

**Sensing and signal processing for non-invasive blood glucose monitoring**

**By**

**Krishna Chaitanya Patchava**

Thesis Submitted for the Degree of PhD

Department of Electronic & Electrical Engineering

December 2017

## **Abstract**

Remote monitoring is required in several applications and generically, a remote monitoring architecture can be separated into 3 distinct but inter-related layers namely: sensing, signal processing and communications. However, this study focuses on the sensing and signal processing aspects in healthcare. In particular, the research is to investigate sensing and signal processing techniques for a non-invasive approach to type-1 diabetic patients monitoring. Diabetes mellitus is a long-lasting disease and the number of people with diabetes is increasing rapidly worldwide. Managing this disease requires continuous monitoring of blood glucose levels. It is stipulated that avoiding the traditional finger prick method could help improve the adherence and overall management.

The present research is concerned with using Fourier transform near infrared spectrometer for the non-invasive measurement of the blood glucose levels. This research has focused on the signal processing and data analysis aspects where a near infrared spectrophotometer has been employed for the sensing to collect practical representative test data. In the signal processing aspects, most of the researchers to date have tended to employ linear regression techniques are the Partial Least Squares Regression (PLSR) and the Principal Component Regression (PCR) based methods and their variants. However, these methods have limitations in practice and have not been translated into a clinical tool. In this project, we target to overcome the current drawbacks of these techniques and in particular their inability to detect the components with low variance by investigating the potential of certain non-linear regression techniques; one of the promising techniques proposed in this research is based on combining a Local Linear Embedded Regression (LLER) with pre-processing. The coupling of bandpass filtering with the novel LLER has been shown to achieve better prediction results than existing methods. A novel regression model called improved support vector regression coupled with Fourier self-deconvolution is also proposed and compared with the linear calibration models under the same conditions. The other proposed model is Partial Least Squares Regression coupled with Frequency self-deconvoluted ReliefF (FSDR-PLSR) which is based on the variance adjustment according to the importance of the features. Finally, two novel pre-processing methods are introduced in this work; i) pre-processing based on Hilbert haung transformation and ii) pre-treatment technique based on coupling digital bandpass filtering with scatter correction techniques.

## **Dedication**

The research work presented in this dissertation is dedicated to my late grandmother Mrs. Patchava Ratnamma whose unconditional love and sacrifice made this journey possible.

To my parents Mr. Patchava Mohan Rao and Mrs. Malleswari; their love, patience and support made this long journey possible.

To my wife: Mrs. Patchava Nandini, for her love, support and encouragement during every phase of my research work.

To my darling daughter: Miss Patchava Moksha Ranhitha, who entered into my life during final phase of my research study.

Finally, this work is dedicated to millions of people who are suffering from diabetes worldwide.

## **Acknowledgements**

Firstly, I would like to express sincere thanks to my supervisor, Dr. Mohammed Benaissa, senior lecturer, EEE Department for agreeing to be my supervisor and for his patience, support and providing kind guidance throughout my research study at the university of Sheffield. Without his feedback, support and guidance, it is not possible to finish this work.

I am very thankful to Dr. Hatim Behairy for his support, valuable feedback and guidance during the progress meetings to enhance my knowledge in the specified subject.

I would like to thank Steve Marsden, EEE technician for his quick response and help to provide technical support at his earliest convenience.

I would like to thank Dianne webster, EEE technician for assisting with fire training and out of hours working which helps to work during outside the normal hours.

I am very thankful to Mrs. Hilary J Levesley for her quick response for all the queries and friendly support during every phase of my research work at the university of Sheffield.

I would like to thank my fellow lab mates; Bilal Malik, Osamah and Zia khan for their friendly support and encouragement.

Contents

[**Abstract** ii](#_Toc505600353)

[**Dedication** iii](#_Toc505600354)

[**Acknowledgements** iv](#_Toc505600355)

[**List of Figures** vii](#_Toc505600356)

[**List of Tables** ix](#_Toc505600357)

[**List of Acronyms** x](#_Toc505600358)

[**Chapter 1: Introduction** 1](#_Toc505600359)

[**1.1** **Preface:** 1](#_Toc505600360)

[**1.2** **Thesis aims and scope** 4](#_Toc505600361)

[**1.3** **Key contributions** 5](#_Toc505600362)

[**1.4** **Thesis structure** 7](#_Toc505600363)

[**1.5** **Publications** 8](#_Toc505600364)

[**Chapter 2: Background** 10](#_Toc505600365)

[**2.1 FTIR Spectrometer:** 10](#_Toc505600366)

[**2.3 Non-invasive blood glucose monitoring:** 16](#_Toc505600367)

[**2.4** **Multivariate calibration models:** 19](#_Toc505600368)

[2.4.2 Principal component regression (PCR): 22](#_Toc505600369)

[2.4.3 Partial Least Squares Regression (PLSR)**:** 24](#_Toc505600370)

[**2.5** pre-processing methods: 29](#_Toc505600371)

[**2.5.1** Mean Centering: 29](#_Toc505600372)

[2.5.2 First derivative as a pre-processing technique: 30](#_Toc505600373)

[2.5.3 Standard normal variate pre-processing technique: 30](#_Toc505600374)

[2.5.4 Fourier Self Deconvolution: 31](#_Toc505600375)

[2.5.5 Digital Bandpass filtering: 32](#_Toc505600376)

[**4.1 INTRODUCTION** 38](#_Toc505600377)

[**4.1 Local Linear Embedding (LLE) Dimensionality Reduction Algorithm:** 40](#_Toc505600378)

[**3.2.1.** **Local Linear Embedding Regression (LLER):** 44](#_Toc505600379)

[**4.4 Discussion of Experimental Results and Comparisons:** 53](#_Toc505600380)

[57](#_Toc505600381)

[**5.1 Introduction:** 60](#_Toc505600382)

[**5.2** Support vector regression (SVR): 61](#_Toc505600383)

[**5.2.1** **Improved support vector regression:** 62](#_Toc505600384)

[**5.2.2** **Fourier Self-Deconvolution** 62](#_Toc505600385)

[**5.2.3** **Experiments:** 63](#_Toc505600386)

[**5.2.4** **Results and Discussions:** 64](#_Toc505600387)

[**Chapter 6: Fourier Self Deconvoluted RReliefF** 70](#_Toc505600388)

[**6.3.1The relevance analysis:** 75](#_Toc505600389)

[**6.3.2 The variance adjustment:** 75](#_Toc505600390)

[**6.4 Fourier Self Deconvolution:** 77](#_Toc505600391)

[**6.5 Fourier Self Deconvoluted RRelifF (FSDR) as a Pre-processing Method:** 78](#_Toc505600392)

[**6.6 Experiments:** 79](#_Toc505600393)

[**6.7 Discussion of experimental results and comparisons:** 79](#_Toc505600394)

[**Chapter 7: NIR spectroscopy using human serum albumin data** 85](#_Toc505600395)

[**7.1 SAVITZKY-Golay coupled with digital bandpass filtering** 85](#_Toc505600396)

[**7.1.1 INTRODUCTION** 85](#_Toc505600397)

[**7.1.2 The RReliefF** 87](#_Toc505600398)

[**7.1.3 The RReliefF processing** 87](#_Toc505600399)

[**7.1.4 The digital bandpass filtering** 89](#_Toc505600400)

[**7.1.5 Savitzky-Golay Filtering** 90](#_Toc505600401)

[**7.1.6 Experiments (Human serum albumin data)** 90](#_Toc505600402)

[**7.1.7 Discussion of experimental results** 90](#_Toc505600403)

[**7.2 Hilbert Haung Transformation pre-processing** 95](#_Toc505600404)

[**7.2.1 Introduction:** 95](#_Toc505600405)

[**7.2.2 Hilbert Haung Transformation (HHT)** 97](#_Toc505600406)

[**7.2.3 Empirical Mode Decomposition (EMD)** 97](#_Toc505600407)

[**7.2.4 Hilbert Spectral Analysis (HSA)** 99](#_Toc505600408)

[**7.2.5 Proposed HHT-PLSR model** 101](#_Toc505600409)

[**5.3.8 Discussion of Experimental Results and Comparisons:** 102](#_Toc505600410)

[**Chapter 8: Conclusions and Future scope** 107](#_Toc505600411)

[**8.1 Summary and Conclusions:** 107](#_Toc505600412)

[**8.2 Future scope** 110](#_Toc505600413)

[**References** 112](#_Toc505600414)

## **List of Figures**

2.1 Schematic diagram of Michelson interferometer used in FTIR spectrometer……………………………………………………………………. (12)

2.2 Schematic diagram of MLR model to predict the concentration of analyte of interest (21)

3.1 Illustration of experimental data preparation for the urea dataset……………………. (35)

3.2Illustration of experimental data preparation for human serum albumin data…………(37)

4.1 Implementation of LLE algorithm …………………………………………………….(43)

4.2 Block diagram of the Gaussian digital bandpass filter................................................ (48)

4.3 Block diagram of the Chebyshev digital bandpass filter............................................ (49)

4.4 Flow chart of parameter optimization for DBPF-LLER model.................................. (52)

4.5 Comparison of the PCR, PLSR, SVR and LLER models without pre-processing...... (54)

4.6 PCR and PLSR with different pre-processing techniques…………………………… (57)

4.7 Comparison of PCR, PLSR, SVR and LLER models with different types of pre-processing techniques…………………………………………………………………… (58)

6.1 Block diagram of the proposed FSDR technique for the PLSR model………………. (78)

6.2 Variance captured plot for the PLSR model………………………………………… (80)

6.3 Comparison of PLSR model with different pre-processing methods………………… (81)

6.4 Clarke Error Grid Analysis plots for PLSR model with different pre-processing methods................................................................................................................................ (83)

7.1.1 variance captured plot for PCR and PLSR models..................................................... (91)

7.1.2 Mean squared prediction error plot for PCR with different pre-processing techniques............................................................................................................................. (92)

7.1.3 Mean squared prediction error plot for PLSR with different pre-processing techniques............................................................................................................................. (94)

5.2.1 3D plot for the collected Human serum spectra.......................................................... (89)

5.2.2: Prediction performance of PCR and PLSR................................................................ (90)

5.2.3 Comparison of PCR and PLSR for different factors................................................... (91)

5.2.4 Comparison of the SVR model with the PCR and the PLSR methods without pre-processing............................................................................................................................. (92)

5.2.5 Comparison of SVR, PLSR and PCR methods using FSD pre-processing method... (93)

7.2.1 The block diagram of the HHT-PLSR model............................................................. (101)

7.2.2 The raw spectra of data1 and the human serum albumin........................................ (102)

7.2.3 The variance captured plot for PLSR model using data1.......................................... (103)

7.2.4 The variance captured plot for PLSR using human serum data................................ (104)

7.2.5 The comparison of PLSR with different pre-processing techniques using data1...... (105)

7.2.6 The comparison of PLSR with different pre-processing techniques using human serum albumin data…………………………………………………………………………… (106)

## **List of Tables**

4.1 The prediction capability of the LLER model for different values of K and d……… (55)

4.2 Comparison of PCR,PLSR, SVR and LLER models ………………………………. (59)

6.1 Comparison of PLSR model with different pre-processing methods……………….. (84)

7.1.1 Summary of results for PCR and PLSR with different pre-processing techniques.... (93)

5.1 Summary of results for PCR, PLSR and SVR........................................................... (69)

## **List of Acronyms**

FTIR Fourier Transform Infrared Spectrometer

EMS Electro Magnetic SpectrumNIR

NIR Near Infrared

PCR Principal Component Regression

PLSR Partial Least Squares Regression

PCA Principal Component Analysis

LLE Local Linear Embedding

LLER Local Linear Embedding Regression

FSD Fourier Self Deconvolution

HHT Hilbert Haung Transformation

IDF International Diabetes Federation

MIR Mid Infrared

MLR Multiple Linear Regression

PLS Partial Least Squares

IR Infrared

ZPD Zero Path Difference

SVM Support Vector Machines

SVR Support Vector Regression

SRM Structural Risk Minimization

ERM Empirical Risk Minimization

RMSEP Root Mean Square Error of Prediction

RMSEC Root Mean Square Error of Calibration

RMSECV Root Mean Square Error of Cross Validation

DBP Digital Bandpass

LPF Low-pass filter

HPF High-pass filter

BRF Band-reject filter

PC Principal Components

LV Latent Variables

RBF Radial Basis Function

SNR Signal to Noise Ratio

FSDR Fourier Self Deconvoluted RReleifF

FWHH Full Width at Half Height

FFT Fast Fourier Transform

IFFT Inverse Fast Fourier Transform

RPLSR Recursive Partial Least Squares Regression

EGA Error Grid Analysis

SNV Standard Normal Variate

ISVR Improved Support Vector Regression

SG Savitzky-Golay

RPLS Recursive weighted Partial Least Squares

PBS Phosphate Buffer Solution

KS Kennard Stone

EMD Empirical Mode Decomposition

HSA Hilbert Spectral Analysis

IMF Intrinsic Mode Function

## **Chapter 1: Introduction**

This chapter briefly describes the key contributions of the thesis in addition to its aims and scope. In this chapter, the concept of non-invasive approach for the blood glucose monitoring and the near infrared spectroscopy has also been presented. The thesis structure is also described in this chapter.

## **Preface:**

Signal processing refers to the techniques or algorithms that deal with the operations on or analysis of signals. Signal processing can be used to enhance the accuracy or quality of the signals, to decompose overlapped signals, to extract specific information from the signals. It is also used to compress the data such that it requires less memory storage units and also used to convert signals from one form of representation to another form so that it is compatible with other applications. Signal processing has been used in a number of fields such as communications, biomedical engineering, data compression, multimedia and image processing. However, signal processing techniques have not been used extensively in the area of Chemometrics. The study of obtaining the information from chemical systems using various multivariate analysis methods is called Chemometrics. Near Infrared (NIR) spectroscopy received much attention in different applications such as biomedical chemistry, environmental analysis, agricultural and food industries after the development of the Fourier Transform Infrared (FTIR) spectrometers. In NIR spectroscopy, the range of Electro Magnetic Spectrum (EMS) starting from 780nm to 2500nm is used to investigate the properties of chemical systems. NIR spectrum can be divided into 3 regions; first overtone region (1540nm to 1820nm), combination region (2000nm to 2500nm) and short wavelength region (700nm to 1330nm). Many biological components have unique response to NIR signals and they have absorption bands in NIR region.

The advantages of NIR spectroscopy include:

1. Low water absorbability of NIR radiation
2. The preparation of samples includes less cost and collection of spectra is simple and the sample path length can be in the order of mm.
3. It can be used to study not only the biological and physical properties but also molecular structure of the components.
4. The spectra collection time is very less with high signal to noise ratio.

Hence NIR spectroscopy has become forefront for many biomedical researchers concerns, in particular with respect to non-invasive blood glucose monitoring [1-6].

As the water absorbability is low and the glucose absorbance band is narrower than the water, particularly at the absorption band centred at 2273nm. Hence most of the recent studies including this study focus on the combination region of the NIR radiation for the non-invasive measurement of the blood glucose levels [7, 8].

Non-invasive sensing is the measurement of concentration of the analyte of interest with no contact between the transduction element and the representative biological fluid.

Diabetes is one of the serious chronic diseases found worldwide and poor control of the disease may cause even death. Diabetes is a metabolic disorder which affects the metabolism of carbohydrates and causes increased sugar levels. Glucose is generated from the breakdown of food and it spread the entire body in the blood, which is treated as the main source of energy. The glucose concentration for healthy people is well regulated with the actions of hormones such as glucagon and insulin. However, the diabetic patients are unable to either produce or utilize the insulin effectively, which results in hypo-glycemia or hyper-glycemia. Hyper-glycemia is the abnormal higher blood sugar level, whereas hypo-glycemia occurs when the blood glucose level falls below the normal range. The complications with the hypo-glycemia include unexpected coma, it may lead to damage of the brain cells or even death. Hyper-glycemia may lead to nerve damage, blindness, heart stroke and kidney failure [9, 10]. Diabetes is classified into mainly 3 types; type 1, type 2 and gestational diabetes. If the sugar level is raised due to pancreas does not produce the required amount of insulin, which is known as type 1 diabetes or the cells do not respond to the insulin produced by the pancreas which is called type 2 diabetes. Type 1 diabetes is also known as juvenile diabetes, which usually occurs in children, young adults or teenagers. Gestational diabetes generally occurs in some women in the later stages of pregnancies and usually disappears after the baby birth [11]. Recent reports indicate that out of diabetic patients, around 10% are type-1 diabetic patients. Type 1 diabetes is also called insulin dependent diabetes mellitus, occurs due to the destruction of beta cells of the pancreas which produces the insulin resulting the body does not produce enough insulin [12]. Type 1 diabetes can occur to people of any age and as there is no cure for the disease, it needs to be well managed for the proper control of the disease. The continuous monitoring of the blood glucose levels is needed for better control of the illness. For the better management of the disease, patient need to measure the blood glucose levels and take insulin several times a day. Traditional method for the measurement of glucose levels is the finger prick method, which is not very convenient to many patients to check their glucose levels continuously. Hence researchers have been trying to come up with the non-invasive glucose monitoring techniques. In particular, our research focuses on the non-invasive glucose measurement for type 1 diabetic patients. NIR Spectroscopy has been recognized as one of the optimistic techniques for non-invasive glucose measurement as it provides a reasonable signal to noise ratio and requires little or no sample preparation. The experimental data is prepared by dissolving urea, triacetin and glucose in a Phosphate Buffer Solution (PBS). To validate the efficacy of the proposed models and to improve the complexity of the dataset, another experimental data is also prepared by dissolving Human Serum Albumin (HSA) and glucose in a PBS.

Signal processing aspects can be applied as a pre-processing technique for developing a regression model. Most of the researchers to date have tended to employ Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) for the determination of glucose concentration from the NIR spectra [13-15].

## **Thesis aims and scope**

The measured NIR spectra are affected by instrumental and environmental variations, noise and chemical components of the sample. Therefore, the main aim of this thesis is to investigate and develop the signal processing algorithms to overcome the effect of unwanted variations in the raw spectra, to improve the quality of spectra and to amplify the information related to the chemical components. The signal processing methods used in this work can be broadly classified into two categories; pre-processing techniques and the multivariate calibration models. However, in this work, the pre-processing methods and the calibration models have been coupled to further improve the prediction performance of the designed models.

The aim of the proposed multivariate methods is to build the calibration models to predict the concentration of analyte of interest from the spectra of a chemical mixture and the aim of the introduced pre-processing techniques is to eliminate the unwanted variations from the spectra such as the high frequency noise components and the baseline variations in order to improve the quality of spectra. These signal processing methods can be used in the quantitative analysis of analyte of interest from NIR spectra.

In this work, the non-invasive detection of blood glucose from an aqueous mixture is investigated from the point of view of signal processing. As the blood components made up of combinations of O-H, C-H or N-H (O – Oxygen, C – Carbon, N – Nitrogen, H – Hydrogen) have the same features in the combination region and the resultant absorbance spectra might be overlapped. The main components of blood that have significant absorbability in NIR region are urea, haemoglobin, water, protein, cholesterol and lactate. Hence in this study, triglycerides, urea, water and glucose were used to generate mixtures and the individual concentration values are selected in a way that spans their physiological range in blood. The proposed models have been compared with the most commonly used calibration models the PCR and the PLSR models under the same conditions using the experimental data. To validate the efficacy of the proposed models another dataset was also prepared by mixing HSA and glucose in a PBS and the proposed models were also tested using this dataset.

## **Key contributions**

The signal processing methods developed in this research can be used in a wide range of biomedical applications. These methods can also be used in the quantitative and qualitative analysis of NIR spectra. However, there is a need for the efficient pre-processing methods and the multivariate methods in NIR spectroscopy. Hence the following novel calibration models and the pre-processing methods have been proposed for the quantitative analysis of NIR spectroscopy.

1. PCR [16, 17] is the most commonly used calibration model in Chemometrics for the quantitative analysis of NIR spectra. In PCR, the principal component analysis (PCA) is used to decompose the absorbance matrix *A* into loading and scores matrices. Then the target element is regressed against the scores obtained instead of the input raw spectra. To minimize the complexity of the calibration model, the PCA method removes the features that have low variance and retain the factors that represent most of the variation of *A*. If the analyte of interest shows the lowest variance in the spectra, then it may be removed during the decomposition process. Thus, this decomposition process may degrade the prediction performance of the calibration model.

To overcome the drawbacks of PCR method, a novel non-linear calibration model called Local Linear Embedding Regression (LLER) is proposed to detect the components that produce even low variation in the spectra. The LLER model is based on local linear embedding (LLE) dimensionality reduction method. In the LLER method, the scores obtained from the LLE method are regressed against the analyte of interest to compute the regression coefficients. (chapter 4).

1. A novel calibration model based on combining the Fourier self-deconvolution (FSD) with the Improved Support Vector Regression (ISVR) is proposed and the model has also been compared with the PCR and the PLSR models under the same conditions and the results show that the ISVR coupled with the FSD outperforms the linear models under the same conditions (chapter 5)
2. FSD is an effective pre-processing method to eliminate the high frequency noise and the baseline variations from the raw spectra, however irrelevant features with high variance may not be removed during this method. Hence a novel pre-processing method based on combining the Fourier self-deconvolution with the feature weighting method (RReliefF) is proposed to improve the prediction performance of the PLSR model. (chapter 6)
3. Two novel pre-processing models are also introduced in this dissertation; i) Since the spectra are affected by scale shifting, noise and baseline variations that make the spectra non-stationary. The Hilbert Haung Transformation has proven successful for the analysis of non-stationary, natural and non-linear signals [18]. Hence a novel calibration model based on combining the Hilbert Haung Transformation (HHT) pre-processing with the PLSR is introduced to eliminate both the baseline variations and the high frequency noise from the spectra and the results show the considerable improvements. ii) A novel pre-processing method based on combining the Savitzky Golay (SG) smoothing and bandpass filtering is proposed to improve the prediction performance of the linear models.

## **Thesis structure**

In Chapter 2, an introduction to non-invasive glucose monitoring and the methods used for non-invasive blood glucose monitoring are presented. This chapter also includes the background on basic calibration models the PCR and the PLSR models in addition to the existing pre-processing methods in Chemometrics are also discussed in detail.

In chapter 3, the process of preparing the samples for the two experimental data sets used in this study are discussed in detail.

The PCR and the PLSR models are widely used in Chemometrics for the quantitative analysis of glucose from near infrared spectra. However, these models perform better when the analyte of interest contributes highest variation to the spectra. If the analyte of interest contributes less variation to the raw spectra, then the information relating to the analyte of interest may be removed during the decomposition process, which will Hence Chapter 3 introduces a novel calibration model called Local Linear Embedding regression (LLER) which can perform better even the analyte of interest contributes to less variance.

In chapter 4, a novel calibration model called Improved support vector regression coupled with Fourier self deconvolution is introduced.

The Chapter 5 introduces a novel calibration called improved support vector regression coupled with fourier self deconvolution for the quantitative analysis of glucose in NIR spectra.

In chapter 6, a novel pre-processing method called Fourier Self deconvoluted RReliefF (FSDR), which is based on combining the FSD with feature weighting algorithm RReliefF has been introduced.

The Chapter 7 introduces two novel pre-processing methods; i) pre-processing method based on Hilbert Haung transformation is presented in this chapter and ii) A novel pre-processing method based on combining the scatter correction techniques with bandpass filtering is also introduced in this chapter.

Chapter 8 concludes the thesis. The findings are summarised here and the recommendations for future work are identified.

## **Publications**

1**.** Patchava, Krishna Chaitanya, Mohammed Benaissa, Bilal Malik, and HatimBehairy. "Local linear embedded regression in the quantitative analysis of glucose in near infrared spectra." *Analytical Methods* 7, no. 4 (2015): 1484-1492.

2. Patchava, Krishna Chaitanya, Mohammed Benaissa, and HatimBehairy. "Improving the prediction performance of PLSR using RReliefF and FSD for the quantitative analysis of glucose in Near Infrared spectra." *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015.

3. Malik, Bilal, Krishna Chaitanya, and Mohammed Benaissa. "Support vector regression with digital band pass filtering for the quantitative analysis of near‐infrared spectra." *Journal of Chemometrics* 28, no. 2 (2014): 116-122.

4. Patchava, K. C., Alrezj, O., Benaissa, M., & Behairy, H. (2016, August). Savitzky-golay coupled with digital bandpass filtering as a pre-processing technique in the quantitative analysis of glucose from near infrared spectra. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the* (pp. 6210-6213).

5. Krishna chaitanya patchava, Mohammed Benaissa and Hatim Behairy. “Partial Least Squares Regression coupled with Hilbert HaungTransformation pre-processing for the quantitative analysis of glucose in near infrared spectra”, submitted to IEEE international workshop on signal processing symposium (SIPS 2017), October 3-5, 2017, Lorient, France.

6. Krishna chaitanya patchava, Mohammed Benaissa and Hatim Behairy. “Improved support vector regression coupled with Fourier self-deconvolution in the quantitative analysis of glucose in near infrared spectra”, to be submitted to the Journal of chemometrics and Intelligent laboratory systems.

7. Osamah, Krishna chaitanya patchava and Mohammed Benaissa. “Scatter correction methods coupled with bandpass filtering as a pre-processing technique in the quantitative analysis of glucose in near infrared spectra”, Accepted for the presentation at IEEE Engineering in Medicine and Biology Society (EMBS’17), July 11-15,2017, S.Korea.

8. Osamah, Krishna Chaitanya patchava, Mohammed Benaissa, S. Alshebeili, “'Pre-processing to Enhance the Quantitative Analysis of Glucose from NIR and MIR Spectra' Accepted for the presentation at ‘The joint conferences: 'European Medical and Biological Engineering' and 'Nordic-Baltic Biomedical Engineering (EMBEC-NBC 2017), 11-15 June 2017, Tampere, Finland.

# **Chapter 2: Background**

This chapter describes the concept of non-invasive blood glucose monitoring and the Fourier Transform Infrared (FTIR) spectrometer and presents the existing calibration and pre-processing methods for the non-invasive monitoring of the blood glucose levels. The traditional linear regression models such as the PCR, the PLSR and the non-linear regression method the Support Vector Regression (SVR) along with its limitations are also discussed.

## **2.1 FTIR Spectrometer:**

Infrared (IR) spectroscopy is a widely used technique for the quantitative analysis of molecular species, in which IR radiation is passed through a sample. Some of the radiation is transmitted through the sample where as some IR radiation is absorbed by the sample. The resulting spectrum represents the molecular finger print of the sample as it constitutes the molecular absorption and transmission. As no two molecular structures develop the same IR spectrum, this makes the IR spectroscopy to be effective method in the quantitative analysis of a single analyte from the spectral mixture.

The IR region spans from 12,800 to 10cm-1 or in wavelength 780nm to 1000µm. The IR region is further categorized into mid, near and far infrared based on the energy and type of vibrational transitions. The Near Infrared (NIR) region covers from 780nm to 2500nm.

FTIR spectroscopy is a preferred method of spectroscopy and its advantages are summarized below.

The following are the advantages of FTIR spectroscopy:

1. It has greater optical throughput and provides a precise measurement
2. Non-destructive technique and mechanically simple with only one moving object.
3. It can measure the spectrum at every second and the collection time for the spectra is very less.

An optical component termed as interferometer is the key part of FTIR spectrometer. In commercial FTIR spectrometers, Michelson interferometer is the most widely used among all the different types of interferometers [19-22].

Schematic diagram of Michelson interferometer is shown in Figure 2.1.

The radiation from an IR source is divided into two equal parts using a beam splitter as shown in Figure 2.1. One part of the beam is directed towards the movable mirror while the other part transmits towards the stationary mirror. The two beams from the two mirrors are reflected and recombined at the beam splitter. The recombined beam is again divided into 2 parts and one beam transmits through the sample and reaches the detector and the other beam is reflected to the IR source as shown in Figure 2.1 [23]. The beam splitter plays vital role in separation of the incident light. Due to difference in travel distances, a destructive or constructive interference occurs between the reflected waves from the two mirrors.

The optical path difference caused due to the relative mirror positions is called retardation (δ). Obtained signal at the output of interferometer is a plot of light output versus retardation which is called an interferogram.

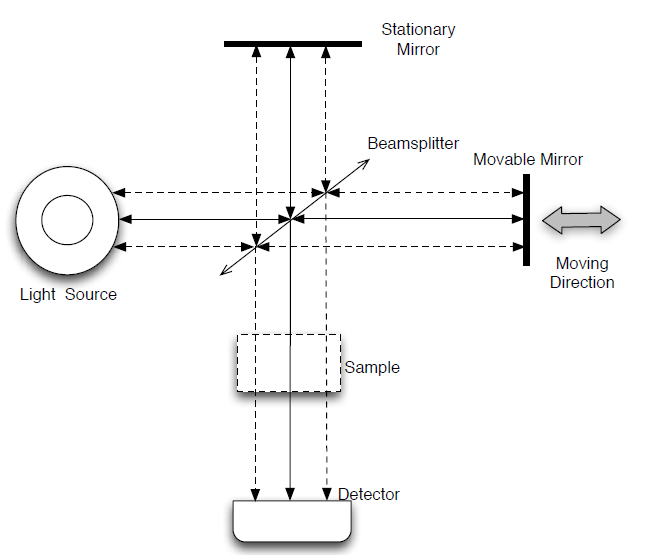


Figure 2.1 Schematic diagram of Michelson interferometer used in FTIR spectrometer [24].

For a monochromatic light source, the interferogram can be denoted by a cosine wave [25]. Intensity of interferogram is a function of retardation δ and is defined as shown in Equation 2.1 below.

*I0 [1+cos(2 π* δ υ)] (2.1)

Where *I*0 is the intensity of interferogram at zero retardation and υ is the wavenumber (cm-1) and equal to 1/λ, where λ is the wavelength of the light.

In most of the FTIR spectrometers, a sample is placed before the detector and a broadband source is used to collect the raw spectrum of the sample.

For a polychromatic or a broadband continuous light source, equation 2.1 can be extended to

I ( = υ) *cos (2 π* δ υ) dυ (2.2)

The location where and all the light frequencies interfere constructively is known as the point of Zero Path Difference (ZPD). As a result, interferogram has maximum intensity at the point of ZPD and is also termed as centerburst.

From Equation 2.2, it is apparent that the sample spectrum *B*(υ) is the Fourier transform of the intensity of interferogram.

*B* (υ) = δ) *cos (2 π* δ υ) dδ (2.3)

As δ) is symmetric at Zero Path Difference, Equation 2.3 can be rewritten as shown below.

*B* (υ) = δ) *cos (2 π* δ υ) dδ (2.4)

In Equation 2.4, integration is performed with respect to δ and the maximum of value of it is chosen by the moving range of movable mirror used in the FTIR spectrometer.

Equation 2.4 demonstrates that infinite retardation can be obtained when the moving mirror moves towards infinity. However, the mirror can only move up to certain distance in practice. Therefore, the interferogram must be multiplied with a window to shorten the range of integration from 0 to δmax . The technique of truncating the interferogram to a finite range is commonly known as apodization. The most commonly used apodization functions in NIR spectroscopy are Gaussian function, sinc square and triangular square.

**2.2 Beer Lambert’s law:**

Beer Lambert’s law states that the absorbance spectrum (A) of the sample is related to the concentration of analyte of interest (C) and transmittance spectrum (T) as shown below

A = -log(T) = -log = ɛ c d ( 2.5)

Where ɛ is the molar attenuation coefficient, I is the intensity of transmitted light through the sample, I0 is the intensity of the incident light and d is the optical path length.

Equation 2.5 indicates that absorbance is related linearly with the concentration of analyte of interest.

Beer Lambert’s law also states that for a homogeneous mixture of q components, the measured spectrum at each wavelength is the linear sum of absorbance spectra of q components as shown in equation below.

(2.6)

where , and are optical path length, concentration and molar attenuation coefficient for each component respectively.

For a homogeneous mixture of sample, the molar attenuation coefficient and optical path length are fixed for each component. Hence the absorbance spectra can be defined as the summation of product of pure component spectra Pi, i = 1, 2, ... , q multiplied by the corresponding individual component concentration as shown in Equation 2.7.

A = C1P1 + C2P2 … + CqPq (2.7)

In the qualitative and quantitative analysis of analyte of interest from NIR spectroscopy, large number of samples need to be measured to extract the required information related to the analyte of interest. Therefore, Beer Lambert’s law for multiple samples (n samples) measured at m wavelengths for multi components can be expressed as

A = CP (2.8)

Where , and

In NIR spectroscopy, the next step is to perform quantitative and qualitative analysis of analyte of interest from the measured raw spectra using various signal processing and multivariate calibration models to extract the information related to the analyte of interest.

As mentioned in the previous chapter, the present study will focus on developing signal processing techniques for the non-invasive measurement of the blood glucose from an aqueous mixture of samples. However, the quality of the collected spectra might be mitigated due to several factors that may have negative impact on the prediction performance of the calibration model. The factors that affect the prediction performance of the calibration model are listed below.

1. Baseline variations due to ambient variations and instruments.
2. Noise which occur in high frequencies such as detector noise
3. The type of light scattering resulting from the skin and other constituents of blood is nonlinear, inhomogeneous and anisotropic; it reduces the signal to noise ratio due to degradation of the optical signal.
4. water has high absorbability in NIR range due to high absorption of OH bond and high concentration of water in human body, resulting in high background spectra of water, which has negative impact on the measured spectra.
5. Overlapping of the other components of blood and skin as the structure of glucose molecules and bonds are like many components in the blood.
6. Very low value of the glucose concentration in blood provides weak signals in comparison to other components of blood (the range of normal blood glucose concentration is from 65 to 120 mg/dL).

## **2.3 Non-invasive blood glucose monitoring:**

Diabetes mellitus is a chronic disease that occurs either when the pancreas cannot produce sufficient insulin or when the body cannot use the produced insulin in the right way. Insulin is a hormone that controls the blood sugar levels and it plays vital role in transferring the glucose to the body cells. Recent reports from International Diabetes Federation (IDF) team have informed that diabetes affects around 415 million people worldwide and is expected to increase to more than 640 million by 2040 [26]. Non-invasive approaches have attracted lot of attention as they improve the quality of life for diabetic patients worldwide. NIR spectroscopy is recognized as one of the promising techniques for non-invasive monitoring of the blood glucose levels [27].

The American Diabetes Association has announced that strict control of the glucose levels in blood minimize the chronic problems by 50-70% [28]. The diabetic patients need to monitor their blood glucose levels a minimum 4-6 times per day, to maintain a strict diet for the proper control of the disease.

The finger prick method is the conventional way of testing the blood glucose levels in which a blood sample is drawn from the patient and analysed using the chemical regents. But this technique is inconvenient, painful and expensive as high quality chemical reagents must be used. To reduce these complications, many researchers have come up with non-invasive methods based on optical radiation for the measurement of the blood glucose levels.

The following are the main advantages of the non-invasive approaches:

1. The non-invasive methods are inexpensive as these do not require any chemical reagents.
2. These methods are simple, painless and convenient as there is no need to draw the blood sample from the patient.

The main idea of non-invasive blood glucose monitoring based on optical measurement techniques is to transmit a finite band of radiation to a selected region of the body. The light intensity or polarization is modified as a function of the tissues through which the light is transmitted [29, 20].

There are several non-invasive approaches in the literature [31-36] for the non-invasive detection of the blood glucose levels, such as scattering spectroscopy, mid infrared (MIR) spectroscopy, near infrared (NIR) spectroscopy, fluorescence and photo acoustic approaches.

Each method has its own advantages and disadvantages. Among all the non-invasive methods, NIR spectroscopy has attracted many researchers as it provides accurate and robust measurements and it can lead to the production of glucose meters that are reliable, cheap and simple which can be used by patiently very easily. NIR spectroscopy has attracted in the process monitoring because of its compatibility with aqueous solutions and simplicity in the preparation of samples. The NIR spectroscopy resides in the range from 780nm to 2500nm. The NIR spectroscopy is divided into 3 regions [37, 38] namely; the first overtone region (1540nm to 1820nm), the short wavelength region (700nm to 1330nm) and the combination region (2000nm to 2500nm).

As each component of blood has unique response to the NIR radiation and the response can be observed using the wavelength range, spectrum profile and the number of peaks and their positions. Hence the NIR spectrum is considered as the fingerprint for each component; however, the components of blood have an overlapped response as they are composed of C-H-N-O bonds.

The NIR spectrum of glucose has three absorbance bands in both the overtone and combination regions and it is quite difficult to observe these bands in short wavelength region as the absorbability of glucose is very low in this region. In the overtone region, these bands are centred at 1.61, 1.69 and 1.73 μm. For the combination region, these bands are centred at 2.326μm, 2.273 μm and 2.115μm [37-40]. The water absorbability is very low in the combination region and the glucose band is narrower than water. Hence most of the research works including the present work in this thesis use the combination region as the glucose features are very clear in this region [41-43].

In NIR region, the spectra are overlapped, weak and broad that make it difficult to extract the useful information from the measured spectra. Hence robust signal processing methods are required to remove the unwanted variations from the spectra to recover the information related to analyte of interest. These signal processing methods should have the capability to suppress most of the variations related to biological components and thus improve the quality of the spectra. These methods should be robust, accurate and have high sensitivity and wide detection range. Regardless of the huge number of studies in this field, accurate, effective and robust signal processing methods yet to be developed for practical representative test data.

The collected spectra might be affected by several factors such as baseline variations, high background spectra of water, overlapping of other components, high frequency noise and light scattering. Hence the work presented in this thesis has focused on the development of signal pre-processing and calibration methods for the non-invasive measurement of the blood glucose levels.

## **Multivariate calibration models:**

Multivariate calibration models are most commonly used in Chemometrics to predict the concentration of analyte of interest from the measured spectra of samples. The principle of multivariate regression model is to find the relation between the concentration of analyte of interest and the measured absorbance spectra to extract the information related to anaylte of interest. Multivariate methods develop calibration model based on response and spectra variables for a known set of data and the model is then used to predict the concentration of analyte of interest for new set of variables. The objective of a multivariate regression model is to develop a mathematical relation between the chemical property of interest to the spectral data encoded across various wavelengths.

In general regression analysis is used to predict the relationship between dependent and independent variables as shown in the equation below:

Y =X β (2.9)

Where Y is dependent or response variable. β is the regression coefficient, X is the input raw spectra.

Multivariate calibration methods, commonly known as regression techniques could be used to predict the glucose concentration from the blood serum [28].

The regression approach is implemented in two phases.

1. Training phase
2. Validation phase

In the training phase, the regression coefficient vector β is computed using training data. In the validation phase, the responses corresponding to the new dataset is calculated based on the value of β computed during the training phase.

The most commonly used existing regression models for the quantitative analysis of NIR spectra are the Principal component regression (PCR) and the partial least squares regression (PLSR) [44, 45].

**2.2.1** Multiple Linear Regression (MLR):

Pearson introduced the concept of Multiple Linear Regression (MLR) in 1908. In general, MLR is used to predict the relationship between a dependent variable and two or more explanatory variables based on Beer Lambert’s law. The linear relation between concentration of analyte of interest and absorbance spectra are given by

Y = X (2.10)

Where Y is the concentration of analyte of interest, X is the measured raw spectra and β is regression coefficients vector.

The regression coefficients vector β is computed using least squares method [46, 47] and is given in Equation 2.11

β = ; (2.11)

The analyte of interest should be correlated with the obtained regression coefficients vector. Hence the performance of MLR prediction model is evaluated by using the correlation between pure component of analyte of interest and regression coefficients vector.

Block diagram of MLR model:

Raw data

YY Y

Compute the regression vector using Xt and Yt by applying least squares method as shown below

Partition of the measured spectra X into Xt and Xs, and Y into Yt and Ys

Prediction error

Predict the concentration of analyte for the new test data. β

**Figure 2.2: Schematic diagram of MLR model to predict the concentration of analyte of interest**.

Multiple linear regression model is developed as shown in the Figure 2.2. Initially the measured raw spectra and the concentration of analyte of interest are divided into training and test datasets. Then the regression coefficient vector for the training dataset is computed using least squares method as shown in Figure 2.2. The computed β vector is used to predict the concentration of analyte of interest for the new samples of test dataset. The capability of the MLR model can be evaluated by computing the Root Mean Square Error of Prediction (RMSEP), and correlation coefficient (coefficient of determination) R2.

RMSEP measure the ability of prediction model to determine the concentration of analyte of interest for a new set of data. RMSEP can be computed as shown in Equation below.

RMSEP = (2.12)

Coefficient of determination (R2) indicates the strength of the linear association between independent values and the dependent values and describes how well the regression line fits the data. R2 measures the amount of variation in the data that is appropriately developed by regression model to the total variation. R2 can be computed as shown below;

R2  = 1 - (2.13)

Where is the total number of test samples, is the actual glucose concentration, is the predicted glucose concentration and is the mean glucose concentration.

Multiple linear regression can be applied only when the rows of X are greater than or equal to its columns. However, when the number of rows is less than the number of columns as in the case of NIR data, MLR cannot be used for the quantitative analysis of NIR spectra. Hence two popular linear regression models principal component regression and partial least squares regression methods have been utilized for the measurement of glucose from the NIR spectra.

## 2.4.2 Principal component regression (PCR):

Principal component regression (PCR) is based on Principal component analysis (PCA) followed by Multiple Linear Regression (MLR) step [48-50]. In PCA, the response matrix is decomposed into two matrices called scoring (*S*) and loading *(L*) matrices as shown below.

*X = SL1 + E* (2.14 )

Where X is the input raw spectra with dimensionality of n × m; n represents the number of spectra and m is the number of wavelengths, S is the scoring matrix with n spectra in the rows and h factors and L is the loading matrix with dimensionality of m × h and E is the residual matrix.

To obtain the loading matrix, Eigen values and the corresponding eigen vectors are computed for the matrix X1X. The eigen vectors can be calculated using Non-linear iterative partial least squares (NIPALS) algorithm or singular value decomposition (SVD) method [51].

The loading vector corresponding to the largest eigen value is called first principal component as it projects most strongly onto X. As all the loading vectors are orthonormal, each vector provides unique information in X.

By taking the loading vectors for h largest eigen values forms h-dimensional basis that can represent the information in raw matrix most efficiently.

The obtained loading matrix is used to compute the score matrix as follows.

The score vector corresponding to eigen vector i, Si = Xli ;

As the loadings are orthogonal, the calculated scoring vectors are also orthogonal.

Equation 2.15 is employed to build the PCR calibration model by replacing X with the calculated scores matrix S.

β= ; (2.15)

In Principal component regression, the principal component analysis is used to decompose the absorbance spectra into scores and loading matrices. The PCA algorithm preserves the factors that have high variance and neglects the features with the less variance [52-55]. Hence the PCA helps to remove the noise in the spectra. However, the analyte of interest may contribute less to the variation of the absorbance spectra. In that case, the useful information related to the analyte of interest is removed during the decomposition state and thus causes prediction errors. To overcome this problem PLS decomposes the absorbance spectra and the analyte concentration simultaneously by using the prior information of the analyte concentration in the decomposition of absorbance spectra and vice versa [56]. Hence, the information related to the analyte of interest is preserved even if it has the low contribution in the absorbance spectra.

## 2.4.3 Partial Least Squares Regression (PLSR)**:**

In Partial Least Squares Regression (PLSR), the features from principal component analysis and multiple regression are combined and generalized to predict a set of dependent variables from a set of predictors or independent variables. This prediction is achieved by extracting a set of orthogonal factors called latent variables from predictors or a set of independent variables as the latent variables have maximum predictive power.

In general, PLSR is useful in the applications where we need to determine a set of dependent variables from a large set of predictors (or independent variables).

X is a measured raw spectra of dimensionality n × m. Y is a matrix of dependent variables with dimensionality of n × l. The main goal of PLSR is to predict Y from X and to explain their common structure. This goal could be achieved using ordinary multiple regression when X is a full rank matrix and Y is a vector. When number of observations are less in comparison to number of independent variables or predictors, then multiple linear regression is not applicable due to multicollinearity. PCR method is one of the approaches used to address this problem. However, the principal components obtained using PCA explain X rather than Y and nothing guarantees that the principal components that explain X are relevant for Y.

In PLSR method, both X and Y are decomposed simultaneously to obtain a set of latent variables that describe maximum covariance between X and Y. Hence the principal components obtained using PLSR method, could be relevant to a set of dependent variables that lead to better prediction accuracy.

The Spectral matrix X and concentration vector Y can be decomposed simultaneously into scores and loading matrices as follows;

(2.16)

S l + e (2.17 )

Where X is a matrix with n measured spectra at p wavelengths. Y is a concentration vector. S is a scores matrix with a dimensionality of and Lis the loading matrix.

In partial least squares regression, the concentration vector Y is also decomposed into scores (S) and loading (l) matrices as shown in Equation.

The scores matrix S is like scores matrix in Equation 2.14 and l is a vector represents the loading values of the concentration vector. After the decomposition of both X and Y matrices into h latent variables, the remaining information can be found in the residual matrices E and e. The obtained scores are orthogonal whereas the spectral loadings are not orthogonal. In PLSR, the scores and loadings are both dependent on concentration vector and the measured spectra, which is different from principal component analysis.

There are several algorithms available to link both concentration vector (Y) and absorbance matrix (X) in computing the loading and scores matrices [57]. In this study, both X and Y are decomposed by computing a set of loading weights w for each spectral loading.

The first weight vector can be computed as shown in Equation 2.18.

(2.18)

*W*1 is a basis vector normalized to the unit length by .

The first scoring vector (s1) and its corresponding loading factor (L1) can be computed as shown in Equations (2.19 and 2.20) below.

(2.19)

(2.20)

Similarly, the first loading factor for the concentration vector can be computed as shown below.

(2.21)

After computing the first loading factor and scores vector, the residuals of the concentration vector and absorbance matrix can be computed as shown in Equation 2.22 and Equation 2.23 respectively.

(2.22)

(2.23)

The residuals e1 and E1 represents the remaining concentration information and spectral information which was not extracted from scores vector.

The loading weights, scores and loading factors for the second latent variable can be computed using Equations 2.19 – 2.21 by replacing X and Y with E1 and e1. Next, spectral residuals and concentration residuals for the second latent variable can be computed using Equations 2.22 and 2.23.

The same procedure is repeated until the loadings and scores for h latent variables are obtained.

PLS is called bilinear latent variable model as it incorporates the information related to concentration of analyte in addition to the absorbance matrix in calculating the PLS factors.

Once the scores and loadings are obtained, the procedure for the development of PLSR calibration model is identical to the development of PCR model as previously discussed in section 2.4.2.

To predict the concentration for new test dataset, the spectral matrix for test data is first mean centred and score matrix is computed using the loading weights (*w*1) computed previously in the calibration model. Then, Using the first computed spectral loading, the contribution of is removed from the spectral matrix. These computations are shown in Equations below.

(2.24)

(2.25)

This procedure is repeated till all the required scores are obtained. Then, the regression coefficient vector can be computed as shown in Equation 2.26 below

(2.26)

The calculated regression coefficient vector is used to predict the concentration for unknown spectra as given below.

(2.27)

Where is the mean concentration of the calibration model.

But PLSR model fails to perform better when the analyte of interest has high variation [58]. To overcome the drawbacks of both models, two calibration models have been proposed in this thesis; Local linear embedding regression combined with the digital band pass filtering and the Fourier Self Deconvoluted RReliefF coupled with PLSR (FSDR-PLSR) models.

2.4.4 Support Vector Regression:

The partial least squares regression and principal component regression are most commonly used methods for the quantitative analysis of NIR spectra [59], however these techniques may not be the best choice to deal with non-linear data. Non-linearity of the raw NIR spectra arises due to several factors such as non-linear detector responses, deviations from Beer-Lambert’s law, interactions between analytes etc. support vector machines [60, 61 have demonstrated better results for non-linear data. SVM was initially developed for classification problems and later it is extended to build regression models. When SVM is used in the regression analysis then the model is known as support vector regression (SVR).

2.4.4.1 Support vector machines theory

Vapnik developed SVM from 1960-1990, which is established in the frame work of statistical learning theory and it was further extended by smola and his co-workers to its current existing form [ref]. SVM are based on the structural risk minimization (SRM) principle. SRM is superior to conventional empirical risk minimization (ERM) principle [61, 62]. ERM seeks to minimize only the training error; where as SRM minimizes an upper bound of the generalization error. Hence, SRM achieves better generalization performance than ERM in many machine learning applications.

The performance of the SVM as a classification or regression model dependent only on support vectors, which are a subset of training data points. SVM uses machine learning theory in order to improve the prediction accuracy of the model and avoids the over fitting of the data automatically. In this work, the aim of SVM is to use for regression task (SVR).

The most commonly used two variations of SVR are nu-SVR and epsilon-SVR [63]. The principal parameters of the SVR model are cost, gamma, epsilon or nu.

## pre-processing methods:

The ability of multivariate calibration model to determine the concentration of analyte of interest depends on the quality of raw spectra. However, the collected spectra from the FTIR spectrometer might be mitigated by several factors such as instrumental noise, scattering and temperature variations. Hence pre-processing techniques should be used to improve the quality of raw spectra before applying to the calibration model. The primary goal of pre-processing methods is to avoid the parasitic variations which has impact on the prediction capability of the calibration model. Several mathematical algorithms can be used as pre-processing methods such as derivatives, smoothing techniques, feature weighting methods and digital bandpass filtering techniques which are most widely used in this present thesis.

## Mean Centering:

Mean centering is commonly used as the first step in the construction of calibration models such as PCR, PLSR and SVR models. The main objective of mean centering is to prevent the mean from dominating the first extracted principal components or latent factors to improve the prediction performance of the calibration model. In this method, the average of each column is subtracted from its corresponding column [64].

If *A* is a data matrix with *m* rows and p columns, then mean centering is obtained as shown below

, i = 1, 2, …, p. (2.28 )

Where is the element of ith column and jth row and is the ith centred column of .

## 2.5.2 First derivative as a pre-processing technique:

Derivatives are the most common used pre-processing technique in the analytical spectroscopy [65].

The basic method of derivative is the finite difference method, which is computed as the difference between two subsequent spectral points.

(2.29)

Where is the derivative of the raw spectra is the absorbance at the wavelength number.

The near infrared spectra are affected by both the baseline variations and the high frequency noise. However, the first derivative can suppress only the baseline variations [66]. A digital bandpass filter could remove both the low frequency baseline variations and the high frequency noises [67].

## 2.5.3 Standard normal variate pre-processing technique:

Standard normal variate pre-processing technique is mostly used to process the scatter data to remove multiplicative interference of scatter [68, 69]. The standard normal variate pre-processing centres each spectrum and scales the spectrum by its own standard deviation (SD). SNV is implemented to function on an individual sample spectrum and the SNV transformation is shown in the equation below.

; j = 1,2, …, m. (2.30)

Where is the average absorbance value of the uncorrected ith spectrum, is the absorbance of the ith spectrum at jth wavelength and SD can be computed as shown below.

(2.31)

## 2.5.4 Fourier Self Deconvolution:

Fourier Self Deconvolution was first introduced by Kauppinnen et al to pre-process the Infrared spectra. The raw spectrum obtained from the NIR spectrometer is considered as the convolution of line shape function of the instrument and the actual pure spectrum. To suppress the effect of line shape function to get back the pure spectrum, a reverse operation called deconvolution needs to be performed. Deconvolution can be usually performed either in frequency domain or in time domain. Deconvolution in frequency domain is often faster and simpler as convolution in time domain is equal to multiplication in frequency domain, whereas implementing deconvolution in time domain is complex, slow and requires high memory storage [70].

In NIR spectroscopy, the Lorentzian function [71] can be used to express the line shape function as shown in the equation below.

(λ) = (2.32)

Where denotes the center frequency of the ith band, is the half of the Full Width at Half Height (FWHH) of the ith band and represent the peak amplitude of the ith band. In Fourier Self Deconvolution, Fast Fourier Transform (FFT) is used to convert the raw spectra into frequency domain and are divided by the FFT of the instrumental line shape function. However, the procedure of deconvolution in frequency domain tends to increase the high frequency noise components that degrades the signal to noise ratio of the spectra [72]. To compensate the effect of SNR degradation, the spectra needs to be truncated using the apodization function. This can be achieved by multiplying apodization window with the deconvolved spectra. Finally, the actual spectra are obtained by performing Inverse Fast Fourier Transform (IFFT) at the output of the multiplier.

## 2.5.5 Digital Bandpass filtering:

In the context of NIR spectroscopy, digital bandpass filtering is used to eliminate the unwanted variations from the measured spectra to extract the maximum information related to analyte of interest. As the raw spectra might be affected by high frequency noise and baseline variations that occur at low frequencies, selection of digital bandpass filtering pre-processing method is a good choice as it is effective in eliminating both the low and very high frequencies from the input spectra. Thus, it could remove both the baseline variations and high frequency components from the raw spectra and improve the quality of spectra. Digital bandpass filter is defined by two parameters; bandwidth and centre frequency. Both the parameters should be optimized to get the required band that contains maximum information relating to the analyte of interest. As baseline variations and high frequency noise do not behave according to a known spectrum, the pre-processing step needs to be coupled with the design of calibration model to calculate the optimum digital bandpass filter parameters.

The two types of digital filtering methods used in NIR spectroscopy are time domain filtering and Fourier domain filtering methods. In time domain filtering, the filtered spectra are obtained by convoluting the input raw spectra with the impulse response function of the specified filter. Fourier domain filtering is a method in which the frequency response function of the desired filter is multiplied with the fast Fourier transform of the raw spectra. To get back the original absorbance spectra, an inverse fast Fourier Transform is applied on the filtered spectra.

Fourier domain filtering is used in applications, where information of a signal is encoded in the frequency, phase and amplitude of the Fourier transform of the signal. Whereas, time domain filtering is used when the information of a signal is encoded in the profile of the raw spectra.

Information in the NIR signals is present in the frequency components of its Fourier transform.

In [73], it is shown that the Fourier transform of NIR spectra composed of frequency components that are related to components of mixture, resides in the mid band range of the spectra, whereas the baseline variations and noise resides in the low frequency and high frequency range of the spectra respectively. Hence the quality of spectra can be improved by eliminating the baseline variations and the high frequency noise components present in the spectra. This can be done using a digital bandpass filter, as it allows only a selected band of frequencies in the spectra by eliminating the very low and high frequencies in the Fourier domain of the spectra.

Time domain filtering is performed by convolving the impulse response of the filter with the raw spectra. Based on the impulse response function of the filter, filters are classified into two types; Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filter. IIR filters are more effective in comparison to FIR filters as they are faster and requires fewer filter coefficients, whereas IIR filters cause phase distortion in the processed spectra while FIR filter cause linear phase shift in the filtered spectra. To overcome this problem, dual filters are used in practice, where the spectra are first filtered using forward filter followed by a reverse filter.

As the NIR spectrum is a frequency encoded signal, Digital Fourier filtering has been widely used in Chemometrics. Initially, the raw spectra are converted into frequency domain using FFT and is then multiplied with the frequency response of the specified filter. IFFT is performed on the output of the multiplier to obtain the filtered spectra. Gaussian function is widely used in Fourier filtering as it has the same profile in both frequency and time domain and it is defined by standard deviation and mean which are identical to bandwidth and centre frequency respectively.

Chapter 3 Experimental Data Preparation

This chapter describes the preparation of the experimental datasets used in this study for the quantitative analysis of glucose in NIR spectra. In this work, two sets of spectral data were acquired from two aqueous mixtures using FTIR spectrometer. First set of aqueous mixture is prepared by dissolving urea, glucose and triacetin in a PBS. To further improve the complexity of the data, the human serum albumin, which is the most important protein component in the blood is used in the preparation of the second dataset. Second set of samples were prepared by dissolving HSA and glucose in a PBS.

3.1 Experimental Data preparation using urea, glucose and triacetin:

As the blood components made with the combination of Carbon, Oxygen and Hydrogen atoms will have the similar absorptivity features in the combination region of the NIR spectra, the resultant spectra may overlap in the NIR region [74]. The primary components of blood that have significant absorptivity in the combination region of NIR spectrum are triglycerides, protein, water, urea, haemoglobin, cholesterol and lactate. Therefore, in this work the glucose, triglycerides, urea in addition to water are used to generate mixtures that are closely related to blood. Triacetin and urea are used to represent the triglycerides and urea in blood respectively. The main objective of this study is to demonstrate the performance of the proposed models in comparison to the existing linear and non-linear calibration models, therefore three main components are used to signify the blood.

For this experiment, samples were prepared by dissolving glucose, urea and triacetin in a PBS. Dry solutes of glucose and urea were dissolved in the buffer to prepare their aqueous solutions whereas triacetin solution was diluted by the buffer solution. Initially the buffer solution was prepared by dissolving 3.4023 grams of potassium dihydrogen and 3.0495 grams of sodium mono hydrogen phosphate in distilled water. The pH of the PBS was adjusted to 7.4, which is approximately equal to pH value of blood. A preservative in the form of fluorouracil was added to the buffer solution. The analytes used in this experiment were purchased from Sigma Aldrich, UK.

The preparation of the aqueous mixtures is illustrated in the Figure

3.4023g of potassium

dihydrogen

Distilled water

3.0495g of sodium mono hydrogen phosphate

Phosphate buffer solution (PBS) pH=7.4

Urea

glucose

triacetin

Aqueous mixture1

Figure 3.1 : Illustration of experimental data preparation for the urea dataset.

In this study 30 samples were prepared by varying concentrations of glucose, urea and triacetin. The concentration of these solutions was chosen in such a manner that it was within physiological range in blood. Concentration of glucose, urea and triacetin ranged from 20 to 500 mg/dL, 0 to 50 mg/dL and 10 to 190 mg/dL respectively. All experiments were carried out in a non-controlled environment to assess the capability of proposed models introduced in this work to deal with uncompensated alterations. In this research area, several previous studies have conducted the experiments in a controlled environment in order to compensate the effect of baseline variations.

After preparing the samples, triplicate spectra for each sample were collected with a Fourier transform spectrophotometer (spectrophotometer Cary 5000 version 1.09) which spanned the spectral region from 2000 nm to 2500 nm with a spectral resolution of 1 nm. In this way a total of 90 NIR spectra were collected from 30 samples. The purpose of using three replicate spectra is to reduce the effect of instrument noise. The absorbance spectra of the buffer solution were used as reference spectra.

3.2 Human Serum Albumin Dataset preparation: In this work, a Fourier transform NIR spectrometer was used to collect 100 NIR spectra for 100 mixtures composed of Human Serum Albumin and glucose in a phosphate buffer solution. The preparation of the buffer solution was done by adding Phosphate powder into 1 litre of distilled water to prepare the buffer solution with a concentration of 0.01M/dl and the PH value was adjusted to 7.4. For each sample four NIR spectra were measured and then averaged without removing the sample from the spectrometer to improve the precision of the measurement [75]. The individual component concentrations were selected such that these cover their physiological range in blood. The ranges of the glucose and Human Serum Albumin concentrations of the prepared 100 samples are from 5 to 500mg/dl and 5 g/dl respectively. The collected NIR spectra spanned the spectral region from 2100nm to 3120nm with a spectral resolution of 1.7nm.

Human serum albumin

Phosphate buffer solution

Aqueous mixture 2

Glucose

**Figure 3.2 Illustration of experimental data preparation for Human serum albumin data.**

**Chapter 4: Local Linear Embedding Regression**

In this chapter, a novel regression method called Local Linear Embedding Regression (LLER) has been proposed for the chemical analysis to determine the concentration of glucose in NIR spectroscopy. LLER model has been introduced to overcome the inability of the PCR to determine the concentration of the components that show lowest variation in the measured spectra. The prediction performance of the proposed model is assessed and compared with the existing calibration models the PCR, the PLSR and the SVR models with and without pre-treatment methods. The prediction performance of the LLER model is demonstrated to determine the concentration of glucose in an aqueous mixture of triacetin, urea and glucose in a PBS.

## **4.1 INTRODUCTION**

In the qualitative and quantitative analysis of NIR spectroscopy, advanced spectral processing methods are required to enhance the quality of measured spectrum before extracting the desired information. Differentiation, resolution enhancement, filtering, dimensionality reduction and outlier detection are the most frequently used data processing methods prior to the use of multivariate methods in the processing chain. In general, multivariate regression algorithms require large number of collected spectra to determine the chemical properties of the samples in an effective manner. To reduce the memory space, dimensionality reduction techniques are used prior to multivariate techniques.

In general, dimensionality reduction techniques use minimum number of dimensions to indicate the objects as accurately as possible. Dimensionality reduction methods are used to identify data patterns, to reduce memory space, to obtain accurate and fast spectra searching system and to simplify the data visualization process [76]. Many algorithms have been used up to date for the dimensionality reduction. Principal Component Analysis (PCA) is a commonly used dimensionality reduction technique in chemometrics, due its popularity in reducing the dimensionality without any loss of information [77]. PCA is one of the most popular linear dimensionality reduction techniques used in chemometrics [55]. PCA method is fast and simple to implement as only one parameter should be set and it is a powerful tool to analyze the spectral data.

The PCR and the PLSR are the commonly used multivariate calibration models for the qualitative and quantitative analysis of glucose in NIR spectra [53-58]. In the PCR method, the absorbance matrix *A* is decomposed into two sub matrices called score and loading matrices using the PCA. The PCA algorithm preserves the factors that have high variance and eliminate the factors that represent lowest variation of *A*. Hence, PCA helps to reduce the dimensionality in addition to noise present in the spectra. However, the component of interest may represent lowest variation in the spectra. In this case, the information relating to the analyte of interest may be lost during the process of decomposition, particularly when there is high interference or signal to noise ratio is low [78]. Hence the elimination process may deteriorate the prediction capability of the multivariate regression method.

Partial Least Squares Regression method overcomes this problem by decomposing the absorbance spectra and the concentration concurrently by the use of previous information about the concentration of analyte in the decomposition spectra. Hence, the information relating to the analyte of interest will be magnified, though it represents the less variation in the spectra. However, the PLSR model fails to achieve better prediction if the component of interest shows much variation [79].

The inabilities of the PLSR and the PCR models discussed above stimulated the development of a novel calibration method that preserves the chemical information related to a component of interest, even if it has a high range of variation or has a low contribution in the absorbance spectra.

This chapter introduces the exploitation of LLER model in the data analysis of glucose from NIR spectra. In this Chapter, the use of the LLER method is investigated for the quantitative analysis of glucose from NIR spectra. A non-linear dimensionality reduction method called Local Linear Embedding (LLE) [80] is used in the LLER model to map the high dimensional data non-linearly into a low dimensional space. LLER is considered one of the robust regression methods for non-linear data due to its advantages such as good representational capacity, no local minima and high computational efficiency [81].

In this work, initially the LLER method is developed, evaluated and compared to the existing calibration models the principal component regression, the partial least squares regression and the support vector regression techniques. Pre-processing methods such as digital bandpass filtering and first derivative are also developed with different calibration techniques and the resulting models are evaluated. It is shown that the LLER method can be an attractive alternative regression model for the determination of glucose from NIR spectra.

## **4.1 Local Linear Embedding (LLE) Dimensionality Reduction Algorithm:**

For a raw absorbance matrix consisting of *N* vectors  with dimensionality *D,*  an LLE analysis can be implemented as illustrated in Figure 4.1 as follows: ­­

The LLE algorithm is defined by two parameters namely the number of nearest neighbours and the dimensionality of the embedded data. Let the dimensionality of the embedded data and the number of nearest neighbours be *d* and *K* respectively. Initially, for each data point, *K*-nearest neighbours are determined by using Epsilon size ball or Euclidean metric. The reconstruction weights  that best represent the data points by their neighbouring points can be computed so that they can be used to reconstruct the original data. where  indicates the contribution of *j’th* point to the *i’th* reconstruction and the reconstruction weights are computed by minimizing the following cost function *E(W)*.

 (4.1)

The cost function *E(W)* is the squared sum of the difference between the reconstructed data and the actual data, which represents the reconstruction error. The reconstruction error can be minimized with the two constraints shown below:

The first constraint is = 0, if Xj is not belongs to one of the *K* nearest neighbouring points; it means that every data point must be reconstructed only from its neighbouring points. The second constraint is the summation of all the reconstruction weights should be equal to unity. i.e.

.

The importance of the two constraints is that for any particular data point, the reconstruction weights are invariant to translation, rescaling and rotation of that data point and its neighbours. The invariance to translations is specifically achieved by the second constraint [82].

Solving equation 4.1 subjected to these two constraints is a least squares problem as shown in [80]. The optimum weights are invariant to translation, rescaling and rotation of the data point and its neighbours.

Finally, the *d* dimension embedded vector , can be computed by minimizing the local reconstruction error  as shown in equation 4.2.

 (4.2)

Where  are the reconstruction weights computed from equation (4.1) and  represents the local reconstruction error which is summed squares of the difference between the embedded vectors and their reconstruction. The local reconstruction error  can be minimized using the following two constraints:

1. 
2. 

where *I* indicate the identity matrix.

Finding the embedded vector , is a well-known problem in linear algebra and it is reduced by solving the sparse *N*×*N* Eigenvector problem [80, 83].

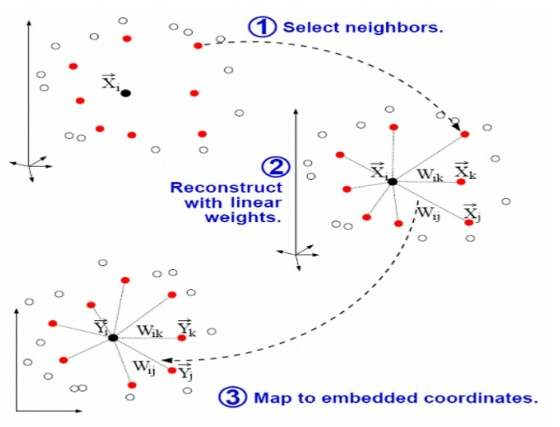


Figure 4.1 Implementation of LLE algorithm (http://youngjae8734.tistory.com/6)

LLE model has to set only single parameter *K* that affect the prediction performance of the LLER model in a direct way. However, incorrect choice of the parameter *K* may degrade the prediction capability of the calibration model. If the value of *K* is selected too high, the small-scale structure of the manifold will be removed, so the data mapping will lose its non-linear property [84]; on the other hand, if the value of *K* is selected too small, the manifold will be divided to disjoint sub-manifolds and hence the LLE algorithm will lose its favorable property of global optimality or convergence [83].

To overcome this problem, two methods are introduced in [84] to optimize the neighborhood size (*K*). In the first method, the residual variance of the embedded data is computed for all the possible values of *K* in the range [1-], which is a straight forward method. The value of *K* corresponding to the minimum residual variance is considered as the optimum value . The drawback of this method is that it is time consuming, as it needs to optimize both and the local reconstruction error and the reconstruction error *E(W)* for every value of *K*. The second method is the hierarchical method, in which the cost function *E(W)* is computed for different values of *K* in the range [1-]; *K\_opt* is considered as the value of *K* corresponds to the minimum residual variance. However, this method produces a set S of potential candidates for *K\_opt* as the residual variance has more than one minimum [84]. Finally, the residual variance for all the values of *K* from the set *S* must be computed and the K\_opt is chosen as the value of K corresponding to the minimum residual variance.

In this work, the first method is used to optimize the parameter *K*.

## **3.2.1. Local Linear Embedding Regression (LLER):**

The LLER is a non-linear calibration model, in which, the LLE dimensionality reduction technique is used to map the high dimensional absorbance spectra (*A*) to a lower dimensional embedded vector (*Y*).

The basic idea in the proposed LLER regression model is that the absorbance matrix *A* is decomposed as the product of the LLE matrix *Y* and the reconstruction factors P.

*A* = *Y*.*P* (4.3)

Where *D* is the number of variables in the raw spectra, *d* is the dimensionality of the embedded vector and *N* is the number of training spectra.

In the LLER method, the embedded vectors computed using the LLE analysis represent the scores and loading matrix is computed by multiplying the pseudo-inverse of the embedded vector with the input absorbance spectra as shown in equation 4. Then the obtained loading and scores matrices can be used in building the LLER calibration model.

The reconstruction matrix can be represented as shown in equation (4.4).

 (4.4)

Where  represents the pseudo-inverse of the embedded data matrix *Y*. The reconstruction factors P and the embedded data matrix Y are considered as the loading factors and scores respectively. As the embedded data Y relates to the concentration of analyte (Cg), the embedded data can be regressed against the concentration of analyte using Multiple Linear Regression (MLR) as shown in equation 4.5.

Cg =Y𝛃lle (4.5)

where 𝛃lle denotes the coefficients of the regression and 𝛃lle is defined by the least squares method as shown in equation 4.6.

 (4.6)

The concentration of analyte  for the new data  can be obtained from the following equation, when both the concentration and training spectra are centered.

 (4.7)

From equations 3 and 5, 𝛃 can be replaced by 𝛃lle

 (4.8)

Where is the average vector of the training spectra,  is the pseudo-inverse of the loading factors of the training spectra and is the average value of the training data concentration.

The LLER model has to set two parameters, one is the dimension of the embedded data *d* and the other one is *K* nearest neighboring points. If *d* is selected too small, different parts of the dataset might be mapped onto each other; conversely, if *d* is selected too high, the mapping reduces the signal-to-noise ratio; [85]. The minimum and maximum possible values of *K* for which the LLER model converges are chosen as the lower and upper limits of *K*.

In this work, the experimental data collected as explained in section 3.1 are used to build the PLSR, the PCR, the SVR and the LLER models. The developed regression models are tested using the test dataset. The standard performance measures used in this study are Root Mean Square Error of Cross Validation (RMSECV), Root Mean Square Error of Prediction (RMSEP) and Root Mean Square Error of Calibration (RMSEC). The error parameters RMSECV, RMSEP and RMSEC are computed for each value of *K.* The optimum parameters of the LLER model are chosen as the values *d* and *K* that together produce the minimum value of RMSECV.

**3.2.2 LLER model Combined with Digital Bandpass Filtering**

The prediction performance of the calibration model can be further improved by the integration of the LLER model with pre-processing techniques such as the bandpass filtering and first derivative pre-treatment method. To our knowledge, this is the first time that the LLER is coupled with digital bandpass filtering for the quantitative analysis of glucose in NIR spectroscopy. In this study, the Chebyshev and Gaussian digital bandpass filters have been used to suppress the baseline variations which occur in low frequencies and high frequency noise components in the raw spectra [86, 87]. In signal processing, digital filter is defined as a system which executes mathematical operations on a discrete time, sampled signal either to reduce or enhance certain properties of that signal. Based on the frequency response, the basic filters are broadly classified into four types; low-pass, high-pass, bandpass and band-reject filters. A low-pass filter is a filter that allows the signal with frequencies lower than a specified cut-off frequency and attenuates all the remaining frequencies and the cut-off frequency is called upper cut-off frequency. In high-pass filtering, the signal with frequencies higher than the cut-off frequency are allowed and the remaining signal frequencies are removed during the filtering process and the cut-off frequency is termed as lower cut-off frequency. Band-reject filter is a filter that allows the signal with frequencies higher than upper cut-off frequency and the signal frequencies below the lower cut-off frequencies and removes all the remaining frequency components. whereas, bandpass filter is a filter that allows the signal with frequencies above the lower cut-off frequency and the frequencies below the upper cut-off frequency. The difference between the upper cut-off frequency and the lower cut-off frequency is called the passband bandwidth. The arithmetic mean of both the cut-off frequencies is called the center frequency. The digital bandpass filtering methods are defined by two parameters [88-91]; the bandwidth and the center frequency. The optimization of these two parameters is very important to select the optimum band of frequencies that contains the maximum information related to the analyte of interest (glucose concentration).

A Gaussian filter can be implemented either in the time domain or in the frequency domain. The Gaussian function exhibit the same profile in both the time and frequency domains [92, 93]. In the frequency domain, the standard deviation and the mean of the Gaussian function are equivalent to the bandwidth and the centre frequency respectively. The Gaussian digital bandpass filter has been implemented in the frequency domain, as shown in Figure 4.2, due to its reduced complexity.



Figure 4.2: Block diagram of the Gaussian digital bandpass filter

Initially, the Gaussian function is multiplied with the Fast Fourier Transform of the input raw spectra. The raw spectra, which is normalized between 0 and 1 is applied as input to the Gaussian function. Finally, the filtered signal is obtained by performing an Inverse Fast Fourier Transform on the result at the output of the multiplier.

Chebyshev filters provide an optimal tradeoff between a steeper roll-off and passband ripples, compared to the other time domain filters [94] and Chebyshev filters can be efficiently implemented in time domain. Figure 4.3 show the block diagram of the Chebyshev digital bandpass filter.

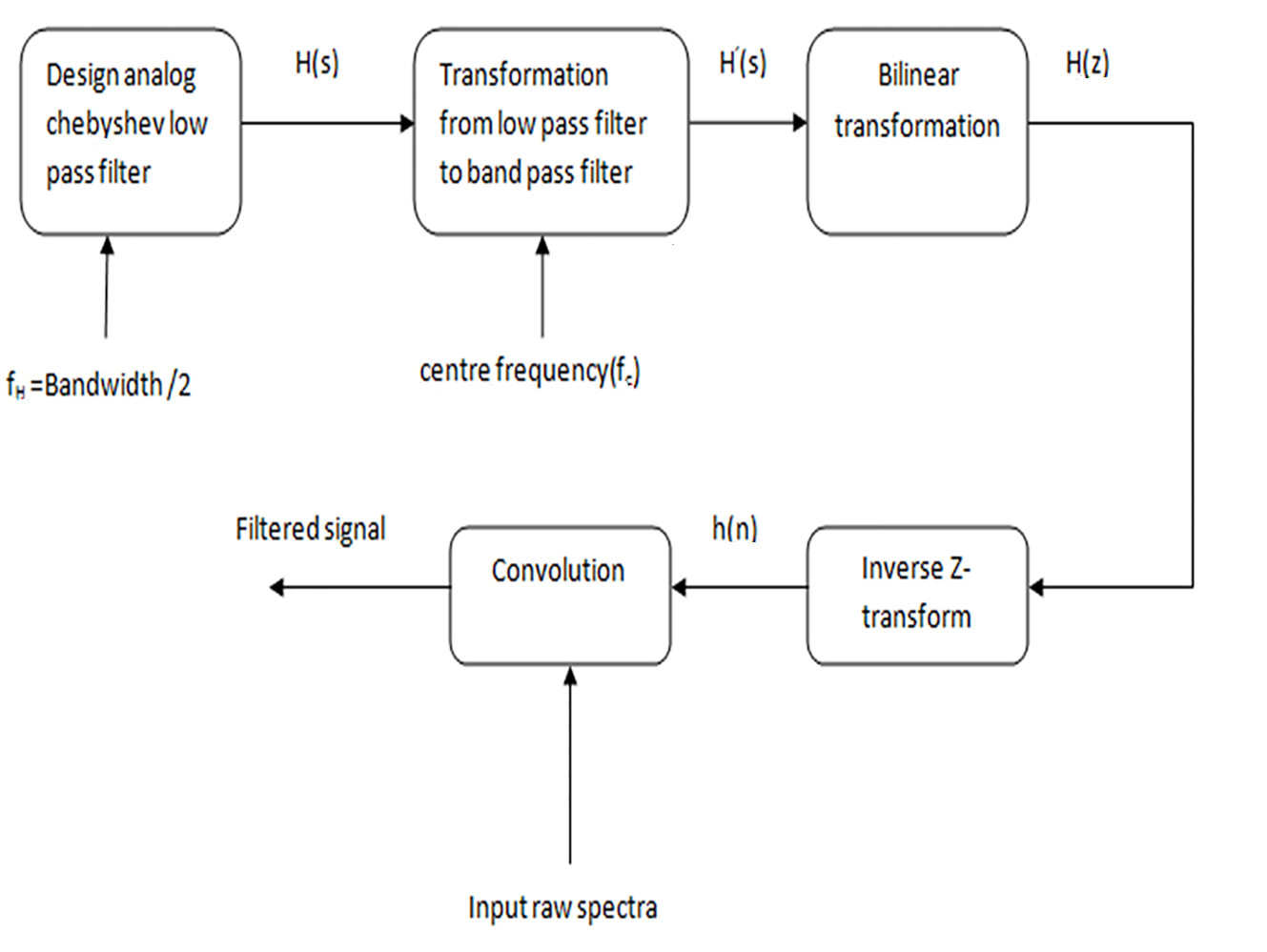


Figure 4.3: Block diagram of the Chebyshev digital bandpass filter

Initially, an analog low pass filter is developed, and the upper cut-off frequency is chosen as the half of the desired bandwidth of the Digital Bandpass (DBP) filter. The implemented low pass filter is then transformed to a bandpass filter by shifting the frequency spectrum to the centre frequency of the bandpass filter. The bilinear transformation is applied to convert the transfer function from analog domain to the digital domain. Inverse z-transform is then applied on the previous output to get the impulse response of the DBP filter. Finally, the filtered signal is obtained by the convolution of raw spectra with the impulse response of the Chebyshev filter.

In this work, the grid search optimization [95] is used to optimize the DBP filter parameters. Initially the RMSECV is computed for all possible values of bandwidth and centre frequency. The prediction performance of the developed models is evaluated by using the Root Mean Square Error of Calibration (RMSEC), the Root Mean Square Error of Prediction (RMSEP), the Root Mean Square Error of Cross Validation (RMSECV) in addition to the coefficient of determination (R2), An optimum model should have a low RMSEC, a high R2 , a low RMSEP, and a low RMSECV. The optimum values of bandwidth (w) and centre frequency (c) are selected as the values of w and c for which the RMSECV has the minimum value.

The collected 90 NIR spectra were divided randomly into calibration and test sets using KS algorithm [96]. The calibration set contained the three replicate spectra of 20 samples and was used to build the calibration model. The test set contained the triplicate spectra of 10 samples and was used in the prediction phase to test the calibration model.

The experiments were carried out in a non-controlled environment. i.e; experiments were not carried under constant temperature. This introduced signiﬁcant baseline variation in the collected spectra to evaluate the ability of the proposed methods in this work to deal with the uncompensated variations. Many previous studies in this area have carried out experiments in a controlled environment to compensate the effect of the baseline variation.

In this study, the Van Der Maaten toolbox [97] has been used to perform the LLE dimensionality reduction on the input raw spectra. The key parameters for LLE model are the number of nearest neighbours (*K*) and the embedded dimension (*d*). The grid search optimization was used to select the optimum values of *K* and *d* in order to prevent the overfitting problem. The doublet (*K*,*d*) with the lowest RMSECV is used to build the final LLER model. The optimum number of PCs and LVs for the PCR and PLSR models were found using “10-fold cross validation” respectively [98]. The key parameters for SVR model using Radial Basis Function (RBF) kernel are cost (C), gamma () and epsilon (). The grid search optimization on C,  and  using 10-fold cross validation was used to avoid over fitting problem as mentioned in LIBSVM (A Library for Support Vector Machines) [99]. The triplet with minimum RMSECV were chosen as the optimum parameters to build the final SVR model.

The grid search optimization [95] is used to optimize the filter parameters (*c*,*w*). In the optimization of the DBP filtering, the centre frequency (*c*) is varied from 0.01 f to 0.5 f and the bandwidth (*w*) is varied from 0.01 f to 0.8 f; where f is the normalized frequency [66]. The values for the filter parameters (*c* and *w*) are chosen in such a way that the filter spans the whole frequencies from *fL*= (*c*-*w*/2) to *fH*= (*c*+*w*/2); where fLis the lower cutoff frequency and fHis the upper cuttoff frequency of the designed digital bandpass filter. In each iteration, the designed digital bandpass filter is combined with the prediction model and the RMSECV is calculated. The computed RMSECV is then stored in the variable called *SECV* and is compared with *SECV\_opt* as shown in the flowchart below; where *SECV\_opt* is the temporary variable used to store the updated minimum RMSECV value in each iteration. The values of *c, w, k* and *d* corresponding to the minimum RMSECV value are chosen as the *c*\_*opt*, *w*\_*opt*, *K*\_*opt*, *d*\_*opt* respectively. The maximum values for *c, w, K* and *d* are considered as *c*max, *w*max, *K*max, and *d*max respectively.

Parameter initialization

*c*=0.01f; *w*=0.01f; *d* = 1;*K*= 3; SECV\_opt= 200;

*c*max=0.5f, *w*max=0.8f, *d*max=30, *K*max=59

DBPF (Gaussian or Chebyshev)

LLER model (*K,d*)

If (*SECV*<*SECV\_opt*)

No

yes

*SECV\_opt*=*SECV*; *(c\_opt,w\_opt)opt=(c,w);*

*(K\_opt,d\_opt)opt=(K,d);*

Increment *K* by 1

If (*K*>*k*max)

No

yes

Increment *d* by 1

No

If (*d*>*d*max)

yes

No

If (*w*>*w*max)

Increment *w* by 0.001f

No

yes

If (*c*>*c*max)

Increment *c* by 0.01f

Figure 4.4 : Flow chart of parameter optimization for DBPF-LLER model

yes

Optimized LLER model

The prediction model with the lower RMSECV is chosen as the optimized digital bandpass filter. The optimum filter parameters for the Chebyshev filter are found to be c=0.03 f, w= 0.04 f and for the Gaussian digital filter, these were c= 0.02 f, *w*=0.01 f .

The selection process of the parameters for the optimum DBPF-LLER model is illustrated in the flow chart as shown in Figure 4.4.

## **4.4 Discussion of Experimental Results and Comparisons:**

For the evaluation, validation, and comparisons, a set of prediction models were developed. Initially the PCR, PLSR, SVR and LLER models were implemented with no pre-processing. The prediction performance of the models was examined by computing the RMSEP, RMSEC, RMSECV and R2 for each model. Figure 4.5 shows the comparison of all the prediction models with no pre-processing; the x-axis shows the reference glucose concentration (mg/dL) and the y-axis represents the predicted glucose concentration (mg/dL). The ‘\*’ symbols correspond to the test samples where as ‘o’ symbols correspond to the calibration. The straight line is the reference line.

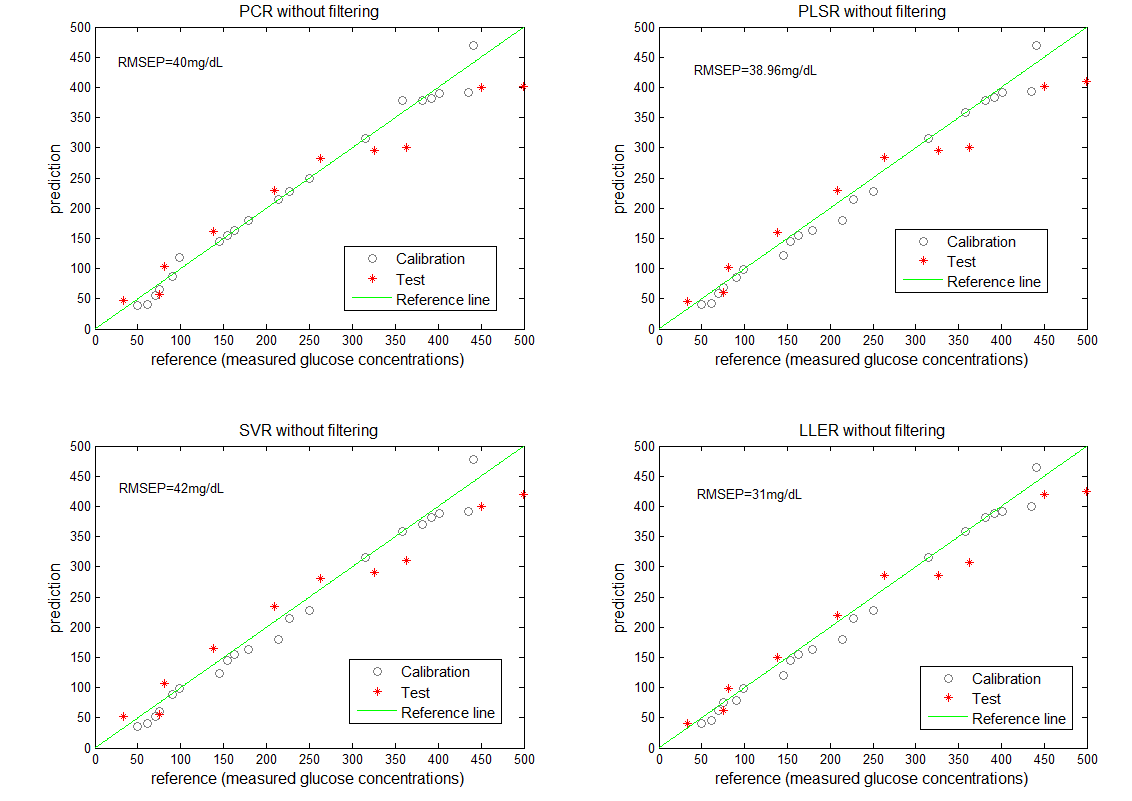
****

Figure 4.5: Comparison of the PCR, PLSR, SVR and LLER models without pre-processing

The results demonstrate that the LLER model gives a better prediction compared to the PCR, PLSR and SVR models when no pre-processing of the raw data is used. This is an interesting result that confirms the advantage of adopting an efficent non-linear dimensionality reduction technique (LLE) in a calibration model when dealing with NIR spectra. Figure 4.5 shows that the LLER model exhibits a more consistent precision of calibration relative to the PCR, PLSR and SVR models, although the testing and training data had a wider range of glucose concentation. The advantage of the LLER method over the PCR, PLSR and SVR models is that it preserves the neighbourhood structure of nearest spectra in the mapped plane. The LLE algorithm maps the high dimensional input coordinates into low dimensional data (Y) by minimising the cost function as given in equation 4.2. The cost function is based on the reconstruction coefficients of K nearest neighbours. Then the mapped data are regressed against the analyte of interest to build the calibration model, which is completely identified by the embedded dimension d and the K nearest neighbours. So, the values of K and d affect the prediction performance of the LLER model. This has been investiagted and Table 4.1 below summarises the impact of these two parameters on the resulting RMSEP and RMSECV values for the LLER model.

Table 4.1: The prediction capability of the LLER model for different values of K and d

|  |  |  |
| --- | --- | --- |
| Calibration model | RMSECV (in mg/dL) | RMSEP (in mg/dL) |
| LLER (K=18 , d=14 ) | 34.90 | 33.20 |
| LLER (K=18 , d= 15) | 36.10 | 36.00 |
| LLER (K=18 , d= 16) | 34.80 | 35.30 |
| LLER (K=19 , d= 14) | 35.70 | 34.60 |
| LLER (K=19 , d= 15) | 32.60 | 31.00 |
| LLER (K=19 , d=16 ) | 33.40 | 35.20 |
| LLER (K=20 , d=14 ) | 38.20 | 36.50 |
| LLER (K=20 , d= 15) | 34.60 | 34.20 |
| LLER (K=20 , d= 16) | 33.70 | 36.80 |

Furthermore, as already mentioned, appropriate pre-processing of the raw data prior to applying the calibration model can yield tangible improvements in prediction, since the raw NIR spectra are affected by baseline shift, background noise, light scattering and instrumental noise in general. Hence, a set of pre-processing techniques including first derivative, Gaussian digital bandpass filtering and Chebyshev digital bandpass filtering are applied and evaluated for each model.

Firstly, the PCR and PLSR models were implemented with the different pre-processing techniques where the number of factors that produce the minimum RMSECV are chosen as the optimum number of principal components and latent variables for PCR and PLSR respectively. The comparison of PCR and PLSR when different pre-processing techniques are applied is shown in Figure 4.6. The y-axis shows the RMSECV and the x-axis represents the number of principal components or Latent variables for PCR and PLSR respectively. The results show that the models with pre-processing of NIR data gives much better prediction accuracy in comparison to models with no pre-processing. From Figure 4.6, it is also observed that models with bandpass filtering achieve better prediction accuracy in comparison to the first derivative pre-treatment. The optimum number of principal components and latent variables are identified to be 6. Information about NIR spectra is prominent in the frequency components in the mid-band range, while the noise and baseline variations tend to occupy the high and the low frequency range respectively, that is why these can be effectively reduced using an optimised bandpass filter rather than the first derivative which tends to reduce the signal to noise ratio (SNR). First derivative pre-processing can eliminate only baseline variations in the raw spectra, whereas the bandpass filter can eliminate both the low frequency baseline variations and the high frequency noise from the spectra.

The PCR, PLSR, SVR and LLER models were then implemented with the raw data pre-processed using the first derivative, the Gaussian, and the Chebyshev digital bandpass filters.

## 

Figure 4.6: PCR and PLSR with different pre-processing techniques

Figure 4.7 illustrates the prediction performance comparison of the PCR, PLSR, SVR and LLER models with the three different pre-processing methods. For each subplot, the x-axis represents the reference glucose concentration (mg/dL) and the y-axis shows the predicted glucose concentration (mg/dL). The ‘o’ symbols correspond to the calibration where as ‘\*’ symbols correspond to the test samples. The reference line is represented by a stright line as shown in Figure 4.7. The results as summarized in Table 4.2, demonstrate that the LLER combined with the Chebyshev filter gives the best prediction accuracy. The advantage of a Chebyshev filter over a Gaussian bandpass filter is that it offers an optimal trade off between a steeper roll off and passband ripples. Hence, it is more effective in reducing the effect of both the high frequency noise and low frequency baseline variations without affecting the mid-band NIR data.

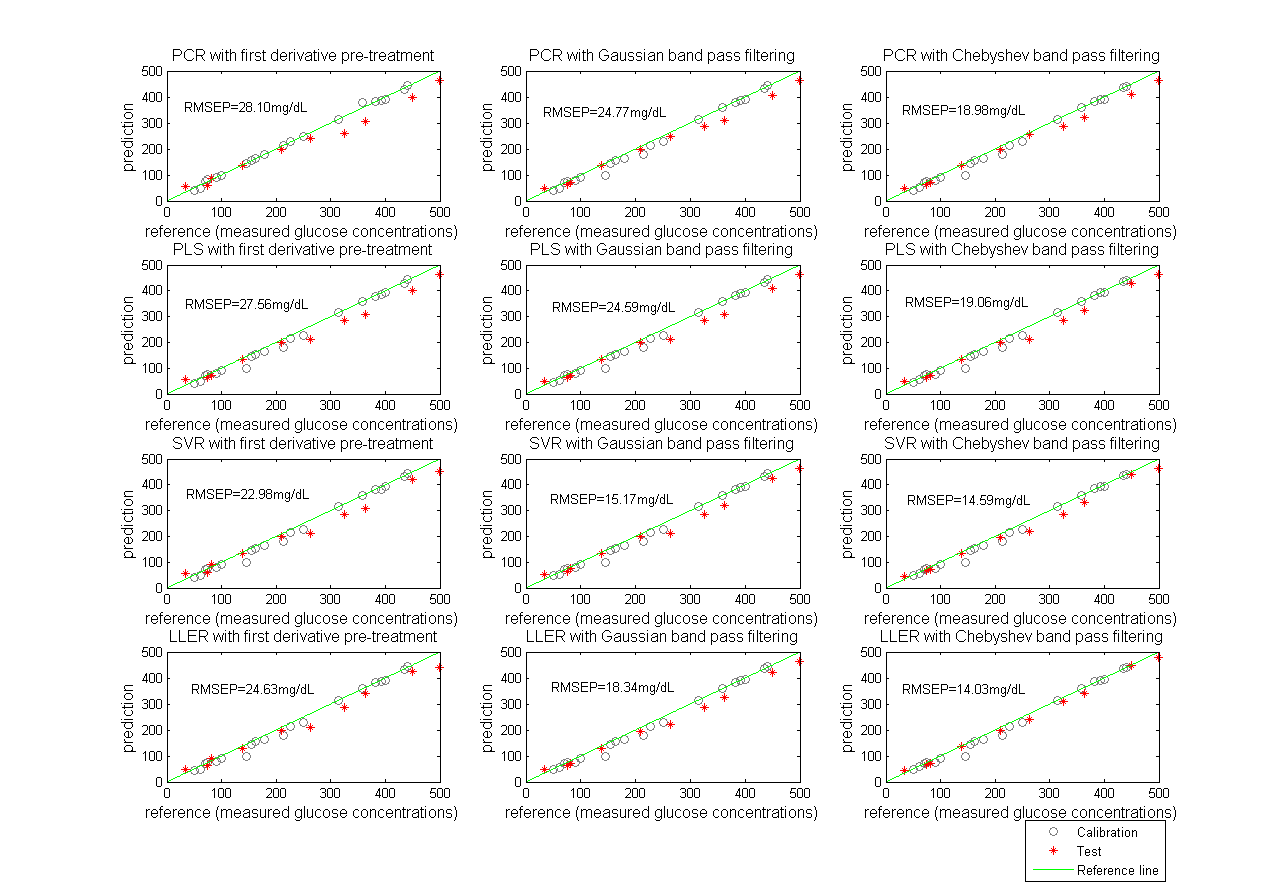
**

Figure 4.7: Comparison of PCR, PLSR, SVR and LLER models with different types of pre-processing techniques.

Table 4.2: Comparison of PCR,PLSR, SVR and LLER models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regression model** | **Pre- processing** | **Optimum parameters** | **RMSEC\*** | **RMSECV\*** |  | **RMSEP\*** |
| **PCR** | None | 6PCs | 25.34 | 67.59 | 0.90 | 40.00 |
| **PCR** | 1st derivative | 6PCs | 24.92 | 51.07 | 0.88 | 28.10 |
| **PCR** | GDBPF | 6PCs | 17.54 | 56.70 | 0.97 | 24.77 |
| **PCR** | CDBPF | 6PCs | 15.93 | 51.23 | 0.98 | 18.98 |
| **PLS** | None | 6LVs | 11.30 | 34.07 | 0.90 | 38.96 |
| **PLS** | 1st derivative | 6LVs | 22.54 | 31.59 | 0.97 | 27.56 |
| **PLS** | GDBPF | 6LVs | 12.00 | 38.30 | 0.96 | 24.59 |
| **PLS** | CDBPF | 6LVs | 15.92 | 28.43 | 0.98 | 19.06 |
| **SVR** | None |  C=0.1\*10^6 | 2.50 | 38.44 | 0.90 | 42.00 |
| **SVR** | 1st derivative |  C=0.2\*10^6 | 13.50 | 28.98 | 0.98 | 22.98 |
| **SVR** | GDBPF |  C=0.04\*10^6 | 12.09 | 28.00 | 0.99 | 15.17 |
| **SVR** | CDBPF |  C=4.5\*10^6 | 12.47 | 27.40 | 0.99 | 14.59 |
| **LLER** | None | K=19, d=15 | 18.52 | 32.60 | 0.95 | 31.00 |
| **LLER** | 1st derivative | K=29, d=25 | 15.55 | 31.50 | 0.97 | 24.63 |
| **LLER** | GDBPF | K=33, d=20,c=0.03f, w=0.04f | 14.92 | 27.80 | 0.98 | 18.34 |
| **LLER** | CDBPF | K=55, d=23 C=0.02f, w=0.01f | 17.80 | 27.12 | 0.99 | 14.03 |

\*=(units are in mg/dL);GDBPF=Gaussian digital bandpass filter;CDBPF=Chebyshev digital bandpass filter*.*

**Chapter 5: Improved support vector regression coupled with Fourier self-deconvolution**

Adopting a non-linear dimensionality reduction method in building a calibration model achieves better prediction results in comparison to linear models as discussed in the previous chapter. In this chapter, a non-linear calibration model based on combining the improved support vector regression coupled with the Fourier self-deconvolution is proposed for the quantitative analysis of glucose in near infrared spectra. The proposed model has been validated to predict the glucose concentration from the mixture of glucose and human serum albumin in a phosphate buffer solution. The PCR and the PLSR methods are also developed under the same conditions and are compared with the proposed model.

### **5.1 Introduction:**

As seen in Chapter 4, determining the concentration of glucose from a mixture of other components, incorporated with spectral variations and the noise remain an issue, which can be resolved by developing the robust multivariate regression methods to predict the concentration of glucose from NIR spectra. The most commonly used linear calibration models the PCR and the PLSR may not be the best choice when the data is non-linear. The non-linearity in the NIR data may exists due to several factors such as non-linear detector responses, interaction between the components and the deviations from the Beer-Lambert Law. Due to strong non-linear capabilities, artificial neural networks have been reported with better results [100-102]. However, the limitation of neural networks is that there exists more than local minimum. SVM have demonstrated promising results similar to neural works for non-linear data but without its pitfalls.

SVM was initially implemented as a calibration model and was later extended to develop the regression models. SVM, when used to build the calibration models is known as Support Vector Regression (SVR). SVR has recently attracted growing research interest in Chemometrics due to its global and unique solution along with its good generalization ability.

The raw data collected from the spectrometer might be contaminated by instrumental noise. Hence a pre-processing step is needed before applying to the calibration model. The derivative, standard normal variate and detrending are widely used pre-processing techniques for the NIR spectra. However, these methods are effective only to remove the baseline variations from the spectra and are ineffective to eliminate the high frequency noise from the spectra [72]. The principal component Regression (PCR) and Partial Least Squares Regression (PLSR) are the basic linear calibration models used in Chemometrics [59]. The Partial least squares regression coupled with Fourier self-deconvolution proved to be effective to remove most of the noise components present in the spectra [71]. In this chapter a non-linear calibration model based on combining the improved support vector regression with the Fourier self-deconvolution is introduced to predict the glucose concentration from a mixture of glucose and human serum albumin in the phosphate buffer solution.

### **5.2** Support vector regression (SVR):

The Support Vector Regression function in its dual form is defined by

S(x) =) d( xj ,x)+b (5.1)

Where are the Lagrange multipliers, x is a vector for a new sample, b represents the bias term and d ( xj ,x) is a non-linear kernel function [103, 104].

### **Improved support vector regression:**

For SVR, many algorithms require a single batch of training samples. For Some applications such as leave one-out cross validation or on-line time series prediction, another model is needed every time when a new sample is removed or added from the training set [105]. Retraining for each new data sample from scratch could be more expensive. In [105], Ma, Junshui etal introduced online support vector regression (also known as improved support vector regression) to solve these kinds of problems.

### **Fourier Self-Deconvolution**

Kauppinen et al introduced the Fourier Self-Deconvolution (FSD) and was used as a pre-processing technique in Infrared (IR) spectroscopy. The raw spectra are treated as the convolution of the line shape function and the actual spectra [70]. The effect of the line shape function should be eliminated to get back the original spectra. Hence the reverse operation of devolution needs to be performed. The deconvolution can be performed either in the frequency domain or time domain. Deconvolution in frequency domain is simpler because the convolution in time domain is equivalent to multiplication in frequency domain and faster, whereas deconvolution in time domain is complex, slow and requires high memory storage [106]. The line shape function can be represented by the Lorentzian function in the NIR range and is defined as shown in equation 5.2.

(λ) = (5.2)

where, is the centre frequency of the lth band, is the half of the Full Width at Half Height (FWHH) of the lth band and is the peak amplitude of the lth band.

In FSD, the Fast Fourier Transform (FFT) is used to convert the raw data into frequency domain and the resultant spectra are divided by the line shape function of the instrument. However, deconvolution in frequency domain degrades the signal to noise ratio (SNR) of the spectra [71] by increasing high frequency components of the noise along with the high frequency components of the spectra. The SNR degradation can be compensated by truncating the deconvolved spectra using the apodization function, which can be done by multiplying the apodization window and the deconvolved spectra. Finally, the actual spectra are obtained by applying an Inverse Fast Fourier Transform (IFFT) on the output of the multiplier.

### **Experiments:**

In this work, mixtures of human serum albumin and glucose are prepared and the NIR spectra of mixtures are measured as discussed in Chapter 3. 100 samples are prepared, and spectra are collected using FTIR spectrometer. Kennard Stone (KS) partition method [96] was used to divide the collected spectra into training dataset consisting of 75 spectra and 25 spectra were used for testing.

### **Results and Discussions:**

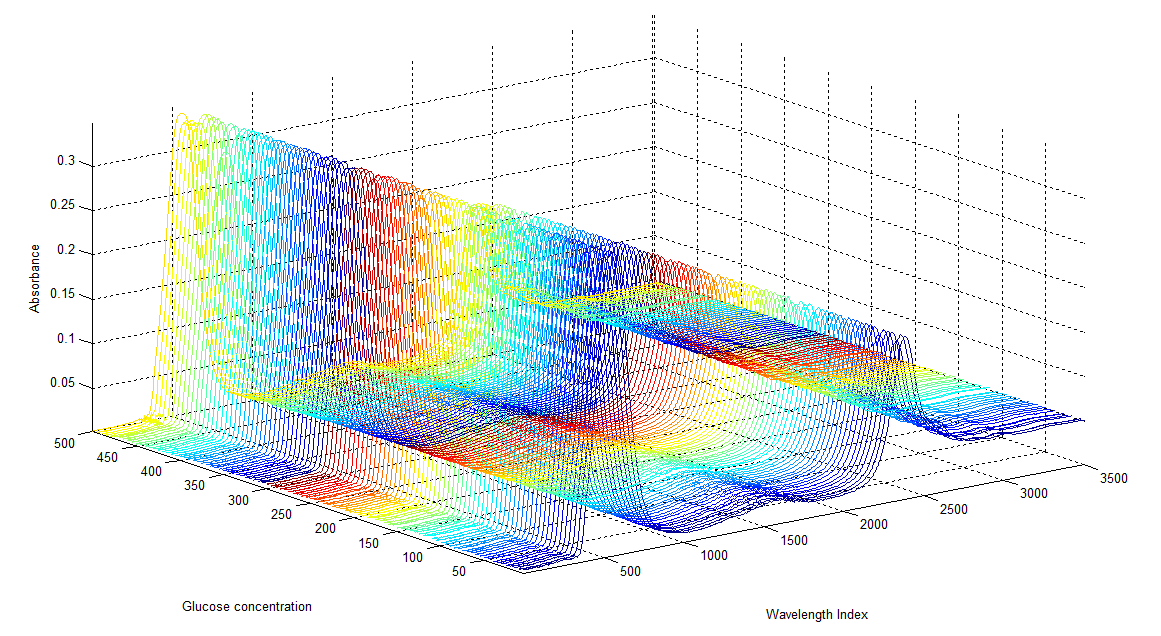
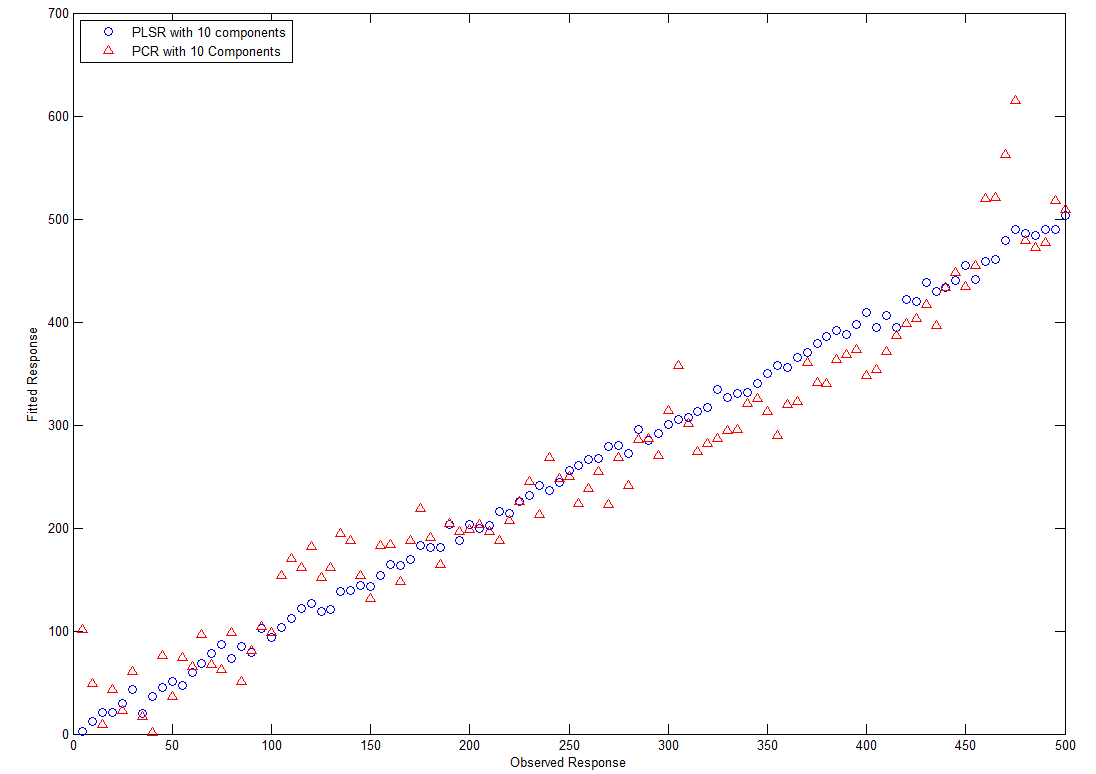
****

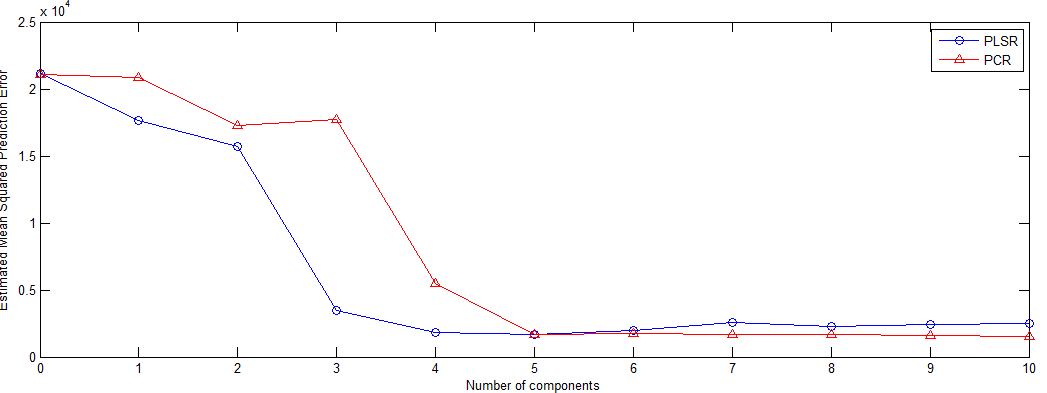
Figure 5.1: 3D plot for the collected Human serum spectra

Figure 5.1 shows the spectra collected from the Fourier transform near infrared spectrometer, which are the spectra for a mixture of human serum albumin and the glucose (with different glucose concentrations) in the phosphate buffer solution.

****

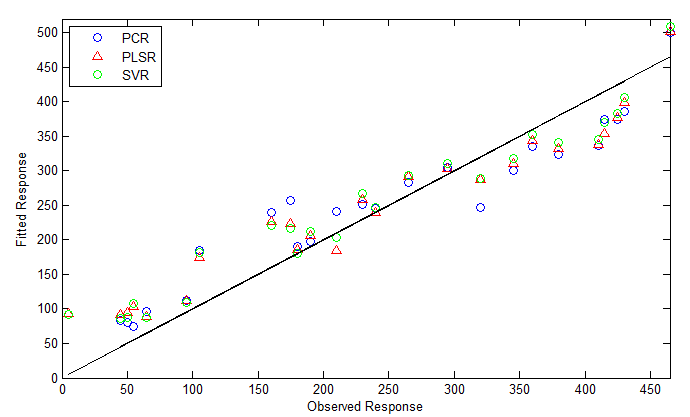
**Figure 5.2: Prediction performance of PCR and PLSR**

Both the PCR and the PLSR models were developed using 10 principal components and the predicted results are plotted and are shown in Figure 5.2.

****

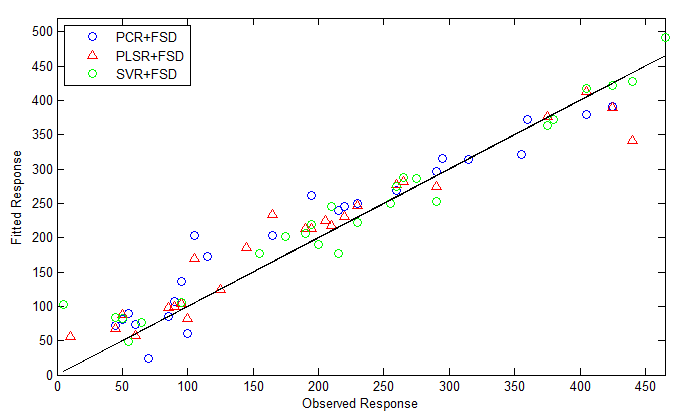
**Figure 5.3: Comparison of PCR and PLSR for different factors**

The linear models the PCR and the PLSR are developed using 10 factors and the mean square prediction error was calculated for the models and are shown in Figure 5.3. The results show that the PLSR model requires minimum 4 latent variables whereas the PCR method needs 5 principal components to build the model.

****

**Figure 5.4: Comparison of the SVR model with the PCR and the PLSR methods without pre-processing**

Figure 5.4 shows the prediction performance of the linear and non-linear methods when no pre-processing is applied. The results show that the non-linear regression model performs better in comparison to both the linear models. The RMSEP of 41.44mg/dL is observed using the SVR method whereas the RMSEP of 43.53 mg/dL and 47.62 mg/dL were observed for PLSR and PCR methods respectively.



**Figure 5.5: Comparison of SVR, PLSR and PCR methods using FSD pre-processing method**

The FSD pre-processing is applied on all the calibration models and the results were shown in Figure 5.5. The FSD combined with SVR has shown to improve better prediction in comparison to the FSD combined with the PCR and the PLSR under the same conditions.

Table 5.1 shows the summary of the results and the corresponding optimum parameters used for the models.

**Table 5.1: summary of results for PCR, PLSR and SVR**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Pre-processing** | **Optimum parameters** | **SEP (in mg/dL)** |
| **PCR** | **None** | **5 PCs** | **47.62** |
| **PLSR** | **None** | **4LVs** | **43.53** |
| **SVR** | **None** | **C=50,000, ɛ = 0.05** | **41.14** |
| **PCR** | **FSD** | **hh=0.0011f, tp=0.5f** | **36.92** |
| **PLSR** | **FSD** | **hh=0.0006f, tp=0.047f** | **33.99** |
| **SVR** | **FSD** | **hh=0.0001f, tp=0.1f, C=4570, ɛ=0.5** | **28.92** |

# **Chapter 6: Fourier Self Deconvoluted RReliefF**

In this chapter, a novel pre-processing method, Fourier Self Deconvoluted RReliefF (FSDR) that is based on coupling Fourier Self Deconvolution (FSD) with the Regressional Relief-F (RReliefF) processing is proposed for the quantitative analysis of glucose in Near Infrared (NIR) spectroscopy. The RReliefF is applied as a feature weighting algorithm and the FSD is used to remove both the high frequency noise and the baseline variations from the input raw spectra. The conventional PLSR was used as the calibration model in this study. The proposed FSDR-PLSR model has been validated to predict the concentration of glucose from NIR spectra of an aqueous mixture composed of urea, traicetin and glucose in a phosphate buffer solution. The concentraions of individual components in the mixture were chosen in such a way that span their physiological range in blood.

**6.1 Introduction:**

NIR spectroscopy has been recognized as one of the optimal techniques for non-invasive glucose measurement as it provides a reasonable signal to noise ratio and requires little or no sample preparation [107, 108]. However, the prediction of glucose concentration from NIR spectra of a mixture remains a challenge. The PLSR is the most commonly used calibration model for the quantitative analysis of glucose from NIR spectra [14, 15].

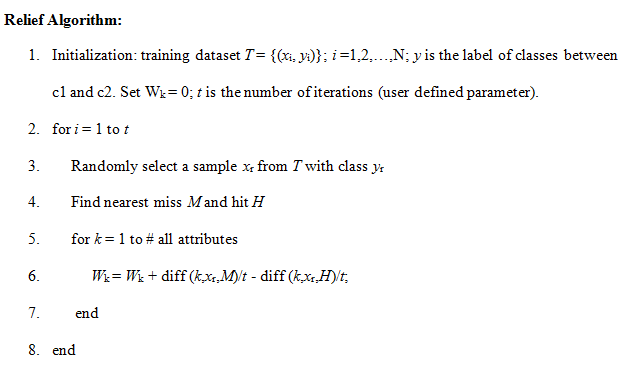
The NIR data collected from a spectrometer might be affected by instrumental noise and baseline variations. So, the spectra need to be pre-processed before applying to the calibration model. The most commonly used pre-processing technique in NIR spectroscopy is taking the derivative of the spectra as it has the capability to remove both the additive and multiplicative noise effects in the raw data [65]. However, the derivative pre-processing can only remove the baseline variations in the spectra. Hence, there is a need to investigate a robust pre-processing technique for the quantitative analysis of NIR spectra.

Partial Least Squares (PLS) is a well known supervised linear feature extraction method [109, 110], but the feature selection or extraction becomes critical when there are large numbers of irrelevant features [78]. Hence applying a feature weighting method before the PLS analysis on the input data would be advantageous. Basically, in feature weighting methods, a real number is assigned to each feature to indicate its relevance to describing the target concept [111]. The Regressional Relief-F (RReliefF) feature weighting technique is robust, noise tolerant and identified as one of the most successful feature weighting algorithms to assess the quality of features due to its effectiveness and simplicity [112, 113]; Thus, a new model called RReliefF-PLSR is introduced in this work, to improve the prediction performance of the PLSR model. However, Pre-processing of the raw data is still required to remove both the baseline variations and the noise from the spectra. Fourier Self Deconvolution (FSD) was used as a pre-processing technique for IR and NIR spectroscopy and shown that it can effectively eliminate noise from the spectra [70, 71]. Hence, additionally in this work, FSD is combined with RReliefF as a new pre-processing technique for PLSR called Frequency Self Deconvoluted RReliefF (FSDR) to enable further improvements in the prediction performance of the model in the context of the determination of glucose concentration from NIR spectra.

**6.2 The RReliefF**

The RReliefF algorithm was proposed by RobnikSikonja et al in [113] which is an extension of Relief and ReliefF algorithms. The Relief algorithm was proposed by Kira and Rendell to assess the quality of features for classification problems [114]. However, the Relief method fails to remove the redundant features and is limited to two classes problems [78]. Later, Kononenko [115] developed an extension to Relief called ReliefF that deals with multiclass problems and able to tolerate noisy data.

For a training dataset *T*= {(*x*i,*y*i with *N* samples and the known class label yi, the Relief algorithm randomly selects a sample (*x*r) and find its two nearest neighbors using Euclidean distance; one from the same class called hit (H) and the other one from the different class called miss (M). The Relief algorithm updates the weight vector based on the values of *M*, *H* and *x*r as shown in the Pseudo code 1 below [92, 93]. The same procedure is repeated for *t* times; where t is a user defined parameter.



**Pseudo code 1: Pseudo code for the basic Relief algorithm [114, 115]**

The function diff (*k*, *s*1, *s*2) performs the difference between the values of attributes for two samples *s*1 and *s*2; and is defined as shown in equation 6.1.

diff (*k*, *s*1, *s*2) **=** (6.1)

It is assumed that *W*= [*w*1, *w*2 … *w*a] be the weight vector for *a* variables in the original space, calculated by the Relief algorithm.

The Relief’s weight estimate *W* [k] of the quality of attribute *k* is an approximation of the difference of the probabilities as shown in equation 6.2.

*W* [*K*] = Probability (different value of *k*| nearest sample from different class)

– Probability (different value of *k*| nearest sample from same class) (6.2)

The main difference between Relief and ReliefF is that Relief uses 1 nearest pair of hit and miss to calculate the weight vector *W*; whereas ReliefF uses k-nearest pairs of hits and misses and averages their contribution to find the weight vector *W*.

Robnik Sikonja et al introduced RReliefF concept for regression problems [113]. The pseudo code of the RReliefF algorithm is shown in Pseudo code 2 below [116]. In each iteration, the RReliefF algorithm randomly selects a sample from the training set and computes its K nearest neighbors. For each neighboring point, it computes three probabilities; PdA, PdP, PdA&dP and are defined as shown in equations 6.3, 6.4 and 6.5 below.

PdA = Probability (different value of attribute |nearest neighbor) (6.3)

PdP = Probability (different value of prediction |nearest neighbor) (6.4)

PdA&dP = Probability (different value of prediction |different value of attribute and nearest neighbor) (6.5)

The RReliefF algorithm contributes all the three probabilities to compute the weight of each feature.

Using Bayes rule, equation 6.2 can be written as

*W* [*k*] = (6.6)

The weights for different attribute, different prediction and different attribute and prediction are assigned to respectively. Finally, the weight vector is computed as shown in lines 13, 14 and 15 of Pseudo code 2.



**Pseudo code 2**: **Pseudo code of the RReliefF algorithm [116]**

The term *d* (*i*, *j*) (lines 6, 8 and 9) is used to calculate the distance between random instance (*R*i) and jth neighbor (*I*j) and is defined in equation 6.7 below.

*d* (i,j) **=** (6.7)

(6.8)

Where rank (*R*i, *I*j) is the rank of Ij in a group of instances arranged by distance from *R*i and *𝜖* is a user defined variable.

**6.3The RReliefF processing*:***

The RReliefF processing involves 2 inter connected steps [117]; the relevance analysis and the variance adjustment.

### **6.3.1The relevance analysis:**

In the relevance analysis, the weights of the features are calculated using the RReliefF feature weighting algorithm as illustrated in Pseudo code 2, for a weight vector *W*, *W*= [,,….], with *t* variables in the original space. As the weights refer to the level of relevance, the feature with the highest weight is considered as the most relevant feature. The features with the weight greater than the defined threshold are subject to the second step. So, the weight vector changes as shown below:

(6.9)

where *Th* is the threshold defined by the user. The threshold *Th* should be positive (0 < *Th* < *w*max) to eliminate the irrelevant features, whose weights are either negative or close to zero. In this work the threshold *Th* is selected to be where is the highest weight in the weight vector calculated by the RReliefF feature weighting algorithm. After removing all the irrelevant features, the relevant features with the weights greater than the threshold (*Th*) are subject to the variance adjustment step.

### **6.3.2 The variance adjustment:**

In the variance adjustment step, the weight vector computed from the relevance analysis is used and the variances of the features are changed in such a way that the most relevant feature becomes the most important feature for PLS analysis. The adjusted or the new variance of ith feature is computed as shown in equation 6.10.

= *r* - (-) (*r*-*z* (*i*)) (6.10)

where ‘*r*’ is the number of features with the weights greater than the threshold *T*, is the maximum weight of the relevant features, *z*(*i*) is the rank given to the weight of the ith feature (*r* is most importance and 1 is least importance) and is the weight of the ith feature. So, to change the variance of ith feature to, the values of features should be multiplied by a specified number ‘*n*’, which is calculated as shown below:

The variance of the features, when multiplied by a specified number ‘n’ becomes as shown in equation 6.11.

= (6.11)

equation 6.11 can also be written as:

(N-1)= (6.12)

From the above equation, the value of *n* can be calculated as shown in equation 6.13.

(6.13)

Hence the raw spectra should be multiplied by the number ‘*n*’ to change the variance of the features in accordance with the relevance of the feature with the target concept. After the variance adjustment step, the most important feature shows the highest variance, whereas the lowest variance is represented by least important feature. Then, the PLSR is performed on the modified spectra.

## **6.4 Fourier Self Deconvolution:**

Fourier Self Deconvolution (FSD) was proposed by Kauppinen et al, and was used for pre-processing of IR spectra. The measured spectra are treated as the convolution of the actual spectra and the line shape function of the instrument [70]. To remove the effect of the line shape function and to get back the original spectra, the reverse operation of deconvolution needs to be performed. In general, deconvolution can be performed either in the time domain or in frequency domain. Deconvolution in time domain is often slow, complex and requires high memory storage [106], whereas performing the deconvolution in the frequency domain tends to be faster and simpler as convolution in time domain is equivalent to multiplication in the frequency domain.

In the NIR range, the line shape function can be expressed by the Lorentzian function [71] as defined in equation 6.14.

(λ) = (6.14) where is the half of the Full Width at Half Height (FWHH) of the ith band, is the centre frequency of the ith band and is the peak amplitude of the ith band.

In FSD, the raw spectra are converted into frequency domain using a Fast Fourier Transformation (FFT) and are divided by FFT of the line shape function of the instrument. However, deconvolution in frequency domain tends to increase high frequency components of the near infrared spectra along with the high frequency components of the noise that degrades the SNR of the spectra [71]. So, the deconvolved spectra need to be truncated using apodization function to compensate the SNR degradation. This can be done by multiplying the deconvolved spectra with the apodization window. An Inverse Fast Fourier Transformation (IFFT) is then performed on the output of the multiplier to get the actual spectra.

## **6.5 Fourier Self Deconvoluted RRelifF (FSDR) as a Pre-processing Method:**

In this proposed model, the FSD is combined with the RReliefF processing method as shown in Figure 6.1 as a pre-processing step to PLSR.

Raw spectra

FSD

FFT

Divider

Line shape function

IFFT

Apodization window

FFT

Actual spectra

Prediction error

RReliefF

Features with modified variance

PLSR

Variance adjustment

Relevance analysis

Relevant features

**Figure 6.1: Block diagram of the proposed FSDR technique for the PLSR model**

Initially both the input raw spectra and the line shape function of the instrument are converted into frequency domain by applying the FFT and are given as inputs to a divider. In practice, deconvolution degrades the SNR of the spectra. The deconvolved spectra are then multiplied with the apodization window to compensate the SNR degradation. The effect of the line shape function is removed by applying an IFFT on the output of the multiplier to get the actual spectra. Relevance analysis is performed on the actual spectra to differentiate relevant and irrelevant features in the spectra by calculating the feature weight vector. The weight vector is calculated using the RReliefF feature weighting method [117]. The features with the weight greater than a defined threshold (see section 6.3.1) are subjected to the next step of variance adjustment. In variance adjustment, the variances of the features are adjusted according to the relevance ranking levels of the features. The high-ranking features become features with higher variance. Finally, pre-processed spectra using the FSD and RReliefF are then applied to the PLSR calibration model.

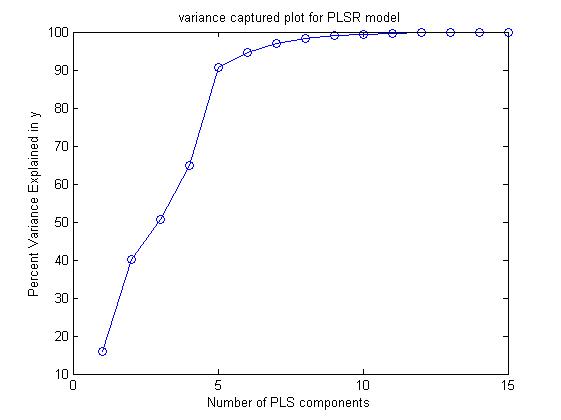
## **6.6 Experiments:**

In this experiment, samples are prepared and NIR spectra are collected as discussed in Chapter 3 (see section 3.1). The individual component concentrations were selected such that these cover their physiological range in blood. The ranges of the glucose, triacetin and urea concentrations of the prepared samples are from 20 to 500mg/dl, 10 to 190mg/dl and 0 to 50mg/dl respectively. The collected NIR spectra spanned the spectral region from 2100 nm to 2400 nm with a spectral resolution of 1nm.

The collected raw spectra are divided into training and testing data subsets using Kennard stone algorithm [96] such that the training set consisting of 60 spectra that are three replicate spectra of 20 samples and the test set has 30 spectra with three replicate spectra of 10 samples.

## **6.7 Discussion of experimental results and comparisons:**

Initially, the PLSR model was implemented without pre-processing and the variance captured by each latent variable is shown in Figure 6.2. The x-axis shows the number of PLS components and the y-axis represents the percentage of variance captured. Figure 6.2 shows that 95% of the variance is captured by 6 Latent Variables (LVs). The optimum LVs for the PLSR model were found by using the 10-fold cross validation [118].

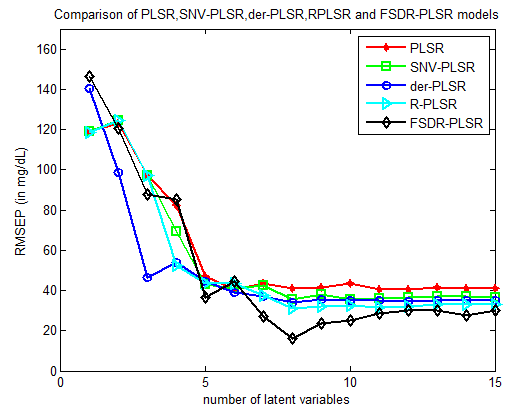
****

**Figure 6.2: Variance captured plot for the PLSR model**

In this work, the libPLS (an integrated library for Partial Least Squares Regression and Discriminant Analysis) toolbox [119] (<http://www.libpls.net/>) was used to perform the PLSR [57]. The PLSR model is then pre-processed using the RReliefF method. The test set was pre-processed using the parameters obtained in the training phase.

Next, the FSD is performed on the RPLSR model to further improve the prediction performance of the PLSR model. The grid search optimization [8] is used to find the optimum parameters of the FSD model. The two parameters of the FSD model; the truncation frequency (*t*p) and the FWHH are varied from 0.001 f to 0.1 f and from 1 cm-1 to 50 cm-1 respectively. The optimum values of FWHH and truncation frequency are found to be 7.5 cm-1 and 0.04 f respectively; where f is the normalized frequency.

The proposed model is further compared with the most common pre-processing methods namely the standard normal variate and first derivative [120]. Figure 6.3 shows the comparison of PLSR model with different pre-processing methods. The results show that the proposed model performs much better in comparison to other pre-processing methods, when intergrated with the PLSR model. The summarized results are given in table 6.1.

****

**Figure 6.3: Comparison of PLSR model with different pre-processing methods**

The Clarke Error Grid Analysis (EGA) is used to quantify the clinical accuracy of predicted glucose concentration in comparison to the reference glucose concentration. The EGA is identified as one of the golden standards to measure the accuracy of blood glucose meters.

In EGA plot, the predicted glucose concentrations by the technique under test are shown on y-axis and the reference glucose concentration values are displayed on x-axis. The diagonal line indicates the ideal fit between the two and the points above and below the line represents underestimation and overestimation of the actual values respectively. Zone A (admissible) indicates the predicted glucose values that are deviated from the actual values by ±20%. Zone B lies above and below the zone A and indicates the points that are outside the ±20% but it is still clinically acceptable. The glucose values that fall in zones C, D and E are clinically inadmissible.

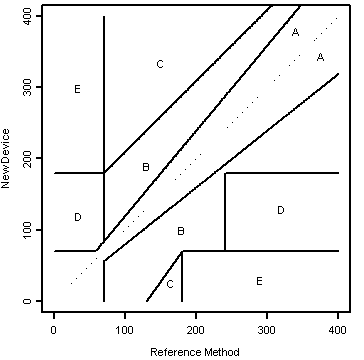
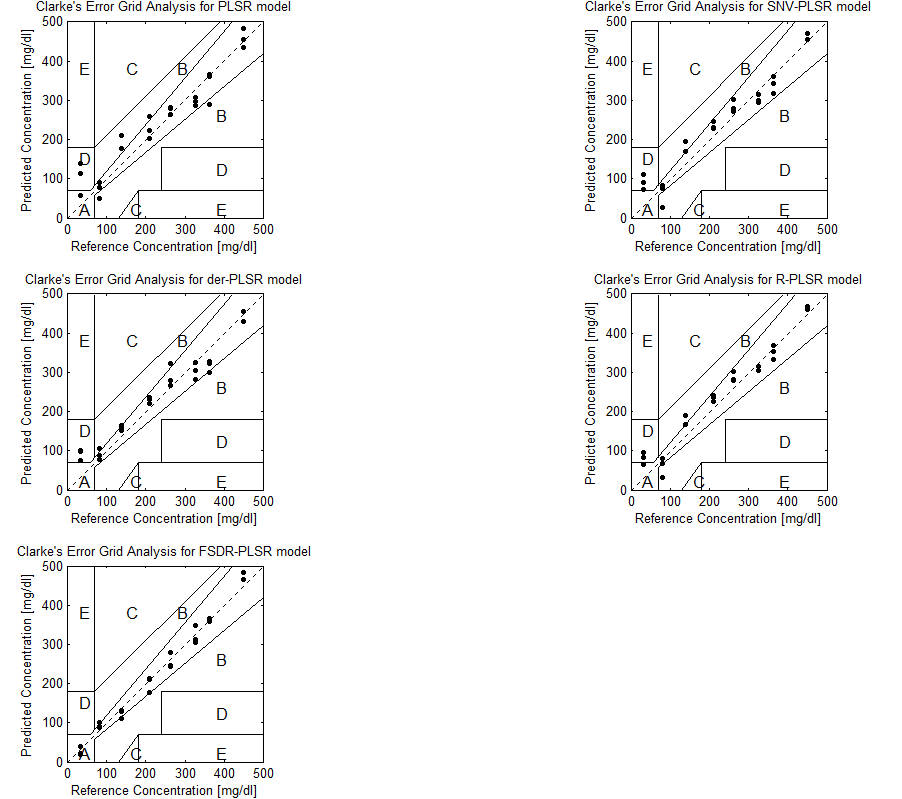


Figure 6.4 shows that all the glucose values predicted by the PLSR with the proposed pre-processing technique fall into Zone A, whereas the values predicted by PLSR model with first derivative, standard normal variate and RReliefF pre-processing methods fall into Zones B and D.

****

**Figure 6.4: Clarke Error Grid Analysis plots for PLSR model with different pre-processing methods**

**Table 6.1: Comparison of PLSR model with different pre-processing methods**

**\*** indicates that the units are in mg/dL.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Pre-processing** | **RMSEP\*** | **RMSECV\*** | **Optimum parameters** |
| PLSR  PLSR  PLSR  PLSR  PLSR | None  SNV  1st derivative  RReliefF  FSDR | 40.30  35.48  34.10  31.30  16.39 | 41.85  39.16  35.93  35.47  31.14 | 6LVs  8LVs  8LVs  8LVs  8LVs, FWHH =7.5 cm-1;  truncation frequency (*t*p) = 0.04 f; |

# **Chapter 7: NIR spectroscopy using human serum albumin data**

In this chapter, two novel pre-processing methods namely Savitzky-Golay coupled with digital bandpass filtering, Hilbert haung transform based pre-processing and one novel calibration model called FSD-ISVR (Fourier self-deconvoluted improved support vector regression) are proposed to determine the concentration of glucose from an aqueous mixture of human serum albumin and glucose in a PBS.

## **7.1 SAVITZKY-Golay coupled with digital bandpass filtering**

A novel pre-treatment technique based on coupling Savitzky-Golay with digital bandpass filtering has been introduced to enhance the prediction capability of the linear regression models the PLSR and the PCR models in NIR spectroscopy. pre-processing method based on combining bandpass with Savitzky-Golay filtering to further improve the prediction performance of the linear calibration models the PCR and the PLSR in NIR spectroscopy. The proposed method is compared to the highly efficient RReliefF pre-processing technique for further evaluation. The developed calibration models have been validated to predict the glucose concentration from near infrared spectra of a mixture of glucose and human serum albumin in a PBS.

### **7.1.1 INTRODUCTION**

Non-invasive blood glucose monitoring is seen as an important development that could potentially lead to improved diabetic control and quality of life. Promising non-invasive approaches include near infrared (NIR) spectroscopy, photo acoustic infrared spectroscopy, Raman infrared spectroscopy and mid infrared spectroscopy. In this study, NIR spectroscopy is chosen as it is identified as one of the optimal non-invasive techniques for the measurement of blood glucose levels. However, the determination of concentration of a single analyte (glucose) from NIR spectra of a mixture remains a challenge. The most commonly used regression models for the quantitative analysis of glucose in NIR spectra are the Principal Component Regression (PCR) and the Partial Least Squares Regression (PLSR) [13, 14]. However, the raw spectra are affected by baseline variations and instrumental noise. Hence there is a need for the data to be pre-processed before applying the regression model [121].

Bandpass filtering is one of the most efficient ways of linear pre-processing in NIR spectroscopy as it can deal with both the baseline variations and the high frequency noise components of the spectra. However, matching the filter parameters to the raw data and the regression techniques used to preserve the important features remains a challenge [65]. An optimal bandpass technique based on combining the filtering method with another pre-processing stage needs to be investigated. As Smoothing technique can eliminate the unwanted noise components and also preserves the peaks in the spectra, it is chosen to pre-process the raw data before applying it to the filtering analysis. Two blocks of pre-processing methods could be more advantageous. Savitzky-Golay (SG) smoothing is a most commonly used pre-processing technique which can eliminate the unwanted noises such as tilt, baseline-drift, reverse and so forth [122]. Hence in this chapter, a novel pre-processing method based on combining the SG smoothing and Bandpass filtering is proposed to improve the prediction performance of the linear models.

Feature weighting methods are used to eliminate the irrelevant features in the raw spectra and can be adopted in pre-processing as a class on non-linear techniques. The Regressional Relief-F (RReliefF) is identified as one of the robust and most successful feature weighting algorithms due to its simplicity and effectiveness. RReliefF pre-processing has proven to be one of the most effective pre-processing methods in NIR spectroscopy to eliminate the irrelevant features from both low and high frequency components and therefore help improve the prediction capability of linear regression models [78]. However, such techniques tend to be complex to implement and difficult to adapt to different noise environments.

In this chapter, a novel linear pre-processing technique is proposed that combines a linear ‘feature preserving’ smoothing technique (Savitzky-Golay) with bandpass filtering to improve the prediction performance of PCR and PLSR. The technique is evaluated against the non-linear RRelief method. Corresponding models are developed, tested and evaluated using experiments to predict the glucose concentration in a mixture of glucose and human serum albumin in a phosphate buffer solution. The results obtained show not only that the proposed technique improves the prediction performance of both PCR and PLSR, but also achieves better prediction performance than RReliefF.

### **7.1.2 The RReliefF**

Robnik Sikonja et al proposed the RReliefF algorithm [113] which is an extension of Relief and ReliefF algorithms. Kira and Rendell proposed the basic Relief algorithm to assess the quality of features for classification problems [111]. However, the Relief method is limited to two classes problems and fails to eliminate the redundant features in the spectra [78]. Later, the extension of Relief called ReliefF was proposed by Kononenko [123] that deals with multiclass problems and also able to tolerate noisy data.

### **7.1.3 The RReliefF processing**

The RReliefF processing consists of 2 inter related steps [78]; the relevance analysis and the variance adjustment. In the relevance analysis, the RReliefF algorithm is used to calculate the weights of the features [124], for a weight vector W, *W*= [1, 2, 3, ….], with *m* variables in the original space. As the level of relevance is referred by weights, the feature corresponding to highest weight is identified as the most relevant feature corresponding to the target concept. Features with the weight lesser than the defined threshold are removed and the remaining features are subject to the second step. So, the weight vector changes as shown in equation 7.1 below:

(7.1)

Where *T* is the threshold (user defined parameter). To eliminate the irrelevant features whose weights are either close to zero or negative, the threshold *T* should be positive (0 <<). In this work, the threshold is chosen as/2; where isthe highest weight in the weight vector computed by the RReliefF algorithm.

In the variance adjustment step, the weight vector computed from equation 1 is used and the variances of the features are modified so that the most relevant feature becomes the most important feature for the linear prediction analysis. The new variance or the modified variance of the feature is calculated as shown in equation 7.2.  
(7.2)

where ‘*m*’ is the number of features with the weights greater than the value of threshold*,* is the highest weight of the relevant features, is the weight of the feature and is the rank given to the weight of the feature (1 is least importance and is most importance). In order to change the variance of feature to, the values of features must be multiplied by a specified number ‘’, which is calculated as shown in equations 4 and 5. When the values of features are multiplied by a specified number then the variance of the features becomes as shown in equation 7.3.

(7.3)   
From equation 7.3, the value of p can be calculated as shown below:

(7.4) So, the input spectra must be multiplied by the number ‘’ to modify the variance of the features corresponding to their relevance level with the target concept. Finally, the PCR and the PLSR analysis are done on the modified spectra.

### **7.1.4 The digital bandpass filtering**

Digital bandpass filtering is one of the well-known pre-processing techniques in NIR spectroscopy which is used to eliminate the baseline variations and the high frequency components in the raw spectra. Bandwidth and centre frequency are the main factors defining the digital bandpass filter and both should be optimized to find the required band that has maximum information about the glucose concentration. The PCR and the PLSR regression models should be combined with the design of the digital bandpass filter because the high frequency noise and baseline variation do not behave according to a known pattern. As a start the raw spectra obtained from the chemical mixture are filtered using digital bandpass filter with pre-calculated parameters then the PCR and the PLSR models are built using the filtered spectra with a specific number of Principal Components (PC) and Latent Variables (LV).

**Chebyshev Filter:** Defined by its transfer function, the Chebyshev filter is a recursive filter that shows better performance than other filters. The output of the Chebyshev filter depends on past output, past input and present input values. The performance of the filtering operation is in the time domain. The filter order and the location and magnitude of the ripple, control the Chebyshev filter characteristics. The filter ripple can be in the passband (cheby10) or in the stopband (cheby2), this thesis will concentrate on the passband only to enhance the prediction of glucose concentration [125].

### **7.1.5 Savitzky-Golay Filtering**

Savitzky and Golay (SG) [126] introduced a popular method for vector numerical derivation including a smoothing step. At any center point the derivative can be obtained by fitting a polynomial in a symmetric window on the raw data and it is easy to find the analytical derivative of any order, if the polynomial parameters are calculated then it can be used as a sequential estimate of the derivative at this centre point.

To calculate the window size and the degree of the fitted polynomial it is required to decide on the number of used points, the Savitzky-Golay derivation originally used a symmetric window smoothing which requires an equal number of data points in both sides of the centre point and as a result the technique discards points on both ends of the spectra during the pre-processing [126].

### **7.1.6 Experiments (Human serum albumin data)**

The collected raw spectra are divided into training and testing data subsets using Kennard stone algorithm [96] such that the training set consisting of 75 spectra samples and the test set has 25 spectra samples.

### **7.1.7 Discussion of experimental results**

Initially, the PCR and the PLSR models were developed without pre-processing and the variance captured by each LV or PC is Figure 5.1.1. The x-axis represents the number of PLSR components in blue and PCR components in red and the y-axis shows the percentage of variance captured by the PCs or LVs. From Figure 7.1.1, it is observed that98% of the variance is captured by 10 PCs, whereas95% of the variance is captured by 5 LVs. 10-fold cross validation was used to find the optimum number of PCs and LVs for the PCR and the PLSR models respectively [118].

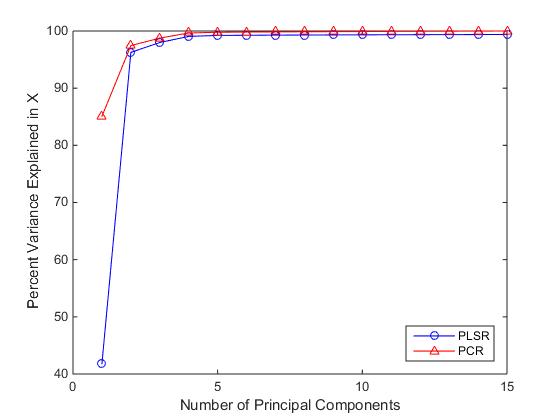


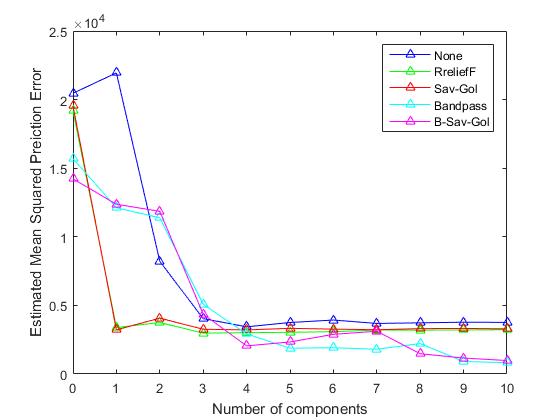
Fig 7.1.1 variance captured plot for PCR and PLSR models

In this work, the PLSR model was implemented using the libPLS (an integrated library for Partial Least Squares Regression and Discriminant Analysis) toolbox [119] (http://www.libpls.net/).

To improve the prediction accuracy of the PCR and the PLSR models, the RReliefF and bandpass filtering methods are combined with the calibration model. The parameters that affect the Chebyshev digital bandpass filtering are the bandwidth (w), centre frequency (c), the filter order (N) and the ripple (R).

The calibration models were implemented with the proposed pre-processing technique based on combining the Savitzky-Golay coupling with the bandpass filtering. Both the PLSR and PCR models were developed with different pre-treatment methods and the obtained results are shown in Table 7.1.1

Also, the mean squared error plots for both PCR and PLSR were shown in Figure 7.1.2 and Figure 7.1.3 respectively.

Table 7.1.1 shows that the glucose prediction is enhanced by using the PCR and PLSR with different pre-processing techniques and the experimental results show the ability of 

7.1.2 Mean squared prediction error plot for PCR with different pre-processing techniques

the PLSR to predict the same Root Mean Square Error of Prediction (RMSEP) value with latent values50% less than the PCR i,e better prediction for the same no of LVs. The coupling of Savitzky-Golay with Chebyshev has improved the prediction accuracy by 41.67% for PCR and 39.43% for PLSR when no pre-processing is applied and the proposed technique has also improved the prediction accuracy by 32.14% for PCR and 30.47% for PLSR in comparison to both the models with RReliefF processing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Pre-processing |  | RMSEP | Factors |
| PCR | None | 0.97 | 47.7 | None |
| PCR | Savitzky-Golay | 0.97 | 43 | deriv=0  order=2 |
| PCR | RReliefF | 0.97 | 41 | K=13 |
| PCR | Chebyshev filter | 0.99 | 30.2 | c=0.005f  w=0.0035f  N=2  R=0.5dB |
| PCR | Savitzky-golay with Chebyshev | 0.99 | 27.8 | c=0.007f w=0.005f R=1.4dB   N=2 deriv=0  order=2 |
| PLSR | None | 0.91 | 47.7 | None |
| PLSR | Savitzky-golay | 0.97 | 43.7 | deriv=0  order=2 |
| PLSR | RReliefF | 0.96 | 41.5 | K=3 |
| PLSR | Chebyshev filter | 0.98 | 31.9 | c=0.0093f  w=0.0051f  R= 2.3dB  N=1 |
| PLSR | Savitzky-golay with Chebyshev | 0.99 | 28.9 | c=0.0192f w=0.0096f  R=1.3dB N=1 |

Table 7.1.1 summary of results for PCR and PLSR with different pre-processing techniques

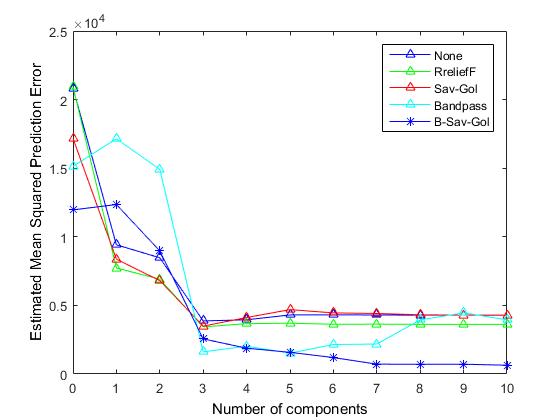


Fig 7.1.3 Mean squared prediction error plot for PLSR with different pre-processing techniques

## **7.2 Hilbert Haung Transformation pre-processing**

## 

In this chapter a novel calibration model based on combining the Hilbert Haung Transformation pre-processing and the Partial Least Squares Regression is also proposed for the quantitative analysis of glucose in near infrared spectra. The pre-processing involves two steps; the Empirical Mode Decomposition and the Hilbert Spectral Analysis. The proposed method was evaluated to predict the glucose concentration from two different mixtures; 1. Aqueous mixture of urea, glucose and triacetin in a phosphate buffer solution. 2. A mixture of human serum albumin and glucose in a phosphate buffer solution. The prediction performance of the proposed calibration model is compared with the PLSR using the first derivative and the standard normal variate pre-treatment methods.

### **7.2.1 Introduction:**

Infrared spectroscopy, near infrared spectroscopy, scatter and polarization changes, photo acoustic spectroscopy and Raman spectroscopy are the optical measurement techniques used in non-invasive technology [36]. Among all the non-invasive optical methods, near infrared spectroscopy is identified as one of the most promising techniques for the non-invasive measurement of blood glucose levels [71].

Partial Least Squares Regression (PLSR) is the most widely used calibration model for the quantitative analysis of glucose in Near Infrared (NIR) spectroscopy [73]. However, the input raw spectrum collected from NIR spectrometer might be complicated by scale shifting, baseline variations and instrumental noise. Hence pre-processing of the input raw spectra is an essential step before applying to the calibration model. The first derivative and standard normal variate methods are the most commonly used pre-processing techniques in NIR spectroscopy to eliminate both the multiplicative and additive noise effects in the raw spectra; however, these methods could remove only baseline variations from the input spectra [71]. Hence, a robust and adaptive pre-processing technique needs to be investigated for the analysis of glucose in NIR spectra.

The effects of scale shifting, noise and baseline variations, make the spectra non-stationary. All the traditional data analysis methods are all based on the assumption that the data is both linear and stationary. Later, few methods have been introduced to analyze non-stationary and non-linear data; for instance, several non-linear time series analysis methods (Tong 1990; Kantz and Schreiber 1997; Diks 1999) were implemented for stationary but non-linear systems [127-129]. Also, Wagner- Ville distribution and wavelet analysis (Gr¨ochenig 2001: Flandrin 1999) were implemented for linear but non-stationary data [130, 131]. However, in most of the realistic systems, either man-made or natural ones, the data are mostly expected to be both non-stationary and non-linear. An important condition to represent non-stationary and non-linear data is to have adaptive basis. In signal processing, the Hilbert Haung Transformation (HHT) has proven to be successful for the analysis of non-stationary, natural and non-linear signals [132]. HHT has been tested and validated completely and the studies confirm that method show results much sharper than the traditional data-analysis methods. Also, HHT provided true physical representation in most of the data inspected [133]. However, to our knowledge no work to date has attempted to apply HHT to NIR spectroscopy. In this thesis, it is proposed to adopt the HHT as a pre-processing step in the quantitative analysis of NIR spectra. HHT is based on the Empirical Mode Decomposition (EMD); EMD splits any complex signal into a set of orthogonal bases called Intrinsic Mode Functions (IMFs). Each IMF has unique information about the signal and the rate of redundancy is zero. However, each IMF might be mitigated due to several unwanted variations such as baseline variations, high frequency noise and scale shifting. Hence the filtering operation on each IMF is needed before applying to the calibration model.

Thus, a novel calibration model based on combining HHT based pre-processing with the PLSR is developed in this thesis and evaluated using actual experimental data based on 2 different mixtures to predict the glucose concentration in these mixtures. The benefit of HHT is non-stationary and the data is decomposed into many non-identical trends which separate the noises from the signals. It could also eliminate more non-stationary noises than stationary methods. The main disadvantage of HHT method is that calculation time is much longer than stationary methods. Models based on the most common pre-processing techniques (Derivative and Standard Normal Variate pre-processing) are also developed with PLSR regression for comparison; the results obtained show that the HHT-based model outperforms in terms of prediction performance of glucose from the NIR spectra of the two mixtures.

### **7.2.2 Hilbert Haung Transformation (HHT)**

In Signal processing, the Hilbert Haung Transformation (HHT) is often used for the analysis of natural, non-stationary and non-linear signals. HHT is generally used to define trends in a spectrum. In most of the studies, we defined the trend that represents the noise and the baseline as a straight line and it is then fitted to the spectrum. To get the zero-mean residue, we remove the straight line. However, such linear trends are not applicable for the real-world applications and non-linear data. Noise is non-stationary and exists non-linearly. when we try to recover the spectra, the line is non-stationary and non-linear in practice. HHT can provide physically meaningful representation of data from both non-stationary and non-linear processes as its basis of expansion is adaptive. HHT comprises of two steps; Empirical Mode Decomposition (EMD) and the Hilbert Spectral Analysis (HSA).

### **7.2.3 Empirical Mode Decomposition (EMD)**

Empirical Mode decomposition is the main feature of HHT, in which any complicated data can be decomposed into a finite number of components.

The EMD splits a signal into a set of orthogonal bases of the same length called Intrinsic Mode Functions (IMF) and the procedure of decomposition is called sifting process [134-136]. Each intrinsic mode constitutes a simple oscillation which will have same number of zero crossings and extrema. Additionally, the oscillation will be symmetric in respect of the local mean.

An IMF should satisfy the following two conditions.

1. The number of zero crossings and the number of extrema must either be equal or differ by at most one.
2. The mean value of the envelope defined by the local minima and the envelope defined by the local maxima at any point is equal to zero.

In general, the number of IMFs is equal to log2N; where N represents the total number of data points. Also, the original data can be represented by the sum of all IMFs.

**EMD procedure**:

If the original signal is *X (t),* the EMD process can be summarized as follows:

1. Find local maxima and local minima in the signal
2. Connect all the local maxima by a cubic spline to get the upper envelope *Emax (t)* and lower envelope *Emin (t)* is obtained by connecting all the local minima.
3. Compute the mean *m1* between both the envelopes and subtract it from the actual signal to get the residue *h1* as shown in equations 7.2.1 and 7.2.2.

(7.2.1)

(7.2.2)

1. Now assume as a new data and repeat the steps 1 to 3 on to get *h11*.

where *h11**m11*(7.2.3)

1. Repeat the process *p* number of times to get the first component that satisfies the IMF condition.

*h1p*  = *h1 (p-1) - m1p* (7.2.4)

C*1=h1p* is the first component and the residue is r1.

*r1= X (t) – C1* (7.2.5)

*r2= r1-c2* (7.2.6)

*rn= rn-1 – Cn* (7.2.7)

Repeat the steps 1 to 5 to get the second component and in this way a total of n components can be extracted from the original signal *X (t).*

The original signal can be expressed as the sum of the IMFs as follows;

*X(t) = + rn*  (7.2.8)

### **7.2.4 Hilbert Spectral Analysis (HSA)**

Hilbert transform on each IMF is applied to get the instantaneous frequency spectrum. Each IMF provides unique information about the absorbance of the analyte of interest. As the raw spectra might be affected by several unwanted variations and each IMF should be filtered to remove the noise effect from the raw spectra. In spectral analysis, a digital filter is used to remove the unwanted variations from each instantaneous frequency spectrum. In this study, Butterworth filter is used as it contains maximally flat frequency response in passband and is simple to implement compared to other filters. Finally, the filtered-out components are combined to get the pre-processed spectra that are applied to the PLSR calibration model.

Derivative pre-processing technique:

In NIR spectroscopy, derivatives are the most commonly used pre-processing method [120].

Finite difference method is the basic method of derivative technique, which is computed as the difference between two successive spectral points as shown below.

(7.2.9)

Where is the absorbance at ith wavelength and is its corresponding derivative.

The NIR spectra are affected by both baseline variations and high frequency noise. However, the derivative pre-processing is effective at eliminating the baseline variations from the raw spectra [120].

**Standard Normal Variate Pre-processing**

Standard Normal Variate (SNV) pre-processing method is widely used to process the scatter data to eliminate the multiplicative effects of scatter [137].

In standard normal variate pre-processing method, each spectrum is centered and scales by its own standard deviation (σ).

SNV transformation on an individual spectrum can be implemented as shown below.

j=1, 2,…,p; (7.2.10)

Where is the absorbance of ith spectrum at jth wavelength and is the mean absorbance value of the ith raw spectrum.

The standard deviation (σ) can be computed as shown below;

σ (7.2.11)

### **7.2.5 Proposed HHT-PLSR model**

Initially the input raw spectra are applied to the EMD process; this step involves the decomposition of input raw spectra into a set of finite orthogonal basis functions called IMFs. Hilbert transform is then performed on each IMF. The noise associated with each IMF is then removed using the Butterworth filter. All the spectra of each IMF are then added using the adder. The pre-processed spectra are then applied to the PLSR calibration model to predict the glucose concentration from the input NIR spectra.

Prediction error

Raw NIR spectra

HSA +

IMF1

PLSR

IMF2

EMD

HSA

IMFn

HSA

Figure 7.2.1: The block diagram of the HHT-PLSR model

### **5.3.8 Discussion of Experimental Results and Comparisons:**

In this study, an integrated library for Partial Least Squares Regression and Discriminant Analysis (libPLS) toolbox(<http://www.libpls.net/>) [119] was used to build the PLSR model.

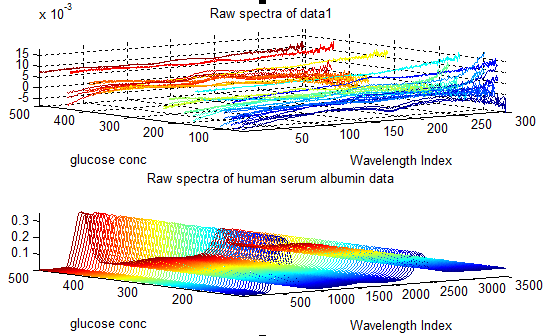


Figure 7.2.2: The raw spectra of data1 and data2 (human serum albumin data).

The raw NIR spectra of both the data1 and the human serum albumin data were shown in Figure 7.2.2 and the graphs show that the human serum albumin data is more informative in comparison to the raw data1 (The peaks are clearly visible in data2).

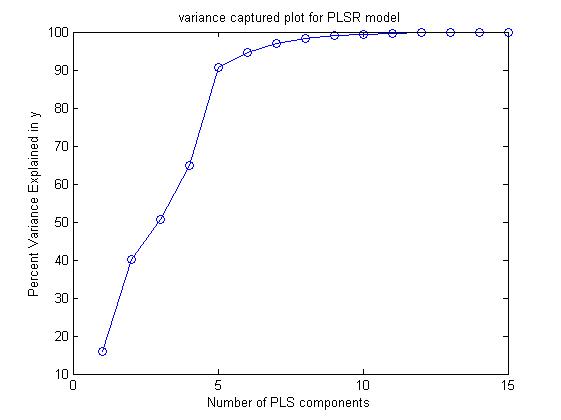


Figure 7.2.3: The variance captured plot for PLSR model using data1.

10-fold cross validation was used to find the optimum number of components for the PLSR model and 95% of total variance is captured by 6 components as shown in Figure 7.2.3.

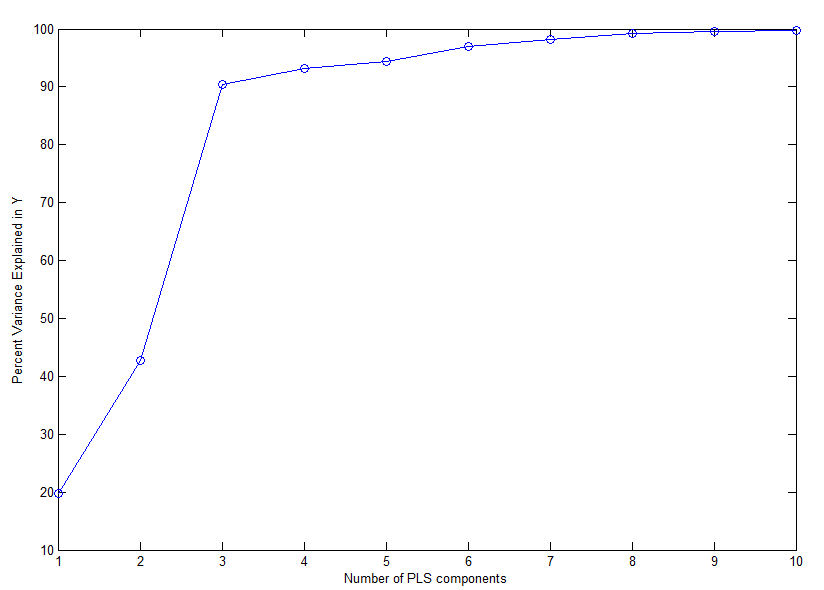


Figure 7.2.4: The variance captured plot for PLSR using human serum data.

Figure 7.2.4 shows that more than 95% of the total variance is captured by 6 components for the human serum data.

Then the PLSR with HHT pre-processing was developed. The comparison of the PLSR model with different pre-processing techniques for data1 is shown in Figure 7.2.5.

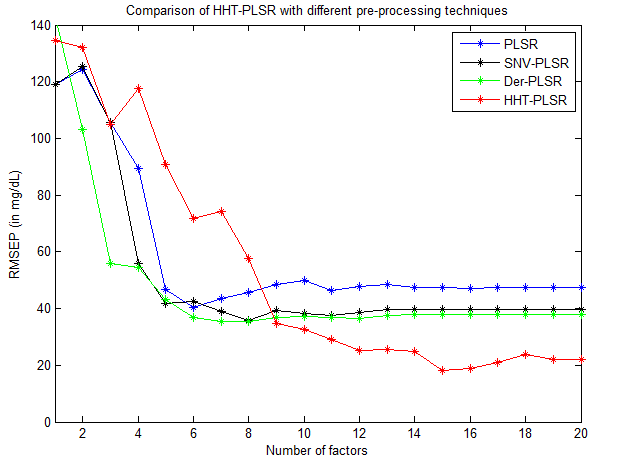
****

Figure 7.2.5: The comparison of PLSR with different pre-processing techniques using data1.

Figure 7.2.5 shows that PLSR with HHT pre-processing performs better prediction accuracy in comparison to PLSR with first derivative and standard normal variate techniques.

The results show that the proposed model offers better prediction results in comparison to the PLSR model with the standard normal variate and the first derivative pre-processing methods. From Figure 5.3.5, it is observed that HHT pre-processing reduces the RMSEP from 40 mg/dl to 18.37 mg/dl for PLSR model.

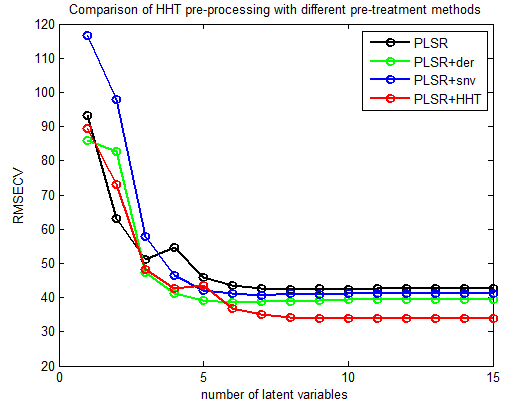


Figure 7.2.6: The comparison of PLSR with different pre-processing techniques using human serum albumin data.

The proposed model was also evaluated on the human serum albumin data and the prediction performance of the PLSR with different pre-treatment methods is shown in Figure 7.2.6.

# **Chapter 8: Conclusions and Future scope**

This chapter presents the conclusions of the research work for this dissertation and the recommendations for the future work are also described.

## **8.1 Summary and Conclusions:**

The research work described in this dissertation has focused on developing non-invasive blood glucose measurement techniques which can facilitate all the diabetic patients worldwide. Diabetic patients need to monitor their blood glucose levels 4-6 times per day for the proper control of the disease. However, the invasive approach is inconvenient and demotivating for many patients to check their glucose levels on a daily basis. These complications can be addressed using the non-invasive glucose measurement. NIR spectroscopy is identified as one of the promising non-invasive glucose measurement techniques due to its non-destructive, painless and non-invasive nature. However, the clinical translation of the technology still remains a challenge. The conventional linear calibration methods used for the analysis of glucose in NIR spectra are the partial least squares regression and the principal component regression models. However, these models may not be the best choice for the analysis of non-linear data. Hence, there is a need for the development of robust non-linear calibration models in chemometrics. In this work, a non-linear calibration model called Local linear embedding regression has been proposed to predict the glucose concentration from the NIR spectra.

The two important parameters that affect the LLER model are the number of nearest neighbouring points (K) and the dimension of the embedded data (d). So, the selection of these parameters is very important to get the optimum prediction results. The important feature of LLER method over the linear calibration methods the PLSR and the PCR is that it preserves the neighbourhood structure of the nearest NIR spectra in the mapped plane. The results discussed in this dissertation has shown that LLER model outperforms PLSR, PCR and SVR models with and without pre-processing. SVR has proven to be robust non-linear calibration model in chemometrics; however, another model is needed every time when a sample is added or removed from the training set specially in some applications such as leave one-out cross validation or online time series prediction. Retraining the model for each training data sample from the scratch could be expensive. Hence another non-linear calibration model called Improved support vector regression coupled with Fourier self-deconvolution has been introduced in this research. The proposed models have been validated to predict the glucose concentration from two different aqueous mixtures; 1. From an aqueous mixture of glucose, triacetin, glucose and urea in a phosphate buffer solution. 2. from an aqueous mixture of human serum albumin and glucose in a phosphate buffer solution. The concentration of the components is selected to be within their physiological range in blood.

As the raw spectra from FTIR spectrometer might be mitigated by several factors such as low frequency baseline variations, high frequency noise and scattering effects. The most commonly used pre-processing methods in NIR spectroscopy are standard normal variate, first derivative and digital filtering techniques. As PLSR model preserves the features that show highest variance in the spectra and eliminate the features that contributes lowest variation to the spectra. The important features might be eliminated due to the decomposition process, if the analyte of interest show lowest variation to the spectra. There is a need for the incorporation of feature weighting method in pre-processing stage. Hence, in this work a novel pre-processing method based on combining FSD with RReliefF feature weighting method has been proposed to improve the prediction performance of the PLSR model. The proposed model combines FSDR and PLSR models, where FSDR decomposes the raw spectra into principal components and the PLSR builds the regression model, which is used for new test data. Grid search optimization is used to find the optimal parameters for the FSDR-PLSR model. The two parameters that affect the Fourier self-deconvoluted RReliefF (FSDR) pre-processing model are the truncation frequency and the Full Width at Half Height (FWHH). Although, the proposed FSDR-PLSR model consumes lot of time in generating the calibration model, it improves prediction capability of the PLSR model. The FSDR-PLSR model has to set three parameters number of latent variables, truncation frequency and FWHH. The results presented in this work confirm that the FSDR model has shown better prediction performance than the first derivative and standard normal variate pre-processing methods.

Bandpass filtering is an effective pre-processing method to remove both the baseline variations that occur in low frequencies and high frequency noise. However, matching the digital filtering parameters to the raw NIR spectra and the calibration models used to preserve the important features remains a challenge. As smoothing techniques preserve the peaks and also eliminate the unwanted noise variations in the spectra, in this work, another pre-processing method based on coupling Savitzky-Golay with bandpass filtering has been proposed to improve the prediction performance of the calibration models. The results described in this thesis confirm that the RReliefF method performs better prediction than Savitzky- Golay alone but combining Chebyshev digital bandpass filtering with Savitzky-Golay outperforms the RReliefF method.

As the raw NIR spectra are affected by noise, scale shifting and baseline variations that make the spectra non-stationary. The Hilbert Haung Transformation (HHT) has proven to be successful for the analysis of natural, non-stationary and non-linear data. Finally, a novel pre-processing method called Hilbert Haung Transformation pre-processing has been introduced in this dissertation to deal with non-stationary data. The results demonstrate that the proposed models yield better prediction results compared to the existing linear calibration models using both the experimental data sets. All the experiments were performed in a non-controlled environment so that the proposed models can effectively eliminate most of the experimental variation.

## **8.2 Future scope**

In this work, Local linear embedding regression coupled with bandpass filtering has been proposed for the quantitative analysis of glucose in NIR spectroscopy. However, the optimum filter parameters computed in this study are based on the analysis of the calibration model. In the future work, an adaptive bandpass filtering can be developed to facilitate the pre-processing stage to be more robust and independent. All the proposed models have been validated to predict the glucose concentration from a mixture of glucose, urea and triacetin in a phosphate buffer solution and from another mixture of glucose and human serum albumin. However, in the future work, all the proposed models can be validated using the real blood plasma or serum using more number of samples.

In NIR spectroscopy, the number of variables is greater than the number of spectra in general. In signal processing, compressive sensing is the optimal solution in cases, where the number of equations is less than the number of unknowns and the compressed regression model can be implemented for the quantitative and qualitative analysis of glucose in NIR spectra.

In the future work, the prediction performance of the calibration models can be further improved by incorporating more robust feature weighting methods in the pre-processing stage.

In the future work, a real time set up for the non-invasive measurement of the blood glucose levels can be developed using a portable spectrometer.

The datasets related to this research area are not publicly available. So, it is recommended that some datasets should be built so that they are accessible to all the research groups involved in non-invasive glucose measurement.

# **References**

1. Mukire J. Wabomba, Gary W. Small, Mark A. Arnold., “Evaluation of Selectivity and Robustness of Near Infrared Glucose Measurements Based on Short Scan Fourier Transform Infrared Interferograms,” Analytical Chimica Acta 490, 325-340,2003.
2. R. W. Waynant, V. M. Chenault, “Overview of Noninvasive Fluid Glucose Measurement Using Optical Techniques to Maintain Glucose Control in Diabetes Mellitus, IEEE. Org/organizations/pubs/newsletters/leos/apr98/overview.htm.
3. Raju Poddar, Joseph Tomas Andrews, Pratyoosh Shukla, Pratima Sen “Noninvasive Glucose Monitoring Techniques: A review and current trends,” Medical Physics, Instrumentation and Detectors eprint arXiv:0810.5755, 10/2008
4. Andrea Tura, Alberto Maran, Giovanni Pacini, “Noninvasive Glucose Monitoring: Assessment of Technologies and Devices According to Quantitative Criteria,” Diabetes Research and Clinical Practice 77, pp16-40,2007
5. Omar S. Khalil, “Spectroscopic and Clinical Aspects of Noninvasive Glucose Measurements,” Clinical chemistry 45:2, pp165-177, 1999.
6. Hoeil Chung Mark A Arnold, Martin Rhiel and David W. Murphammer, “Simultaneous Measurement of Glucose and Glutamine in Aqueous Solutions by Near Infrared Spectroscoping,” Applied Biochemistry and Biotechnology, Vol. 50, page 109- 125, 1995
7. Mutua J. Mattu, Gary W. Small, and Mark A. Arnold, “Determination of Glucose in a Biological Matrix by Multivariate Analysis of Multiple Bandpass Filtered Fourier Transform Near- Infrared Interfeograms,” Ana. Chem., Vol. 69, 22, pp 4695- 4702, 1997
8. Mark A. Arnold and Gary W. Small, “Determination of Physiological Levels of Glucose in an Aqueous Matrix with Digitally Filtered Fourier Transform Near-Infrared Spectra,” Analytical chemistry, American Chemical Society 62:1414, pp 1457-1464, 1990.
9. Alberti, Kurt George Matthew Mayer, and PZ ft Zimmet. "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation." *Diabetic medicine* 15.7 (1998): 539-553.
10. Gabbay, Kenneth H. "Hyperglycemia, polyol metabolism, and complications of diabetes mellitus." *Annual review of medicine*26.1 (1975): 521-536.
11. Kjos, Siri L., and Thomas A. Buchanan. "Gestational diabetes mellitus." *New England journal of medicine* 341.23 (1999): 1749-1756.
12. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. "Continuous glucose monitoring and intensive treatment of type 1 diabetes." *New England Journal of Medicine* 359.14 (2008): 1464-1476.
13. Næs, Tormod, and Harald Martens. "Principal component regression in NIR analysis: viewpoints, background details and selection of components." *Journal of chemometrics* 2.2 (1988): 155-167.
14. Haaland, David M., and Edward V. Thomas. "Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information." *Analytical Chemistry* 60.11 (1988): 1193-1202.
15. Lin Zhang, Gary W. Small and Mark A. Arnold “ calibration standardization algorithm for partial Least Squares Regression: Application to the determination of physiological levels of glucose by Near-Infrared spectroscopy,” Anal.Chem.2002,74,pp 4097-4108.
16. I.T. Jollife, “Principal Component Analysis,” Second Edition, Springer, 2002.
17. R. Kramer, “Chemometrics Techniques for Quantitative Analysis,” Marcel-Dekker, (1998).
18. Wu ZH, Huang NE, Long SR, Peng CK. On the TREND, detrending, and variability of nonlinear and nonstationary time series. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:14889–14894.
19. Faix, O. "Fourier transform infrared spectroscopy." *Methods in lignin chemistry*. Springer, Berlin, Heidelberg, 1992. 83-109.
20. Shichiri, Motoaki, T. Uemura, and K. Nishida. "Non-invasive Fourier transformed infrared spectroscopy for the measurement of submucosal tissue glucose concentration—application of chalcogenide optical fiber system." *IEEE LEOS Newslett* 12 (1998): 14-16.
21. Mendelson, Yitzhak, et al. "Blood glucose measurement by multiple attenuated total reflection and infrared absorption spectroscopy." *IEEE transactions on biomedical engineering*37.5 (1990): 458-465.
22. Malin, S.F., Ruchti, T.L., Blank, T.B., Thennadil, S.N. and Monfre, S.L., 1999. Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy. *Clinical chemistry*, *45*(9), pp.1651-1658.
23. Persky, M. J. "A review of spaceborne infrared Fourier transform spectrometers for remote sensing." *Review of scientific instruments* 66.10 (1995): 4763-4797.
24. Guo, Qiaohan. *Quantitative infrared spectroscopy in challenging environments: applications to passive remote sensing and process monitoring*. IOWA UNIV IOWA CITY, 2012.
25. Thermo Nicolet, Introduction to Fourier Transform Infrared Spectrometry. www.thermonicolet.com.
26. International Diabetes Federation diabetes ATLAS 7th edition <http://www.diabetesatlas.org/>
27. Poddar, Raju, et al. "Non-invasive glucose monitoring techniques: A review and current trends." *arXiv preprint arXiv:0810.5755* (2008).
28. Fredric M. Ham, Glenn M. Cohen, Ivica Kostanic and Brent R. Gooch, “ Multivariate determination of glucose concentrations from optimally filtered frequency-warped NIR spectra of human blood serum,” Physiol. Meas printed in UK, vol. 17, pp.1- 20, 1996.
29. Raju Poddar, Joseph Tomas Andrews, Pratyoosh Shukla, Pratima Sen “Noninvasive Glucose Monitoring Techniques: A review and current trends,” Medical Physics, Instrumentation and Detectors eprint arXiv:0810.5755, 10/2008.
30. Andrea Tura, Alberto Maran, Giovanni Pacini, “Noninvasive Glucose Monitoring: Assessment of Technologies and Devices According to Quantitative Criteria,” Diabetes Research and Clinical Practice 77, pp.16-40, 2007.
31. R. W. Waynant, V. M. Chenault, “Overview of Noninvasive Fluid Glucose Measurement Using Optical Techniques to Maintain Glucose Control in Diabetes Mellitus, IEEE. Org/organizations/pubs/newsletters/leos/apr98/overview.htm.
32. Fredric M. Ham, Glenn M. Cohen, Ivica Kostanic and Brent R. Gooch, “ Multivariate determination of glucose concentrations from optimally filtered frequency-warped NIR spectra of human blood serum,” Physiol. Meas printed in UK, vol. 17, pp.1- 20, 1996.
33. Omar S. Khalil, “Spectroscopic and Clinical Aspects of Noninvasive Glucose Measurements,” Clinical chemistry 45:2, pp.165-177 (1999).
34. Mark A. Arnold and Gary W. Small, “Noninvasive Glucose Sensing,” Anal. Chem., 77, 5429-5439, 2005.
35. R. Anthony Shaw and Henry H. Mantsch, “Infrared Spectroscopy in Clinical and Diagnostic Analysis,” Encyclopedia of Analytical Chemistry Edited by Robert A. Meyers. Ó John Wiley & Sons Ltd, Chichester. ISBN 0471 97670 9.
36. Jaspreet Kaur, Jagdish Kumar, H. K. Sardana, R. Bhatnagar, N. S. Mehla, “Non Invasive Blood Glucose Measurement Using Optical Method: Feasibility Study And Design Issues,” ICOP 2009-International Conference on Optics and Photonics CSIO, Chandigarh, India, 30 Oct-1 Nov 2009.
37. Airat K. Amerov, Jun Chen, and Mark A. Arnold, “Molar Absorptivities of Glucose and Other Biological Molecules in Aqueous Solutions over the First Overtone and Combination Regions of the Near-Infrared Spectrum,” Applied Spectroscopy, Volume 58, Number 10, pp.1195-1204, 2004.
38. Jun Chen, Mark A. Arnold and Gary W. Small, “Comparison of Combination and First Overtone Spectral Regions for Near-Infrared Calibration Models for Glucose and Other Biomolecules in Aqueous Solutions, ” Anal. Chem., 76, pp.5405-5413, 2004.
39. Fredric M. Ham, Ivica Kostanic, Glenn M. Cohn, and Brent R. Gooch, “Determination of glucose concentrations in an aqueous matrix from NIR spectra using optimal time domain filtering and partial least squares regression,” IEEE Transactions on Biomedical Engineering, Volume 44, Issue 6, pp.75–485, June 1997.
40. Ndumiso Cingo, Gary W. Small and Mark A. Arnold, “Determination of Glucose in a Synthetic Biological Matrix with Decimated Time Domain filtered Near- Infrared Interferogram Data,” Vibrattional Spectroscopy 2000, vol. 23, pp.103-117.
41. Mark A. Arnold and Gary W. Small, “Determination of Physiological Levels of Glucose in an Aqueous Matrix with Digitally Filtered Fourier Transform Near-Infrared Spectra,” Analytical chemistry, American Chemical Society 62:1414, pp.1457-1464, 1990.
42. Mutua J. Mattu, Gary W. Small, and Mark A. Arnold, “Determination of Glucose in a Biological Matrix by Multivariate Analysis of Multiple Bandpass Filtered Fourier Transform Near- Infrared Interfeograms,” Analy Chem, vol. 69, issue 22, pp.4695- 4702, 1997.
43. Gary W. Small, Mark A. Arnold, and Lois A. Marquadt, “Strategies for Coupling Digital Filtering with Partial least squares Regression: Application to the Determination of Glucose in Plasma by Fourier Transform Near- Infrared Spectroscopy,” Anal Chemistry 1993, 65, pp.3279 – 3289.
44. Næs, Tormod, and Harald Martens. "Principal component regression in NIR analysis: viewpoints, background details and selection of components." *Journal of chemometrics* 2.2 (1988): 155-167.
45. Haaland, David M., and Edward V. Thomas. "Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information." *Analytical Chemistry* 60.11 (1988): 1193-1202.
46. Grégoire, G. "Multiple linear regression." *European Astronomical Society Publications Series* 66 (2014): 45-72.
47. Preacher, K.J., Curran, P.J. and Bauer, D.J., 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of educational and behavioral statistics*, *31*(4), pp.437-448.
48. Liu, R. X., Kuang, J., Gong, Q., & Hou, X. L. (2003). Principal component regression analysis with SPSS. *Computer methods and programs in biomedicine*, *71*(2), 141-147.
49. Næs, T. and Martens, H., 1988. Principal component regression in NIR analysis: viewpoints, background details and selection of components. *Journal of chemometrics*, *2*(2), pp.155-167.
50. Liu, R.X., Kuang, J., Gong, Q. and Hou, X.L., 2003. Principal component regression analysis with SPSS. *Computer methods and programs in biomedicine*, *71*(2), pp.141-147.
51. Lin Zhang, Gary W. Small and Mark A. Arnold “ calibration standardization algorithm for partial Least Squares Regression: Application to the determination of physiological levels of glucose by Near-Infrared spectroscopy,” Anal.Chem.2002,74,pp 4097-4108.
52. Risvik, H., 2007. Principal component analysis (PCA) & NIPALS algorithm.
53. Golub, G.H. and Reinsch, C., 1970. Singular value decomposition and least squares solutions. *Numerische mathematik*, *14*(5), pp.403-420.
54. Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), pp.37-52.
55. Moore, B., 1981. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE transactions on automatic control*, *26*(1), pp.17-32.
56. Haaland, David M., and Edward V. Thomas. "Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information." *Analytical Chemistry* 60.11 (1988): 1193-1202.
57. Geladi, P. and Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. *Analytica chimica acta*, *185*, pp.1-17.
58. Martens, H. and Martens, M., 2001. Multivariate analysis of quality. An introduction.
59. Patchava, K.C., Benaissa, M., Malik, B. and Behairy, H., 2015. Local linear embedded regression in the quantitative analysis of glucose in near infrared spectra. *Analytical Methods*, *7*(4), pp.1484-1492.
60. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), pp.18-28.
61. Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and computing*, *14*(3), pp.199-222.
62. Malik, B., Chaitanya, K. and Benaissa, M., 2014. Support vector regression with digital band pass filtering for the quantitative analysis of near‐infrared spectra. *Journal of Chemometrics*, *28*(2), pp.116-122.
63. Basak, D., Pal, S. and Patranabis, D.C., 2007. Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), pp.203-224.
64. Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., Van Ommen, B. and Smilde, A.K., 2006. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Analytical chemistry*, *78*(2), pp.567-574.
65. Rinnan, Åsmund, Frans van den Berg, and Søren Balling Engelsen. "Review of the most common pre-processing techniques for near-infrared spectra." *TrAC Trends in Analytical Chemistry* 28.10 (2009): 1201-1222.
66. Robinson, M.R., Eaton, R.P., Haaland, D.M., Koepp, G.W., Thomas, E.V., Stallard, B.R. and Robinson, P.L., 1992. Noninvasive glucose monitoring in diabetic patients: a preliminary evaluation. *Clinical chemistry*, *38*(9), pp.1618-1622.
67. Parks, Thomas W., and C. Sidney Burrus. *Digital filter design*. Wiley-Interscience, 1987.
68. Barnes, R. J., Mewa Singh Dhanoa, and Susan J. Lister. "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra." *Applied spectroscopy* 43.5 (1989): 772-777.
69. Bi, Y., Yuan, K., Xiao, W., Wu, J., Shi, C., Xia, J., Chu, G., Zhang, G. and Zhou, G., 2016. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Analytica chimica acta*, *909*, pp.30-40.
70. Kauppinen JK, Moffatt DJ, Mantsch HH, Cameron DG. Fourier self-deconvolution: a method for resolving intrinsically overlapped bands. Applied Spectroscopy. 1981 May;35(3):271-6.
71. Al-Mbaideen, Amneh, and Mohammed Benaissa. "Frequency self deconvolution in the quantitative analysis of near infrared spectra." *Analytica chimica acta* 705.1 (2011): 135-147.
72. Patchava, K.C., Benaissa, M. and Behairy, H., 2015, August. Improving the prediction performance of PLSR using RReliefF and FSD for the quantitative analysis of glucose in Near Infrared spectra. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 2379-2382). IEEE.
73. Small, Gary W., Mark A. Arnold, and Lois A. Marquardt. "Strategies for coupling digital filtering with partial least-squares regression: Application to the determination of glucose in plasma by Fourier-transform near-infrared spectroscopy." *Analytical chemistry* 65.22 (1993): 3279-3289.
74. Al-Mbaideen, A. and Benaissa, M., 2011. Determination of glucose concentration from NIR spectra using independent component regression. *Chemometrics and Intelligent laboratory systems*, *105*(1), pp.131-135.
75. De Maesschalck, R., et al. "The development of calibration models for spectroscopic data using principal component regression." Internet Journal of Chemistry 2.19 (1999): 1099-8292.
76. Mizuta, M., 2012. Dimension reduction methods. In *Handbook of computational statistics* (pp. 619-644). Springer, Berlin, Heidelberg.
77. Abdi, H. and Williams, L.J., 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), pp.433-459.
78. Yazdani, Samaneh, JamshidShanbehzadeh, and Mohammad TaghiManzuriShalmani. "RPCA: a novel preprocessing method for PCA." *Advances in Artificial Intelligence* 2012 (2012).
79. de Jong, S. and Phatak, A., 1997. Partial least squares regression. *Recent advances in total least squares techniques and errors-in-variables modeling*, pp.311-338.
80. Roweis, Sam T., and Lawrence K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." *Science* 290.5500 (2000): 2323-2326.
81. Vin de silva and Joshua B. Tenenbaum, “Global versus Local methods in Non linearDimensonalityReducton,” *proceedings of the conference on Advances in Neural InformatonProcessng Systems (NIPS), 2003.*
82. Chang, H. and Yeung, D.Y., 2006. Robust locally linear embedding. *Pattern recognition*, *39*(6), pp.1053-1065.
83. Dick de Ridder and Robert P.W. Duin, “Locally Linear Embedding for classification,” *IEEE transactions on pattern Analysis and Machine Intelligence.* [*http://www.ph.tn.tudelft.nl/~dck*](http://www.ph.tn.tudelft.nl/~dck)
84. Olga Kouropteva, Oleg Okun and Mattipietik , “Selection of the optimal parameter value for the Locally Linear Embedding algorithm,” *In: Proc of the 1st International conference on Fuzzy Systems and Knowledge Discovery (FSKD’02),pp.359-363.*
85. de Ridder, Dick, and Robert PW Duin. "Locally linear embedding for classification." *Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology, Delft, The Netherlands, Tech. Rep. PH-2002-01* (2002): 1-12.
86. Gary W. Small , Mark A. Arnold and Lois A. Marquadt, “Strategies for coupling Digital filtering with partial Least Squares Regression: Application to the determination of glucose in plasma by Fourier Transform Near- Infrared spectroscopy,” *Analytical Chemistry 1993,65,pp 3279-3289.*
87. Ham, Fredric M., et al. "Determination of glucose concentrations in an aqueous matrix from NIR spectra using optimal time-domain filtering and partial least-squares regression." *Biomedical Engineering, IEEE Transactions on* 44.6 (1997): 475-485.
88. Mitra, Sanjit KK. ‘’*Digital signal processing: a computer-based approach’’*. McGraw-Hill Higher Education, 2000.
89. Parks, Thomas W., and C. Sidney Burrus. *Digital filter design.* Wiley-Interscience, 1987,*.*
90. Christiano, L.J. and Fitzgerald, T.J., 2003. The band pass filter. *international economic review*, *44*(2), pp.435-465.
91. Bolton, J.P. and Burgin, K.N., Inmos Ltd, 1993. *Digital signal processing*. U.S. Patent 5,202,847.
92. Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. Discrete-time signal processing. *Vol. 5. Upper Saddle River: Prentice hall, 1999.*
93. Smith, S. W. Digital signal processing: a practical guide for engineers and scientists*. Newnes,2003.*
94. Belle A. Introduction to digital signal processing and filter design. *Wiley-Interscience, 2005.*
95. Arnold, Mark A. "Non-invasive glucose monitoring." *Current opinion in biotechnology* 7.1 (1996): 46-49.
96. R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics, 11 (1969), pp. 137-148
97. Van der Maaten toolbox for dimensionality reduction[*http://homepage.tudelft.nl/19j49/Matlab\_Toolbox\_for\_Dimensionality\_Reduction.html*](http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html)
98. Refaeilzadeh, P., Tang, L. and Liu, H., 2009. Cross-validation. In *Encyclopedia of database systems* (pp. 532-538). Springer US.
99. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)*2.3 (2011): 27.
100. F. Despagne and D. L. Massart, "Neural networks in multivariate calibration," Analyst, vol. 123, pp. 157R-178R, 1998.
101. F. Marini, R. Bucci, A. Magrě, and A. Magrě, "Artificial neural networks in chemometrics: History, examples and perspectives," Microchemical Journal, vol. 88, pp. 178-185, 2008.
102. N. B. Karayiannis and A. N. Venetsanopoulos, Artificial Neural Networks: Learning Algorithms, Performance Evaluation, and Applications vol. 209: Springer, 1993.
103. *V. N. Vapnik, The nature of statistical learning theory: Springer Verlag, 2000.*
104. Gunn, S.R., 1998. Support vector machines for classification and regression. *ISIS technical report*, *14*(1), pp.5-16.
105. Ma, Junshui, James Theiler, and Simon Perkins. "Accurate on-line support vector regression." *Neural Computation* 15.11 (2003): 2683-2703.
106. S.W. Smith, The Scientist and Engineer’s Guide to Digital Signal Processing, *California Technical Publishing*, 1997.
107. M. Blanco and I. Villarroya, "NIR spectroscopy: a rapid-response analytical tool," TrAC Trends in *Analytical Chemistry*, vol. 21, pp. 240-250, 2002.
108. Malin, Stephen F., et al. "Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy." *Clinical Chemistry* 45.9 (1999): 1651-1658.
109. Garthwaite, P.H.: An interpretation of Partial Least Squares. *Journal of the American Statistical Association* 89, 122 (1994)
110. Höskuldsson, A.: PLS Regression methods. *Journal of Chemometrics* 2, 211–228 (1988)
111. Sun, Yijun. "Iterative RELIEF for feature weighting: algorithms, theories, and applications." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*29.6 (2007): 1035-1051.
112. Dietterich, Thomas G. "Machine-learning research." *AI magazine* 18.4 (1997): 97.
113. Robnik-Šikonja, Marko, and Igor Kononenko. "An adaptation of Relief for attribute estimation in regression." *Machine Learning: Proceedings of the Fourteenth International Conference (ICML’97)*. 1997.
114. Kira, Kenji, and Larry A. Rendell. "A practical approach to feature selection."*Proceedings of the ninth international workshop on Machine learning*. *Morgan Kaufmann Publishers Inc.,* 1992.
115. Kononenko, Igor. "Estimating attributes: analysis and extensions of RELIEF."*Machine Learning: ECML-94*. Springer Berlin Heidelberg, 1994.
116. Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." *Machine learning* 53.1-2 (2003): 23-69.
117. Liu, Huan, and Hiroshi Motoda, eds. Computational methods of feature selection. *Chapman and Hall/CRC*, 2007.
118. Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *IJCAI*. Vol. 14. No. 2. 1995.
119. *Li H.-D., Xu Q.-S., Liang Y.-Z. (2014)* libPLS: An Integrated Library for Partial Least Squares Regression and Discriminant Analysis*. PeerJPrePrints 2:e190v1*, source codes available at [www.libpls.net](http://www.libpls.net)
120. Brown, Christopher D., Lorenzo Vega-Montoto, and Peter D. Wentzell. "Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration." *Applied Spectroscopy* 54.7 (2000): 1055-1068.
121. Chen, Huazhou, et al. "The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models." *ISRN Spectroscopy* 2013 (2013).
122. Chen, Huazhou, et al. "The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models." *ISRN Spectroscopy* 2013 (2013).
123. Kononenko, Igor. "Estimating attributes: analysis and extensions of RELIEF."*Machine Learning: ECML-94*. Springer Berlin Heidelberg, 1994.
124. Kononenko, Igor. "Estimating attributes: analysis and extensions of RELIEF."*Machine Learning: ECML-94*. Springer Berlin Heidelberg, 1994.
125. Al-Mbaideen, A. a., Rahman, T., &Benaissa, M. (2010). Determination of glucose concentration from near-infrared spectra using principle component regression coupled with digital bandpass filter. *2010 IEEE Workshop On Signal Processing Systems*, 243–248. doi:10.1109/SIPS.2010.5624795.
126. A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627.
127. Tong, H., 1990: Nonlinear Time Series Analysis. Oxford University Press, 564 pp.
128. Kantz, H., and T. Schreiber, 1999: Nonlinear Time Series Analysis. Cambridge University Press, 304 pp.
129. Diks, C., 1999: Nonlinear Time Series Analysis: Methods and Applications. World Scientific Press, 180 pp.
130. Gr¨ochenig, K., 2001: Foundations of Time-Frequency Analysis. Birkh¨auser, 359 pp.
131. Flandrin, P., 1999: Time-Frequency/Time-Scale Analysis. Academic Press, 386 pp.
132. Wu ZH, Huang NE, Long SR, Peng CK. ‘’detrending, and variability of nonlinear and nonstationary time series. Proceedings of the National Academy of Sciences of the United States of America. 2007;104:14889–14894.
133. Huang, Norden E. "Introduction to the Hilbert-Huang Transform and its related mathematical problems." *Interdisciplinary Mathematics* 5 (2005): 1-26.
134. Huang, Norden E. "Introduction to the Hilbert-Huang Transform and its related mathematical problems." *Interdisciplinary Mathematics* 5 (2005): 1-26.
135. Huang, Norden E., et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 454. No. 1971. The Royal Society, 1998.
136. Huang, Norden Eh. *Hilbert-Huang transform and its applications*. Vol. 16. World Scientific, 2014
137. Patchava, Krishna Chaitanya, et al. "Savitzky-Golay coupled with digital bandpass filtering as a pre-processing technique in the quantitative analysis of glucose from near infrared spectra." *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016.