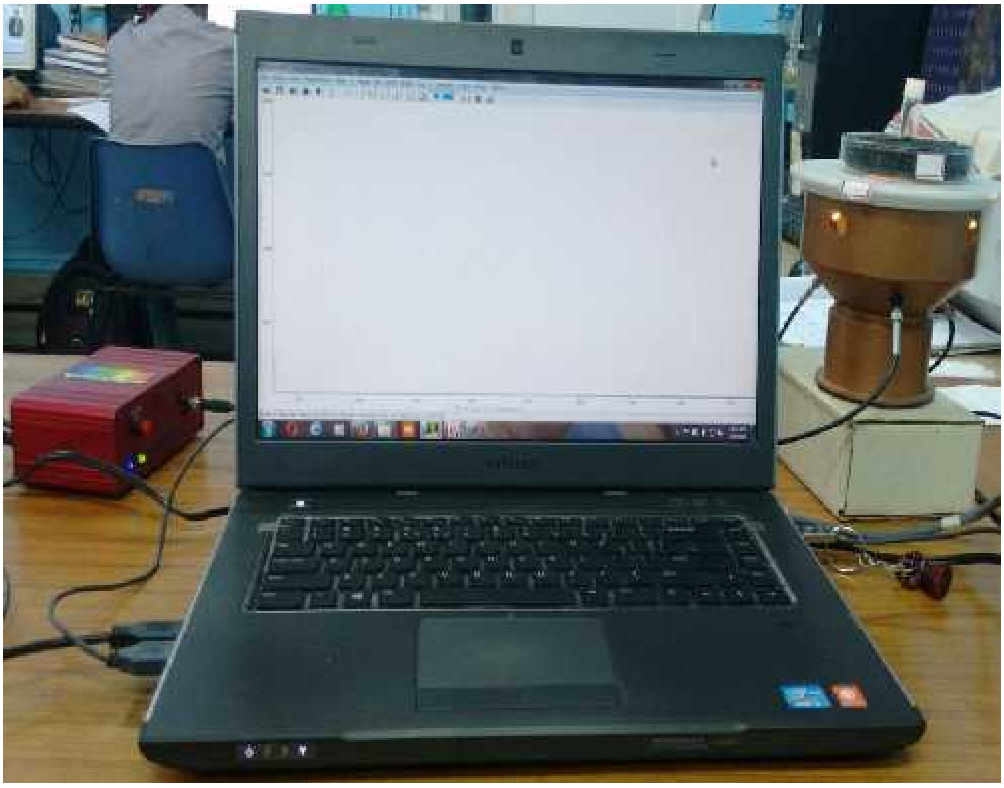


Figure 4.8 NIR instrument setup.

4.2.2. Principal Component Analysis of Data

Principal component analysis is a technique which is used for way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of large dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. Observed data is having large dimension which is 458 different frequencies with 264 sample observation (22 samples×12 observation for each).

4.2.2.1. PCA of Ten Samples.

Before starting PCA, arrangement of all data in excel sheet is done. Next data of each sample is stored in different sheets. So finally, twenty-two sheets (twenty-one tea and one pure caffeine) have been arrenged.

Multivariate analysis compatible software like MATLAB2012 is used as a tool for analysis. First time only nine tea samples and pure caffeine were used to obtain PCA plot. For each sample twelve observations are there and total 120 observations at 458 different wavelengths between 900 to 1700 nm.

For any raw data of an analysis basically two things were obtained such as observation and variable. The samples are treated as observation whereas the 458 different frequencies can be considered as variables. So finally obtained data had dimension 458×120.

$$\begin{array}{ccc} & \overset{\text{Observations}}{\longleftrightarrow} & \\ X = & \begin{bmatrix} s_{11} & \cdots & s_{10,1} \\ \vdots & \ddots & \vdots \\ s_{1.458} & \cdots & s_{10,458} \end{bmatrix} & \updownarrow \quad \text{Variables} \end{array} \quad \text{..............................} \; 4.1$$

Since first few principal components contain most of the variance, so the plot of PC1 and PC2 are obtained as follow:

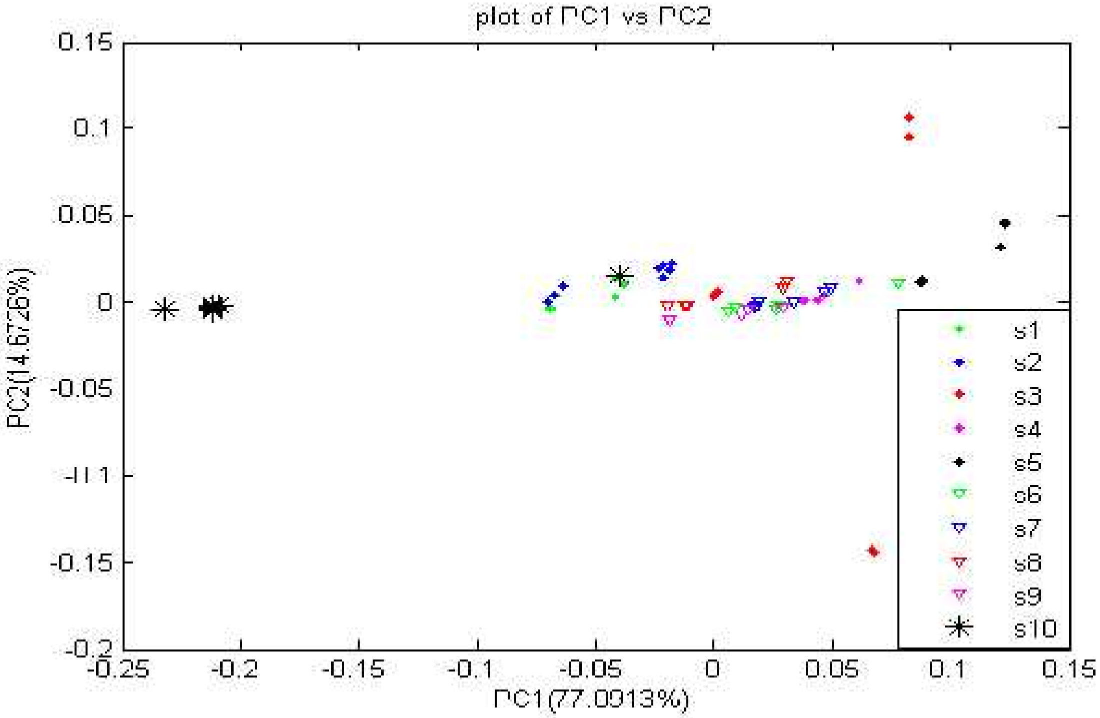


Figure 4.9. PC1 versus PC2 of 10 samples without avg.

In the next analysis average of each group of four observations (taken at same position of sample) has done then PCA applied. The PC1 versus PC2 graph is:

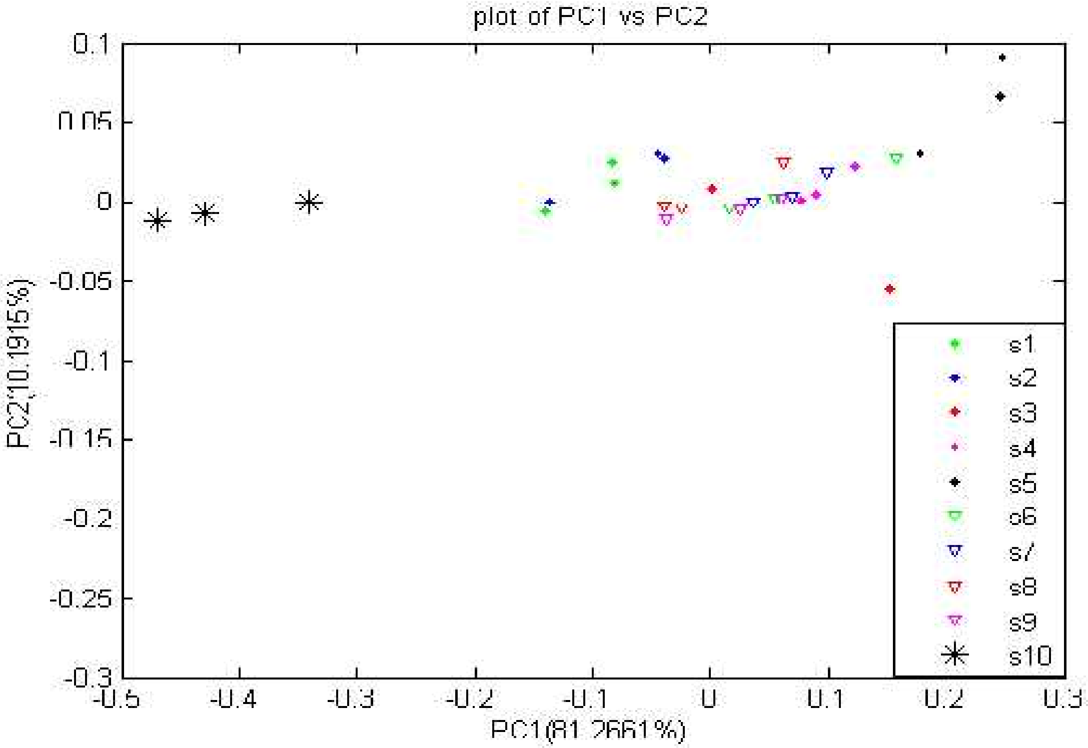


Figure 4.10. PC1 versus PC2 of 10 samples with avg.

4.2.2.2. PCA of All Samples

Same procedure is repeated with twenty one tea samples and one caffeine sample.

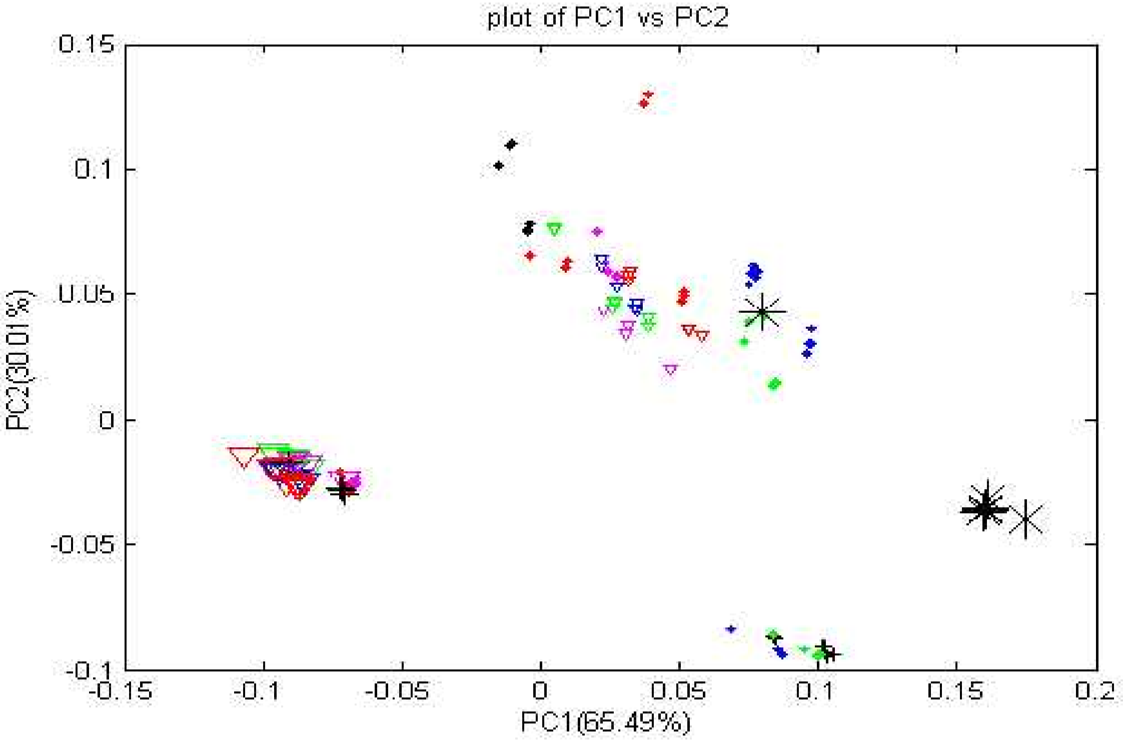


Figure 4.11. PC1 versus PC2 of all samples without averaging.

Again here we can do PCA by taking average of four observations taken at same position so that to make clustering well.

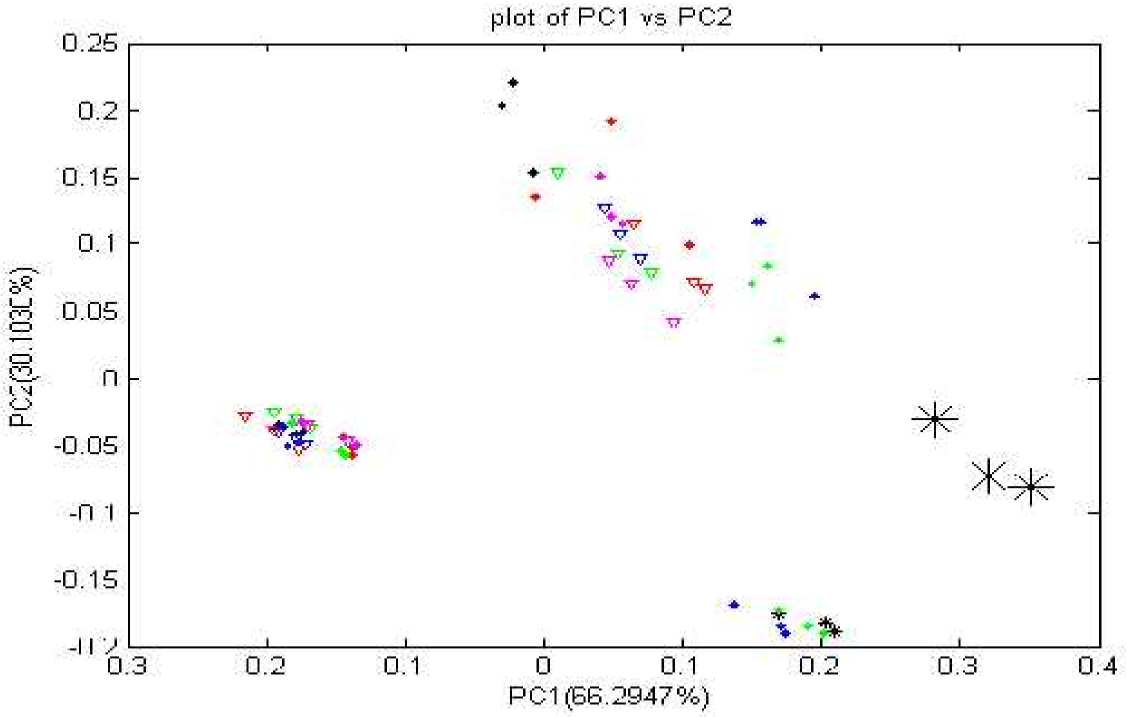


Figure 4.12. PC1 versus PC2 of all samples with averaging.

Table 4.1. Number of samples versus variance of first and second principal component.

(Without averaging the observation)

Table 4.1. (a)

Number of Sample	%Variance explain by first PC	%Variance explain by second PC
10	77.09	14. 67
22	65.49	30.01

(With averaging the observation)

Table 4.1. (b)

Number of Sample	%Variance explain by first PC	%Variance explain by second PC
10	81.26	10.19
22	66.29	30.10

Above table shows that most of the variance is explained by first PC and second PC.

4.3. Principal Component Regression

PCR is a two-step multivariate calibration method: in the first step, a Principal Component Analysis, PCA, of the data matrix X is performed. The measured variables (e.g., absorbances at different wavelengths) are converted into new ones (scores on latent variables). This is followed by a multiple linear regression step, MLR, between the scores obtained in the PCA step and the characteristic y to be modelled.

$$X=\begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad \text{................................ 4.2}$$

where X_1= [X_{11} X_{12}...X_{1P}] is the row vector containing the absorbance measured at p wavelengths (the spectrum) for the first sample, X_2 is the row vector containing the spectrum for the second sample and so on.. For each sample twelve observations at 458 different wavenumber (between 900 to 1700 nm) have been obtained. It is needed to construct a model which gives best prediction without overfitting.

Basic regression model equation

$$y = f(t) = b_0 + \sum_{i=1}^{a} b_i t_i \qquad \text{...} \; 4.3$$

In PCR the scores obtained from PCA is used to construct the regression model, t_i represent the scores where b_i is coefficients.

4.3.1. PCR Steps

4.3.1.1. Averaging the Raw Data and Standardization

Twenty one samples were collected, for each sample twelve data, so before doing PCA averages of four group of observations taken at the same position, it gives three observation for each sample so again averaging is done of these three observations, finally for each sample one observation obtained.

Standardization has been done by subtracting each element to mean of each column and then dividing by its standard deviation or number of rows. Finally a matrix of 458 rows and 21 columns (458×21) is constructed.

4.3.1.2. Principal Component Analysis

Principal component analysis is applied to above matrix. Since different frequencies were assumed as variable and samples as observation so for 458 different variables, 458 principal components can be possible. The scores obtained are used to construct the regression model.

4.3.1.3. Leave One Out Cross-Validation (LOOCV)

While making a model it is necessary to take care of flexibility of the model, it is needed to construct a generalized model which has capability of perform prediction for unknown data set with an adequate accuracy. Cross validation is a technique which is used to construct an optimized model.

Leave One Out Cross-Validation is a technique where each sample data is used to construct training set. All the data were used while making a training set except one and the left data is nothing but the test data set. PCR method is used to construct a model and each time test data set is predicted by trained data set model. Out of twenty one samples for each prediction, training set contains twenty data samples whereas remaining one is test data. Since each model is constructed through principal components, the component (i.e. score) and while making the model these score is used to obtain β value. Using these β values we can predict the value of unknown sample.

After PCA, 458 principal components were obtained, where most of the variance contained by first few PCs. For twenty one samples total number of possible model is 231, it can be explained by as follow:

1:1, 1:2, 1:3, ... 1:21, total 21 combinations

2:2, 2:3, 2:4, ... 2:21, total 20 combinations

...

...

...

20:20 20:21, two combinations

21:21, one combination

Number of possible principal component combinations is 21+20+19...+1= 231. So we need to perform 231 LOO-CV. These combinations are used to obtained β and then these β are used to predict unknown sample prediction, and on the basis of model performance we select the model.

4.3.1.4. Optimisation and Sample Selection

Out of these 231 models only that model is being selected which gives best prediction of the samples, it was found that combination of component 14:15 gives fifteen sample prediction with maximum accuracy (more than 80%), whereas the first few PC combination gives the prediction around 13 samples with accuracy greater than 80%. The root mean square value of first few PC model are as follow:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{cv})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \quad \text{................................ 4.4}$$

where y_i is true value and y_{cv} Is cross-validation predicted value.

PC combination to construct the model	RMSECV
1:1	0.7040
1:2	0.7247
1:3	0.7237
14:15	10.5712

Table 4.2. RMSECV value of selected model.

Since the RMSECV value of first few PC combinations are less as compare to later combinations but the prediction accuracy is not so good as compare to 14:15 PC combination, So it was concluded that those samples having high accuracy at will be selected to make a new model, the samples at the last model (14:15 PC combination) are used to construct a new model, and the sample data of these sample were extracted and kept save for further analysis.

4.3.1.5. Model Formation

The selected samples were taken and used to make new model, these samples can be used to make training data set. All possible models have been constructed and twenty-one samples are being tested. The total number of possible model for fifteen samples is 15+14+13+....+1= 120. The RMSEP value for all model have been calculated the one which have lowest RMSEP with best accuracy was selected among all model.

$$RMSEP = \sqrt{MSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \quad \text{.................................... 4.5}$$

where y_i is true value and $\hat{y}_i$ is predicted value.

Table 4.3. **Number of component used, the RMSEP, and number of sample prediction.**

Number of PC used to form the model	Root Mean Square Error Prediction (RMSEP)	Number of sample predicted (above 80%)
1:1	0.684	14
1:2	0.808	11
1:3	0.707	15
1:4	0.704	15
1:5	0.766	14
1:6	1.213	11
1:7	1.085	12
1:8	1.212	13
1:9	1.250	12
1:12	**1.400**	**16**
1:13	**1.632**	**15**
2:2	0.760	12
2:3	**0.686**	**15**
2:4	**0.684**	**15**
2:5	0.756	13
2:6	1.137	12
2:7	1.0216	13
2:8	1.148	13
2:9	1.170	13
2:12	1.410	14
2:13	1.636	14
3:5	0.740	14
3:6	1.063	14
3:7	0.995	15
3:8	1.132	14
4:4	0.663	14
6:12	1.497	14
7:7	0.625	14
7:8	0.622	14
7:9	0.755	14
8:8	0.662	14
8:9	0.847	14
9:9	0.875	14
13:13	0.806	14

From above table it is clear that model constructed by the components number one to twelve (1:12) can have highest prediction among all with RMSEP value 1.40043, whereas the other component like 1:13, 2:3 and 2:4 component are also have good predictions with prediction fifteen samples (accuracy greater than 80%).

Out of twenty one data sample fifteen samples has used for making the model and prediction of all samples (fifteen used to make training set plus six unknown data set) were obtained. Sixteen predictions with good accuracy have been observed. Out of sixteen twelve sample is having more than 95% accuracy, two sample with accuracy in between 90-95%, and remaining two have accuracy around 82%. Out of remaining five samples four samples have accuracy around 40-60 % which is not acceptable.

The final predicted sample percentage, values and true value table and plot shown below:

Table 4.4. Predicted value, true value and percentage.

True value	Predicted value	Percentage accuracy(%)
2.6	3.862	51.45
3.92	**4.04**	**96.83**
3.78	2.283	60.40
2.16	7.458	-145.29
3.3	**3.419**	**96.38**
3.49	**3.608**	**96.60**
4	**4.121**	**96.97**
4.7	**4.764**	**98.64**
3.65	**3.788**	**96.19**
3.75	1.640	43.75
3.03	2.485	82.02
3.76	1.735	46.17
2.72	**2.849**	**95.25**
3.7	**3.738**	**98.96**
4.01	**4.177**	**95.83**
3.88	**3.962**	**97.87**
2.58	3.020	82.93
2.4	2.223	92.64
2.69	2.936	90.83
2.19	4.172	99.57
3.73	3.852	96.72

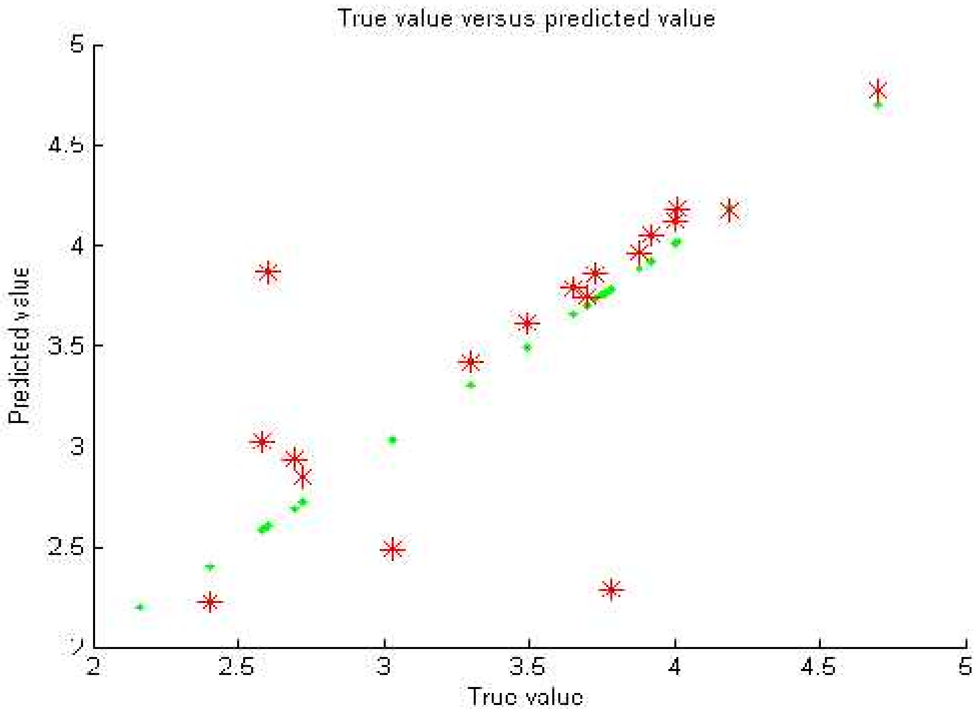


Figure 4.13 Plot of True Versus Predicted value.

4.3.2. Conclusion

In the above figure green dots '.' indicated the over-fitted values and the red '*' indicates the generalised model. This generalised PCR model can be used to predict caffeine content in any tea sample with a good accuracy (more than 80%). The downside of this model is that it cannot be used to predict every sample because only twenty one samples have been used in our analysis if more number of samples were taken the efficiency of model will increase.

Since PCR model uses the Principal component of raw (data which is X), so it provides information regarding X only. To increase model accuracy some other prediction technique such as Partial Least Square method have been used where Principal component of X as well as principal component of Y(dependant variable) both will be used to form the model. The model contains information of X (independent variable) as well as Y (dependent variable).

4.4. Partial Least Square Regression

The general regression method involve the multivariate X (say n number of in dependant variable), and multivariate Y (say m number of dependant variable). It is needed to construct a model using both independent and dependent variable and this model is used to predict the unknown samples without the need of Y.

In order to understand PLS model, we can think of an example of doing PCA not only on X matrix as well as Y matrix.

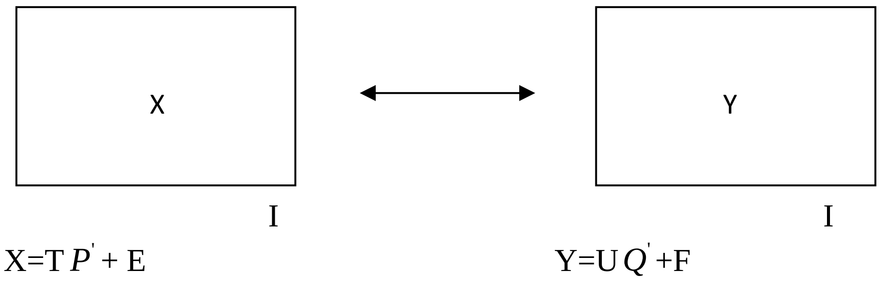


Figure 4.14. PCA on X and Y.

Where T and P are the score and loadings of X respectively, and U and Q is the score and loadings of Y respectively. Once we can predict U matrix we use that U matrix to predict Y value by multiplying with Y-loadings (Q).

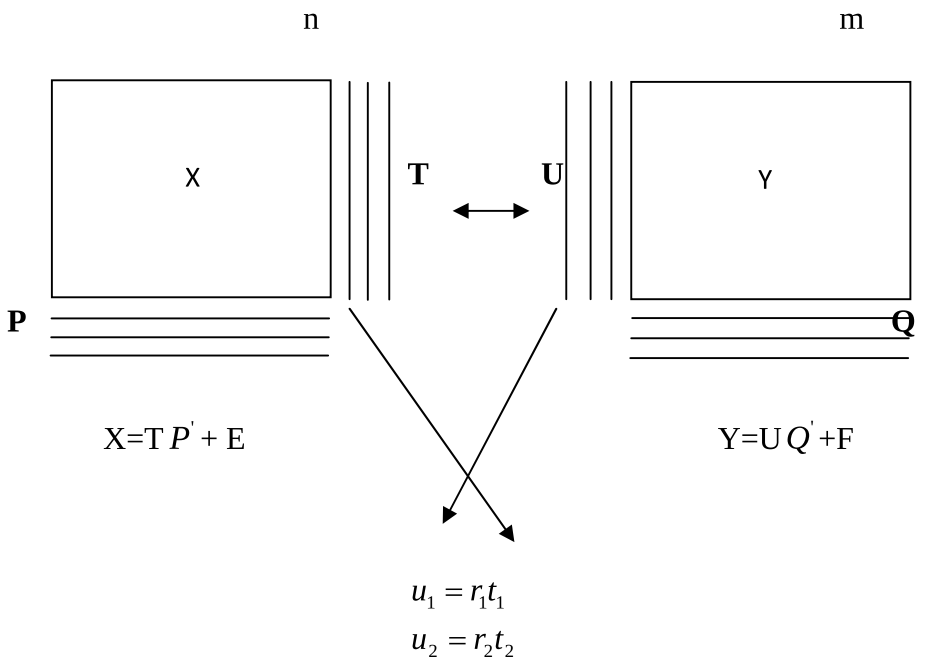


Figure 4.15. Correlation between scores of X and Y.

The loadings of X and Y are rotated from a PCA solution, and it is what exactly PLS does. As the loading changes the scores of X and Y changes correspondingly and correlation between scores of X and Y increase.

$u_1,\ u_2$ are the scores of Y. t_1, t_2 is the scores of X, $r_1,\ r_2$ correlation coefficient. Once correlation between scorers of X and Y increases, U matrix can be predicted up to a good accuracy and by using U matrix w2 can predict Y. This is the fundamental of PLS regression technique.

4.4.1. Steps of PLS Regression Model

4.4.1.1. Averaging the Raw Data and Standardization

Out of twenty one samples and for each sample twelve data were obtained, average of each four group of observation (taken at the same position) has done, three observations for each sample are left, again averaging is done of these three observations, finally for each sample one observation is left.

Standardization has been done by subtracting each element to mean of each column and then dividing by its standard deviation or number of rows. Finally a matrix of 458 rows and 21 columns (458×21) is obtained.

4.4.1.2. Leave One-Out Cross-Validation (LOOCV)

While making a model it is necessary to take care of flexibility of the model, it is needed to construct a generalized model which has capability of perform prediction for unknown data set with an adequate accuracy. Cross validation is a technique which is used to construct an optimized model.

Leave One-Out Cross-Validation is a technique where each sample data is used to construct training set. While making a training set, all sample data except one is used and the left data is nothing but the test data set. A model using PLS regression method is obtained and each time test data set was predicted using trained data set model. For twenty one samples, training set contains twenty samples whereas remaining one is for test. For a training set containing twenty sample nineteen PLS component can be possible (PLS component= number of sample-1), these PLS component are also called latent vector and by selecting appropriate PLS component we can obtain an optimized model.

LOO-CV has been applied to select the appropriate model and filtering those samples which can give best prediction.

4.4.1.3. Optimisation and Model Selection

Training set contains twenty sample so nineteen component is possible for each model. LOO-CV technique has been applied for all nineteen possible models. In first model only first component is used whereas in second first two PLS components have been used and so on.

Out of these nineteen models I have calculated RMSECV values by using below formulae:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{cv})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \quad \text{......................................4.6}$$

where y_i is true value and y_{cv} Is cross-validation predicted value.

Table 4.5. PLS component, RMSECV value, number of sample predicted.

Number of PLS component	RMSECV	Number of sample predicted with accuracy greater than 80%
1	0.733	12
2	0.781	11
3	0.788	11
4	0.865	11
5	0.908	11
6	**0.886**	**13**
7	1.479	11
8	1.452	10
9	1.868	09
10	1.284	07
11	3.276	10
12	3.276	9
13	11.490	10
14	15.028	9
15	14.958	10
16	20.452	9
17	22.251	9
18	21.425	9
19	21.466	9

In above table, it can be seen that at sixth PLS component number LOO-CV gives maximum number of sample prediction with RMSECV value 0.8856. Those thirteen samples will be selected for making new model.

4.4.1.4. Model Formation

Using extracted thirteen samples, Training set was constructed for the PLS model. For all possible models (eleven in this case) RMSEP were calculated.

Trained models will be used to predict all twenty one samples (thirteen used for trained as well as the remaining eight samples). The RMSEP if each model can be calculated using formulae given below:

$$RMSEP = \sqrt{MSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \qquad \text{..4.7}$$

The table below contains RMSEP values of each model.

Table 4.6. Number of PLS component, RMSEP, Number of predicted sample.

Number of PLS component	RMSEP	Number of sample predicted
1	0.657	13
2	1.631	11
3	49.750	1
4	1.827	7
5	3.329	8
6	1.3123	9
7	1.008	8
8	0.7846	15
9	**0.552**	**17**
10	1.036	2
11	0.9778	5
12	0.5050	13

From above table it can be seen that, nine PLS component is having best prediction about seventeen samples with RMSEP of 0.552. The RMSEP of twelve components is less than o.552 but it can predict only thirteen samples with accuracy above 80%. So the model with nine PLS component will be the best predictor among all. The final predicted sample percentage, values and true value table and plot shown below:

Table 4.7. Predicted value, true value and percentage.

True value	Predicted value	Percentage accuracy (%)
2.6	3.713	98.272
3.92	**3.701**	**98.598**
3.78	3.727	97.891
2.16	3.489	93.030
3.3	**3.303**	**88.086**
3.49	**3.4759**	**92.688**
4	**3.390**	**90.406**
4.7	**3.388**	**90.371**
3.65	**3.445**	**91.888**
3.75	3.581	95.502
3.03	3.527	94.057
3.76	3.257	86.860
2.72	**3.033**	**80.894**
3.7	**2.949**	**78.648**
4.01	**2.827**	**75.376**
3.88	**2.708**	**72.223**
2.58	2.461	65.633
2.4	3.230	86.129
2.69	3.324	88.656
2.19	3.405	90.815
3.73	3.543	94.479

Out of twenty one predictions, twelve predictions above 90% and five between 80 to 90%, three between 70 to 80 % and one lowest percentage about 65% were obtained.

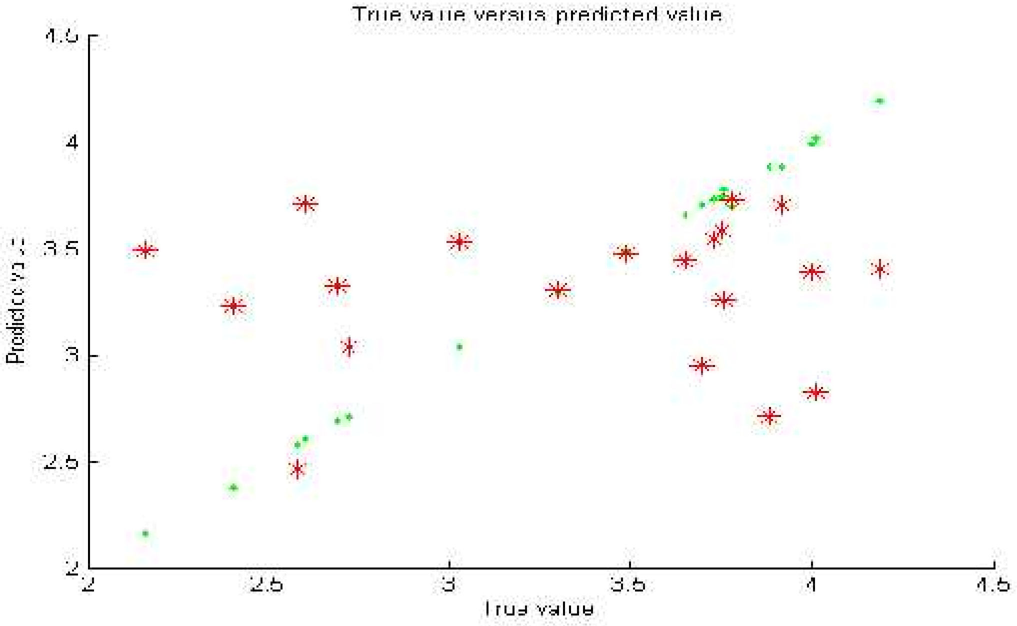


Figure 4.16. True Value versus Predicted value plot.

4.4.2. Conclusion

PLS regression model have predicted caffeine content of seventeen samples with accuracy greater than 80%, so it can be concluded that this model is far better than PCR model because its efficiency is much better as compare to it. In the above model eight samples are completely unknown for the model where prediction of four samples ware very well (greater than 90%) and regarding others two is having accuracy between 75 to 80 %.

It can be concluded from above discussion this model is marginally accepted for the prediction of caffeine content in tea sample

Referances

[1] http://www.stellarnet.us/spectrometers/dwarf-star-miniature-nir-spectrometer/

[2] http://www.stellarnet.us/spectrometers/dwarf-star-miniature-nir-spectrometer/

[3] http://www.stellarnet.us/spectrometers/dwarf-star-miniature-nir-spectrometer/

[4] http://www.stellarnet.us/light-sources/#SL1TungstenHalogen.

[5] http://www.stellarnet.us/light-sources/#SL1TungstenHalogen.

Chapter 5

Conclusions and Future Scopes

5.1. Model accuracy

5.1.1. Principal Component Analysis (PCA)

As twenty-one tea sample containing caffeine and one sample containing pure caffeine have been arranged. During first analysis nine tea samples and one caffeine sample, total ten samples were taken and PCA has been done. From the PCA plot (figure 4.1) it can be seen that the clustering of pure caffeine is far away from other tea sample, it proves that caffeine concentration in pure caffeine sample is higher than other tea sample. The segregation of cluster of sample is also very clear and clustering of sample data proves the better repeatability of instrument. The variance explained by first principal component is very high about 78.45% on the other hand variance explain by second principal component is about 11.54%.

Similarly same procedure is done for all twenty-one samples and segregation of all samples in plot is found well with first PC variance equals 66.49% and for second PC 30.10%.

5.1.2. Principal Component Regression (PCR) Model

The principal component obtained by twenty-one model is used to construct the model called PCR model. All possible models were constructed and leave one out cross validation technique has been used to identify those sample which can be used to construct final model, finally fifteen samples was selected to construct training set and again all predicted all sample with extracted optimized model with principal component one to twelve RMSEP 1.4.

Out of twenty one data set fifteen samples have been used for making the model and prediction all samples (fifteen used to make training set plus six unknown data set) and sixteen predictions with good accuracy was obtained. Out of sixteen twelve sample is having more than 95% accuracy, two sample with accuracy in between 90-95%, and remaining two have accuracy around 82%. Out of remaining five samples four samples is having accuracy around 40-60 % which is not acceptable.

Efficiency of above model is not much good as it was expected so another PLS regression model have been constructed to get better accuracy.

5.1.3. Partial Least Square (PLS) Regression Model

Here a PLS regression model have been constructed by using nine PLS component with RMSEP of 0.552. Out of twenty one, seventeen sample were predicted well. It can be concluded from above result that this model is far better than PCR model because its efficiency is much better as compare to PCR model. In the above model eight samples are completely unknown for the model where the prediction of four samples are very well and regarding others two is having accuracy between 75 to 80 %.

Out of twenty one predictions, twelve predictions have accuracy above 90% and five between 80 to 90%, three between 70 to 80 % and one lowest percentage about 65%.

5.2. Future scopes of NIR Spectroscopy and MVA Regression Model

Only twenty-one samples have been used for prediction model in the analysis. If the number of sample to construct the model will increase the accuracy will also increase up to an acceptable level. The most noticeable thing about this analysis is that if an efficient model can be constructed then needless to go for time consuming technique to quantify chemicals in sample. Others prediction models can be explored with NIR spectroscopy. Optimization techniques for the selection of wave length for particular application can be explored.