PERFORMANCE OF OPTICAL SENSING FOR BLOOD GLUCOSE MEASUREMENT

SO CHI FUK

Ph.D

The Hong Kong Polytechnic University

2013

The Hong Kong Polytechnic University

School of Nursing

Performance of Optical Sensing for Blood Glucose Measurement

SO CHI FUK

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

September 2012

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

SO CHI FUK (Name of student)

ABSTRACT

Diabetes mellitus is an intractable condition in which blood glucose levels cannot be regulated normally by the body alone; it has many complications, including heart disease and stroke, kidney failure, blindness or vision problems, diabetic neuropathy and diabetic foot. Treatment methods include dietary regulation to control blood glucose levels, oral medication, and insulin injection, and all of these treatments should rely on blood glucose measurement. Diabetes mellitus was the tenth most common cause of deaths in Hong Kong in 2010. Currently, the most common means of checking is by using a finger prick glucose meter, but many people dislike using sharp objects and seeing blood because there is a risk of infection. Over the long term, this practice may also result in damage to finger tissue. Given these realities, the advantages of a non-invasive technology are easily understood.

The race for the next generation of painless and reliable glucose monitoring for diabetes is on. As technology advances, so diagnostic techniques and equipment improve. Near infrared (NIR) spectroscopy has become a promising technology, among others, for blood glucose monitoring. While advances have been made, the reliability and the calibration of non-invasive instruments could still be enhanced, and the search has continued to the present without a clinically or commercially viable product emerging. The aim of this study is to evaluate a self-monitoring medical device

adopting NIR spectroscopy, which is able to detect glucose concentrations non-invasively. Moreover, the precision in blood glucose measurement will also be validated.

The objective of this study was to set up a non-invasive blood glucose measurement device that is stable and easy to use to detect the spectral response from human tissue. It was then used to examine different locations of the human body for non-invasive measurement, and the temperature difference of human tissues was inspected. In addition, various pre-processing methods were compared and a series of NIR wavelengths was identified. After that, robust mathematical models for classification and regression approaches were constructed that are able to classify and predict glucose concentrations in blood vessels non-invasively.

Partial least squares (PLS) is widely used in multivariate calibration methods. Partial least squares discriminant analysis (PLS-DA) is a variant of PLS when the dependent variable is binary. They are particularly useful in spectral analysis because the concurrent inclusion of large spectral data for the analyte can greatly improve the precision and applicability of multivariate analysis. Very often, only one single quantitative model is constructed to predict the relationship between the response and the independent variables. This approach can easily misidentify, under or over estimate the important features contained in the independent variables. The results obtained by a single prediction model are thus unstable or correlated to spurious spectral variance, particularly when the training set for PLS is relatively small. New algorithms developed by applying the Monte Carlo (MC) method to PLS and PLS-DA, namely MC-PLS and MC-PLS-DA respectively, are proposed to classify spectral data obtained from NIR

blood glucose measurement. Noise in the data is removed by randomly selecting different subsets from the whole training dataset to generate a large number of models. The new algorithms are then used in determining the mean value over the models with high correlation and small prediction errors for MC-PLS, or the mean sensitivity and specificity of these models are then calculated to determine the model with the best classification rate for MC-PLS-DA. The results show that both the MC-PLS and MC-PLS-DA methods give more accurate prediction results when compared with other multivariate methods used for NIR spectroscopic data of blood glucose. Additionally, the stability of the MC-PLS and MC-PLS-DA models are enhanced compared with the conventional PLS and PLS-DA models.

The MC-PLS and the MC-PLS-DA methods are proposed in this study to tackle the problems in which accuracy is limited by the use of one single prediction model. These methods integrate the Monte Carlo method into the conventional PLS and PLS-DA to improve performance. The proposed algorithms exhibit better performance and accuracy rates when compared to other multivariate methods, as evident from the prediction results on the NIR spectral data. The prediction of the relationship between the response and the independent variables is more accurate, thus enhancing the reliability of the regression model. These advantages make MC-PLS and MC-PLS-DA a promising approach for non-invasive estimation of blood glucose.

PUBLICATIONS

So, C.F., Choi, Kup-Sze, Chung, J.W.Y., Wong, T.K.S. (2013). Modified sequential floating selection for blood glucose monitoring using near infrared spectral data. *Journal of Applied Spectroscopy*, 80(1), 291-294.

So, C.F., Choi, Kup-Sze, Chung, J.W.Y., Wong, T.K.S. (2013). An extension to the discriminant analysis of near-infrared spectra. *Medical Engineering & Physics*, 35(2), 172-177.

So, C.F., Choi, Kup-Sze, Wong, T.K.S., Chung, J.W.Y. (2012). Recent advances in noninvasive glucose monitoring. *Medical Devices: Evidence and Research,* 5, 45-52.

So, C.F., Chung, J.W.Y., Siu, M.S.M., Wong, T.K.S. (2011). Improved stability of blood glucose measurement in humans using near infrared spectroscopy. *Spectroscopy an international journal, 25,* 137-145.

Siu, M.S.M., So, C.F., Choi, Kup-Sze, Chung, J.W.Y., Wong, T.K.S. (2011). Effect of customized diabetes care consultation. *The 2nd International Symposium on Diabetes, Obesity and Cardiovascular Diseases*, September 2011.

Chung, J.W.Y., Siu, M.S.M., Lee, F.S., Choi, M.T., So, C.F., Wong, T.K.S. (2008). Non-invasive glucose measurement using infrared spectroscopy. *Eighth Annual Diabetes Technology Meeting, Journal of Diabetes Science Technology, Diabetes Technology Society*, November 2008.

ACKNOWLEDGEMENTS

I must extend my heartfelt thanks to those who gave me the possibility to complete my study. I am deeply indebted to my supervisors Dr. Kup-Sze Choi, Professor Thomas Wong and Professor Joanne Chung for their intellectual guidance, generous support and valuable criticism at every stage of this study. A research work of this kind would never be realized without their acute foresight. I consider myself very fortunate to pursue reasearch under their guidance.

Special thanks also go to Miss Maggie Siu, Mr. Marcy Choi and Mr. Teddy Lee for their superb support and simulating suggestions concerning this work. The study would not be easy without their efforts. In addition, I am grateful to all the members of the CIDH team for creating a peaceful and harmonious environment from which I could draw comfort when I faced difficulties. We have lots of great times, activities and discussions together so that the hard time passed by easily.

Finally, I would like to express my deepest gratitude to my wife, June, my daughter and son, Clarice and Ernest, for invaluable support and encouragement given to me. Without their understanding and tolerance, I would have never been able to finish this thesis.

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Background

Diabetes mellitus (DM) is a major cause of mortality and morbidity in every country. In 2012, more than 347 million people had DM worldwide (World Health Organization, 2013). Due to the world's increasingly ageing populations, increasingly unhealthy diets, sedentary lifestyles and obesity, it is estimated that the prevalence of DM will double by 2030 (World Health Organization, 2013; Wild, Roglic, Sicree, King, & Green, 2004). It was the tenth most common cause of death in Hong Kong in 2010 (Department of Health, The Government of the Hong Kong Special Administrative Region, 2012). DM is an intractable condition in which blood glucose levels cannot be regulated normally by the body alone. It has many complications, including heart disease and stroke, kidney failure, blindness or vision problems, diabetic neuropathy and diabetic foot. Treatment methods include dietary regulation to control blood glucose levels, oral medication and insulin injection, but all of these have adverse effects on the sufferer's quality of life.

Type 1 diabetes, Type 2 diabetes, and gestational diabetes are three main types of diabetes although some other forms of DM do exist, including congenital diabetes,

cystic fibrosis related diabetes, several forms of monogenic diabetes, and steroid diabetes induced by high doses of glucocorticoids (World Health Organization, 2012). Type 1 diabetes is an autoimmune disease with pancreatic islet beta cell destruction. It is an autoimmune disorder in which the body cannot produce sufficient insulin. Type 2 diabetes is the most prevalent form resulting from insulin resistance due to insulin secretary defect. Both Type 1 and Type 2 diabetes are chronic conditions that usually cannot be cured easily. Gestational diabetes is the term used when a woman develops diabetes during pregnancy. Generally, the situation resolves itself after delivery but it may proceed into the development of Type 2 diabetes later in life.

The control of blood glucose levels relies on blood glucose measurement. Diabetic patients, no matter Type 1 or Type 2, are encouraged to check their blood glucose levels several times per day (International Diabetes Federation, 2011). Currently, the most common means of checking is by using a finger prick glucose meter (American Diabetes Assoication, 2012). In this way, diabetic patients can obtain a clear picture of their blood glucose levels for therapy optimization and for insulin dosage adjustment for those who need daily injections.

Finger-pricking, however, has several disadvantages. Many people dislike using sharp objects and seeing blood; there is a risk of infection. Over the long term, this practice may also result in damage to the finger tissue. Given these realities, the advantages of a non-invasive technology are easily understood. Further, the finger prick glucose meter is a discrete glucose measurement device. It is not practical for continuous monitoring of blood glucose. Continuous glucose monitoring (CGM) is used to monitor blood glucose

levels throughout the day and night (Klonoff, 2005). The CGM system typically consists of a disposable sensor that is inserted into the skin, a wire that connects the sensor to a receiver, and a receiver that records and displays the blood glucose levels (National Diabetes Information Clearinghouse, 2008). The biggest advantage of CGM devices is the real time monitoring that is particularly useful for intensive insulin management or hypoglycaemia control. However, traditional finger-pricking is still needed for device calibration and the output results have a lag time that cannot reflect actual blood glucose levels for CGM devices. Therefore, some incidences of hyperglycemia, or hypoglycemia between the measurements, may not be recorded. Thus, the resultant monitoring cannot fully represent the blood glucose pattern. In this regard, the idea of non-invasive glucose measurement was initiated to eliminate the painful pricking experience, risk of infection, and damage to finger tissue.

Many researchers have attempted to develop a variety of non-invasive methods that monitor blood glucose. Near infrared (NIR) spectroscopy has become a promising technique for blood glucose monitoring among those potential non-invasive approaches, because NIR spectroscopy is a comparably stable process that has low interference caused by other biological components and a higher signal-to-noise (SNR) ratio under room conditions. Figure 1.1 shows the NIR spectra of distilled water, 20% and 50% glucose solution in absorbance measurement ranging from 730 – 2,500nm. However, the search for a promising solution began more than 30 years ago and has continued to the present without a clinically or commercially viable product emerging. Many problems remain unsolved; for example, changes in metabolism and body temperature may consequently affect the blood glucose level. Error sources due to the measurement site

from different parts of body locations may also introduce uncertainty for non-invasive measurement.

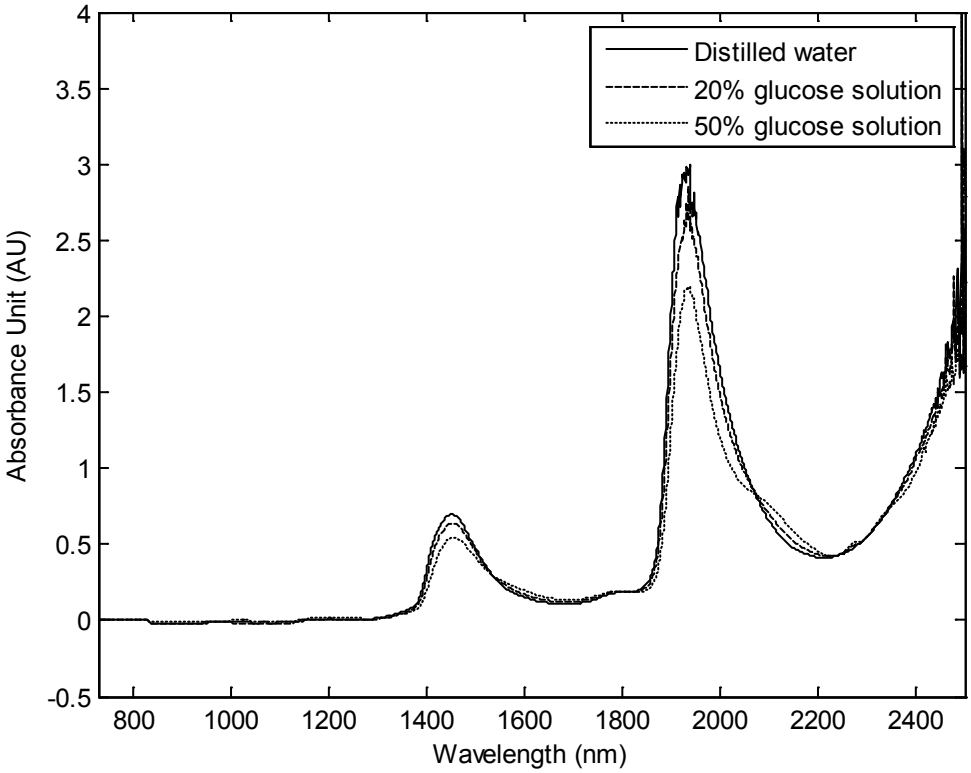


Figure 1.1 NIR spectra of distilled water, 20% and 50% glucose solution in absorbance measurement ranging from 730 – 2,500nm

NIR spectroscopy is based on molecular overtone and combination vibrations. NIR light refers to radiation in the wavelength range from 730nm to 2,500nm. Glucose concentrations are measured through NIR optical sensing. It applies the Beer-Lambert Law, simply known as Beer's Law, that relates the amount of light absorbed by a sample to the concentration of absorbing species in the sample (Smith, 2002). The linear equation of Beer's Law shown in Equation (1.1) demonstrates that the amount of light

absorbed by a sample depends on the absorption of light through the analyte, the thickness of the sample, and the concentration of the analyte.

$$A = \log\left(\frac{I_0}{I}\right) = \varepsilon l c \quad (1.1)$$

where A is the absorbance, I_0 is the intensity of the incident light, I is the intensity, ε is the absorptivity, l is the pathlength, and c is the concentration.

NIR spectroscopy is used to determine the concentration of a substance by measuring how it interacts with light. When light is absorbed during passage through a material, the amount of depletion of the light is measured (which is mostly known as absorbance). This relationship in Equation (1.1) allows absorbance measurements to be used to predict concentrations. For non-invasive blood glucose measurement by optical sensing technology, the absorbance takes place for the spectral measurement. The key component of NIR spectral analysis is dependent on multivariate training methods that require sufficient useful training spectra and sufficient spectral data points to allow analytical information to be extracted from spectra accurately. However, the analytical data matrix generally includes unexpected experimental errors or measurement noise and normally contains hundreds of samples in the NIR spectra. Hence, an appropriate model of spectral response in humans is yet to be determined.

1.2 Objectives

The objective of this study is to construct a non-invasive blood glucose measurement device that is stable and easy to use to detect the spectral response from human tissue. It then examines different human body locations for non-invasive measurement and temperature differences of human tissues. Various pre-processing methods are also compared and the wavelengths relevant to the determination of blood glucose concentration are identified. Finally, robust mathematical models for classification and regression approaches are constructed that are able to classify and predict glucose concentrations in blood vessels non-invasively.

1.3 Conceptual framework

This study adopts the four-stage conceptual framework of bio-signal processing shown in Figure 1.2, (van Bemmel, Musen, & Helder, 1997) which shows that processing bio-signals should normally consist of:

1. Signal acquisition or measurement
2. Signal transformation or signal pre-processing
3. Parameter selection or variable/feature selection
4. Signal classification or signal interpretation

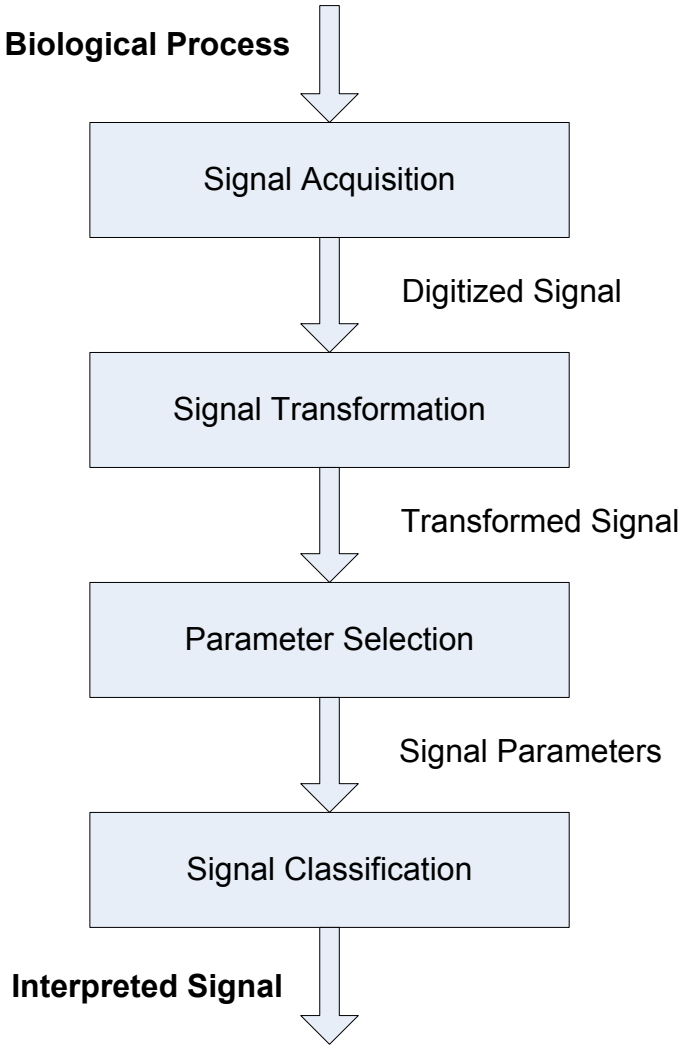


Figure 1.2 The four-stage conceptual framework of bio-signal processing

In the first stage, signal acquisition, a non-invasive blood glucose measurement device based on NIR spectroscopy is set up. NIR light is transmitted through an optical cable to the human skin surface and the reflected light is collected for computing blood glucose concentration in human blood. At this stage, it is most important to maintain a single measurement point in the human body and to obtain the signals with low disturbance so that the effect of measurement sites in different body parts and the temperature at the measurement sites can be investigated.

In the second stage, also called pre-processing, the signals are transformed in a way that such problems can be simplified through a suitable choice of pre-processing method, and misleading results can be avoided. This is because the signals contain much other information, i.e. redundancy, which are not needed to derive parameters in the later stages. By doing so, disturbances are decreased and the amount of unwanted data reduced so that signal quality can be improved for further analysis. Since the pre-processing of NIR spectral data is important for the subsequent multivariate analysis, various normalization schemes and pre-processing techniques are evaluated and reported.

The third stage delivers relevant parameters, also called features, which can be used for decision making. The features are extracted by some complex searching algorithms to distinguish those signal features that contain discriminatory power, and normally they can reduce the size of data so as to compute the diagnostically most significant parameters. Once the signal parameters have been obtained, they are used for further decision making in the interpretation stage. To deal with the difficulty caused by the high dimensionality of the spectral data in multivariate analysis, optimization searching algorithms are applied for features selection to identify the relevant NIR wavelengths.

The interpretation or classification stage of bio-signal processing is the stage of identifying which set of categories a new measurement belongs to, on the basis of a set of training data containing relevant signal parameters whose category membership is known. They can have a logical basis, follow multivariate analysis or be a combination of different methods. This study introduces improved algorithms to deal with the signal

classification stage that can provide more accurate classification or interpretation of results as well as enhance the stability of signal interpretation.

1.4 Contribution

The major contributions the thesis makes include: a large clinical trial on human subjects that uses an assembled non-invasive blood glucose measurement device for data collection; an investigation of the variations in measurement settings (measurement sites and temperature differences) that may affect the calculation of blood glucose concentration; the selection of pre-processing method(s) and the algorithm for feature selection; and developing computational algorithms to improve the performance and accuracy rate of the calculation. In the last case, the proposed algorithms outperform conventional multivariate methods, whereby predicting the relationship between the response and the independent variables is more accurate, thus enhancing the reliability of the regression model. The findings obtained in the thesis provide a useful reference for future development in non-invasive NIR blood glucose measurement.

1.5 Organization of the thesis

The remainder of this thesis is organized as follows. Chapter 2 presents a review of the literature concerning non-invasive technologies and the devices that appear in the market. Chapter 3 reviews the existing multivariate analysis approaches for classification and

regression. Chapter 4 describes the study design and data acquisition. Chapter 5 illustrates the ground work for implementing the non-invasive measurement device, which includes data pre-processing, the selection of a measurement site from different body locations, and the temperature effect for human tissues. Chapter 6 describes the improvement work on the classification approach and regression approach based on the Monte Carlo method as applied to the algorithm. Finally, the discussion and conclusion sections presented in Chapter 7 summarize the overall findings in this study.

CHAPTER 2

NON-INVASIVE GLUCOSE MONITORING

The non-invasive concept was launched more than 30 years ago (Rabinovitch, March, & Adams, 1982). Nevertheless, most non-invasive technologies are still in their early stages of development. Many non-invasive technologies have been described in the literature, and there is an increasing volume of recent research results. Keeping up with the current situation requires constant updating (Khalil, 2004). The results of an Internet search provide much information on this topic, such as overviews of non-invasive technology (Waynant & Chenault, 1998), the future development of meters and monitors for diabetes (The Diabetes Mall, 2010), and information about research centers that are developing this technique (Optical Science & Technology Center, 2010). However, the scope of devices is so broad that no single website can keep up. Much information is outdated (Mendosa, 1995). Therefore, the focus of this chapter is not just to review the related literature but to present the current state of the art of non-invasive glucose monitoring for diabetes. It will describe the technologies being used, technologies under development, devices being used, and the companies producing these devices.

2.1 The technologies

This study considers non-invasive glucose measurement as any technique that does not involve pricking (breaking) the skin. The different techniques/technologies are listed in alphabetical order. The principle of each technology, together with its advantages and limitations, are discussed.

2.1.1 Bioimpedance spectroscopy

Principle

Bioimpedance is a measure of the resistance to electric current flowing through the tissues of a living organism (Sverre & Orjan Grottem, 2008). Measuring the bioelectrical impedance has proved useful as a non-invasive method for measuring body composition (Tao & Adler, 2009). The impedance spectrum or dielectric spectrum is measured in the frequency range of 0.1 to 100MHz. Variations in plasma glucose concentration induce a decrease in sodium ion concentration in red blood cells and an increase in potassium ion concentration (Hillier, Abbott, & Barrett, 1999). These variations cause changes in the membrane potential of red blood cells, which can be estimated by determining the permittivity and conductivity of the cell membrane through the dielectric spectrum (Caduff, Hirt, Feldman, Ali, & Heinemann, 2003; Ermolina, Polevaya, & Feldman, 2000; Polevaya, Ermolina, Schlesinger, & Ginzburg, 1999). In 2003, Pendragon Medical Ltd. (a company in Zurich, Switzerland) developed a wrist band-based glucose monitor called “Pendra” based on this technology. This product was soon withdrawn from the

market, however, because of poor reliability. Currently, Caduff's research group is still working on this technology (Caduff, et al., 2009).

Advantages

Bioimpedance spectroscopy does not require the use of statistically-derived, population-specific prediction models. It has the potential advantage of being able to differentiate between extracellular water and intracellular water, thereby providing an estimate of body cell mass, and thus characterizing the blood bioimpedance properties. The instrument is easy to use and low in cost compared to other devices.

Limitations

The limitation of this technology is that it requires an equilibration process, where the user must rest for 60 minutes before starting the measurements (Caduff, Hirt, Feldman, Ali, & Heinemann, 2003). In addition, it will ionize the body's molecules when using bioimpedance technology. Moreover, some problems remain to be clarified, such as the effects of temperature and body water content (e.g., skin moisture, sweat, overall hydration) on readings (Caduff, et al., 2009).

2.1.2 Electromagnetic sensing

Principle

Like bioimpedance spectroscopy, this technology assesses dielectric parameters of blood. The difference between them is that an electric current is used in bioimpedance

spectroscopy, while the electromagnetic coupling between two inductors is used in electromagnetic sensing (Gourzi, et al., 2005; Tura, Sbrignadello, Cianciavicchia, Pacini, & Ravazzani, 2010). The sensor uses electric currents to detect variations of the dielectric parameters of the blood, which may be caused by glucose concentration changes (Moran, Jeffrey, Thomas, & Stevens, 2000). The frequency range used in this technique is 2.4 – 2.9MHz. However, depending on the temperature of the investigated medium, an optimal frequency exists at which the sensitivity to glucose changes reaches the maximum. Determining this frequency is important for the efficacy of the device. Gourzi's research group suggested the optimal frequency is 2.664MHz at 24°C (Gourzi, et al., 2005). However, another study of this technology using pig blood suggests that the optimal operating frequency is 7.77GHz at 25°C (Melikyan, et al., 2011).

Advantages

Using a specific frequency range can isolate the effect due to blood glucose and minimize the interference caused by other substances like cholesterol that may skew the readings. In addition, the method is relatively safe because it will not ionize the body's molecules as in the case of bioimpedance spectroscopy.

Limitations

Temperature has a strong effect on this form of measurement because it influences the optimal investigation frequency. Furthermore, Moran's research group reported that the blood dielectric parameters depend on several components other than glucose (Moran, Jeffrey, Thomas, & Stevens, 2000). Therefore, more study on the potential confounders is needed before this technology can be considered reliable.

2.1.3 Fluorescence technology

Principle

This technique uses fluorescence reagents to track the presence of glucose molecules in blood. Many approaches exist, such as measuring changes in fluorescence resonance energy transfer between a fluorescent donor and an acceptor or measuring glucose-induced changes in intrinsic fluorescence of enzymes (Pickup, Hussain, Evans, Rolinski, & Brich, 2005). One study reported that glucose levels in tears reflect concentrations similar to those in blood, and thus fluorescence of tears can be used for non-invasive glucose monitoring. Khalil reported that this approach can track blood glucose with an approximate 30-minute lag time and does not suffer from interference caused by variations in the light intensity of the ambient environment (Khalil, 2004a). The photonic sensing is achieved with polymerized crystalline colloidal arrays that respond to different glucose concentrations through the diffraction of visible light.

Advantages

This technology is very sensitive and can be used to detect single molecules. It causes little or no damage to the body. In addition, glucose concentration is measured in terms of fluorescence intensity and decay times, both of which are independent of light scattering and fluorophore concentration, which can reduce loss through diffusion or degradation.

Limitations

Photonic sensing can suffer from strong scattering phenomena, especially in fluorescence technology. Moreover, limitations such as short lifetimes and biocompatibility of the sensor devices have to be dealt with, possibly through the use of colorimetric assays (Moschou, Sharma, Deo, & Daunert, 2004).

2.1.4 Mid-infrared spectroscopy

Principle

Mid-infrared (MIR) spectroscopy employs the same principles as infrared spectroscopy. It is the absorption measurement of MIR frequencies through a sample positioned in the path of an MIR beam, which is based on light in the 2,500 – 25,000nm wavelength region of the spectrum. The absorption differences when MIR light meets human tissues can be represented by certain modelling techniques in spectral quantitative analysis. A partial least squares algorithm is now commonly used for multivariate calibration.

Advantages

MIR light exhibits decreased scattering and increased absorption when compared with NIR spectroscopy because of the larger wavelengths (Lilienfeld-Toal, Weidenmuller, Xhelaj, & Mantele, 2005). Light can only penetrate into skin for a depth of a few micrometers. As a result, only reflected light can be considered, because there is no light transmitted through a body segment. Moreover, another possible advantage of MIR

spectroscopy is that the response peaks of glucose and other compounds are stronger than those in NIR.

Limitations

Poor penetration is the main limitation of MIR because it may not be able to reach blood vessels under the skin. Other limitations, as with NIR, include problems caused by confounding factors such as water content in blood (Brancaleon, Bamberg, Sakamaki, & Kollias, 2001).

2.1.5 Near infrared spectroscopy

Principle

Near infrared (NIR) spectroscopy is located in the wavelength region of 730 to 2,500nm. The principle is similar to that of MIR spectroscopy. NIR spectra are made up of broad bands corresponding to overlapping peaks: the overtones (i.e. first, second, third and combination overtones) formed by molecular vibrations. It allows blood glucose measurement in tissues by variations of transmittance and reflectance of light. Heise, one of the pioneers in noninvasive blood glucose monitoring, has conducted considerable research on NIR techniques (Heise, Bittner, & Marbach, 1998; Heise & Marbach, 1998; Siesler, Ozaki, Kawata, & Heise, 2002). It was also reported that glucose generates one of the weakest NIR absorption signals per concentration unit of the body's major components (Raghavachari, 2001). The research group of Maruo demonstrated the efficacy of this approach in vivo, using NIR diffuse reflectance

spectroscopy through fiber optics on diabetes patients' forearms (Kasemsumran, Du, Maruo, & Ozaki, 2006; Maruo, Tsurugi, Chin, Ota, Arimoto, & Yamada, 2003). The results showed a positive sign on the correlation between predicted values and the reference glucose levels. In addition, Arnold's research group reported that although measurement errors of NIR spectroscopy are too large for clinical purposes, these experimental results demonstrate the potential of non-invasive blood glucose measurements (Arnold & Small, 2005; Liu & Arnold, 2009; Tarumi, Amerov, Arnold, & Small, 2009).

Advantages

The high sensitivity of the photo conductive detectors is the main advantage of NIR spectroscopy. Water is reasonably transparent to the signal bandwidth used by NIR, which makes it possible to use it for blood glucose monitoring. In addition, the measured signal is stronger than that in MIR spectroscopy. This method is also less expensive than MIR; the cost is relatively lower, and the materials required can be obtained from a wide range of commercially available products. These advantages make NIR popular in this research area.

Limitations

Despite much promising work, researchers still cannot overcome important shortcomings, namely and particularly: the scanning location of the human body and the force that must be applied to the measurement site; physiological differences not related to blood glucose; the relatively small fraction of glucose in blood (Arnold & Small,

2005); weak correlation; as well as hardware sensitivity and stability (David & Julie, 2010; Liu, Deng, Chen, & Xu, 2005).

2.1.6 Optical coherence tomography

Principle

Optical coherence tomography is an optical signal acquisition method based on the use of a low coherence light, such as a super luminescent light and an interferometer, which determines the depth of the scattering feature by measuring the delay correlation between a sample arm and a reference arm with a moving mirror to achieve the depth measurement (Larin, Eledrisi, Motamedi, & Esenaliev, 2002). Light backscattered from tissues is combined with light returned from the reference arm of the interferometer, and the resulting signal is detected by the photodetector. The delay correlation between the backscattered light in the sample arm and the reflected light in the reference arm is then measured. An increase of glucose concentration in the interstitial fluid causes an increase in refractive index, which in turn creates a decrease in the mismatch between sample and reference indices to provide a higher correlation.

Advantages

This technology has high signal to noise ratio, high resolution and depth of penetration, because the interferometric signal can be formed only within the coherence length of the source.

Limitations

Optical coherence tomography is sensitive to the individual's body movement. In addition, while slight changes in skin temperature have negligible effects, changes of several degrees have a significant influence on the signal (Yeh, Hanna, & Khalil, 2003). Additionally, there is no clear evidence that this method has advantages compared to other scattering-based techniques.

2.1.7 Optical polarimetry

Principle

Some researchers are trying to apply optical polarimetry in non-invasive glucose monitoring. Because the high scattering coefficients produce complete depolarization when the beam strikes the skin, attention has been focused on the eye, which offers a clear optical medium with a reasonable path length in relation to blood glucose (Malik & Cote, 2010). The polarization is expected to rotate several degrees depending on the concentration of glucose when a light transverses vitreous humor.

Advantages

As light absorption and scattering in the eye are low, and there are virtually no large protein molecules in the aqueous humor, the main component in the aqueous humor is glucose; therefore, stronger correlation may exist to determine the blood glucose concentration. In addition, this technique makes use of visible light, and the optical components can be easily miniaturized.

Limitations

This technique is sensitive to the scattering properties of the investigated tissue, as scattering depolarizes the light. As a result, skin cannot be investigated by polarization technology, because it shows high scattering effects, particularly in the stratum corneum. Moreover, eye movement and motion artifacts are general sources of errors for this technique. Furthermore, the specificity of this technique is poor, as several optically active compounds are present in human fluids containing glucose such as albumin and cholesterol.

2.1.8 Raman spectroscopy

Principle

Raman spectroscopy is based on the use of a laser light to induce oscillation and rotation in human fluids containing glucose. Because the emission of scattered light is influenced by molecular movement, it is possible to estimate glucose concentration in human fluids (Berger, Koo, Itzkan, Horowitz, & Feld, 1999). This effect depends on the concentration of the glucose molecules. This technique can measure very weak signals even in human fluids. The wavelength range of Raman spectrum is considered to be 200 to 2,000cm^{-1} (Hanlon, et al., 2000). Raman spectrum of glucose can be differentiated from those of other compounds in this band.

Advantages

Raman spectroscopy usually provides sharper and less overlapped spectra compared to NIR spectroscopy. The intensity of spectral features is proportional to the concentration of the particular species, and the spectra are less sensitive to temperature changes. Moreover, it is comparatively less sensitive to water, and the interference from luminescence and fluorescence phenomena is only modest.

Limitations

The main limitations are related to instability of the laser wavelength and intensity, and long spectral acquisition times. In addition, as the power of the light source must be kept low to prevent injury, the signal-to-noise ratio is significantly reduced. Moreover, like NIR spectroscopy, interference from other compounds remains a problem.

2.1.9 Reverse iontophoresis

Principle

Reverse iontophoresis is based on the flow of a low electrical current through the skin, between an anode and cathode positioned on the skin surface. An electric potential is applied between the anode and cathode, thus causing the migration of sodium and chloride ions from beneath the skin towards the cathode and anode, respectively. In particular, it is sodium ion migration that mainly generates the current (Sieg, Guy, & Delgado-Charro, 2004). This measurement is possible because neutral molecules, such as glucose, are extracted through the epidermis surface during this convective flow. This

flow causes interstitial glucose to be transported, collecting at the cathode where a traditional glucose sensor is placed to measure glucose concentration directly. The "GlucoWatch" device (Cygnus Inc., Redwood City, CA) is based on this technology and has been approved by the US Food and Drug Administration (FDA). The device collects glucose molecules through the cathode disk and measures the amount by a sensor that contains enzyme glucose oxidase. Blood glucose concentration is predicted by comparing the pre-measured blood glucose value with the signal generated by glucose molecules collected at the cathode. However, this product was withdrawn from the market due to poor accuracy, skin irritation, and long procedural problems.

Advantages

The advantage of this technology is that the electrodes are easily applied to the skin, by which a physiologically relevant fluid sample is collected which has a correlation between glucose concentration in physiological fluid and glucose concentration in blood.

Limitations

While reverse iontophoresis technology has great potential, the only device ever marketed using it had such serious practical drawbacks that it was withdrawn from the market. First, the electrodes irritated the skin. Second, the electrodes needed to be in place for at least 60 minutes, which exceeded the patience of many users. Third, readings were inaccurate, especially when the subject was sweating. Fourth, it was not able to detect the rapid changes in blood glucose, due to its long "wake up" time.

2.1.10 Ultrasound technology

Principle

Ultrasound technology is based on low frequency ultrasound that penetrates the skin for blood glucose monitoring. While this approach has theoretical potential, it seems that no further work has been done since Lee's group reported their laboratory results on rat skins (Lee, Nayak, Dodds, Pishko, & Smith, 2005). A variation, named photoacoustic spectroscopy (PAS), is being used, which is based on the use of a laser light for the excitation of a fluid and for measuring the resulting acoustic response (MacKenzie, et al., 1999). The fluid is excited by a short laser pulse with a wavelength that is absorbed by a particular molecular species in the fluid. Light absorption causes microscopic localized heating in the medium, which generates an ultrasound pressure wave that is detected by a microphone. The principle of the photoacoustic method is that an energy source irradiates the skin surface, causing thermal expansion in the illuminated area. An acoustic wave is released because of the energy of the thermal expansion. Detecting glucose with this technique is based on measuring the changes of the peak-to-peak value of the signal, which will vary according to the glucose content of the blood.

Advantages

This technology can provide higher sensitivity than traditional spectroscopy in the determination of glucose because of the relatively better photoacoustic response of blood, as compared with water. This makes it easier to distinguish hydrocarbons and glucose

(MacKenzie, et al., 1999). Also, the laser light wavelengths that can be used vary in a wide range from ultraviolet to NIR.

Limitations

The technology is sensitive to interference from some biological compounds, temperature fluctuations, and pressure changes. Moreover, when the laser light transverses a dense medium, the photoacoustic signal may be affected by scattering phenomena, which may possibly cause an adverse effect similar to NIR spectroscopy. Another disadvantage is that the instrumentation is expensive and sensitive to environmental parameters.

2.2 Current development for non-invasive glucose monitoring

Various non-invasive technologies were discussed in the previous section. Clearly, many research groups are exploring a wide variety of approaches, trying to develop a blood glucose measurement device that can provide stable and reliable results, conveniently and economically. Table 2.1 shows the most recent work and internet references. Discontinued products such as GlucoWatch, Diasensor (Biocontrol Technology Inc., Pittsburgh, PA) and Pendra, are not listed. In addition, Appendix A shows the devices listed in Table 2.1 for reference.

Device / Company	Technology	Status	URL
BioSensors Inc.	SEMP Technology (Bioimpedance Spectroscopy)	Appeared in 2010 and is under development	http://www.biosensors-tech.com/
ClearPath DS-120, Freedom Meditech	Fluorescent Technology	Appeared in 2007 and is said to be delivered to FDA for approval in 2011	http://freedom-meditech.com/
Cnoga Medical	NIR Spectroscopy	Appeared in 2010 and is said to be delivered to FDA for approval in 2011	http://www.cnoga.com/Medical/Products/Glucometer.aspx
C8 MediSensors	Raman Spectroscopy	Appeared in 2011 and the current status is investigational device	http://www.c8medisensors.com/us/home.html
Easy Check, Positive ID	Chemical sensing in exhaled breath	Appeared in 2010 and is under development	http://www.positiveidcorp.com/products_easycheck.html
EyeSense	Fluorescent Technology	Appeared in 2008 and is still in R&D phase; plan is to launch the device in 2013	http://www.eyesense.com/en/konzept.htm
Glucoband, Calisto Medical Inc.	Bio-electromagnetic Resonance	Appeared in 2005 and claimed under pilot production in 2011	http://www.calistomedical.com/
GlucoTrack, Integrity Applications Ltd.	Ultrasonic, Conductivity and Heat capacity Technology	Under clinical trials phase (last checked: 2011)	http://www.integrity-app.com/
Glove Instruments	NIR Spectroscopy (Optical Bridge Technology)	Appeared in 2008 and is said to be commercialized soon, (last checked: 2012)	http://groveinstruments.com/

OrSense Ltd.	Occlusion Technology (Proprietary Technology)	Appeared in 2006, the company has stated that this product is for market awareness purposes only	http://www.orsense.com/Glucose
SCOUT DS,VeraLight Inc.	Fluorescent Spectroscopy	Appeared in 2011 and has received approval from Health Canada for commercial distribution	http://www.veralight.com/products.html

Table 2.1 Information regarding non-invasive glucose monitoring devices

It is worth noting that very little evidence has proven the analytical feasibility of glucose monitoring by the non-invasive devices listed in Table 2.1. The supporting documentation provided by the research groups is severely limited. Most of the technologies are proprietary, and limited information is disclosed. Although some of the technologies are mentioned in refereed papers, very little specific relevant information is provided. In particular, the judgments of measurement accuracy are completely omitted.

At the same time, Table 2.1 shows that many research groups are working on this problem, trying to develop new measurement technologies and methods to measure blood glucose non-invasively. One of the main reasons is that existing technologies, such as absorption spectroscopy, are relatively poor in signal-to-noise ratio in relation to blood glucose concentration and spectral response. Due to the huge anticipated market for a successful non-invasive glucose monitoring device, the race for research teams to develop more precise and accurate spectroscopic equipment is heated.

2.3 Summary

This chapter described the latest technologies and devices for non-invasive glucose monitoring. It seems that NIR spectroscopy has become a promising technology, among others, for non-invasive blood glucose monitoring. Unfortunately, none of these technologies has produced a commercially available, clinically reliable device; therefore, much work remains to be done. In addition, as multivariate analysis is commonly used to extract relevant information from different types of spectral data to predict analyte concentrations, it plays a critical role in absorption spectroscopy; therefore, the most commonly used multivariate techniques for classification and regression approaches will be reviewed in the next chapter.

CHAPTER 3

REVIEW OF MULTIVARIATE ANALYSIS

The spectroscopic method has become a widely used non-destructive measurement technique for agrochemical, pharmaceutical, and medical applications (Burns & Ciurczak, 2008; Ciurczak & Drennen, 2002; Ozaki, McClure, & Christy, 2007), and multivariate analysis is commonly used to extract relevant information from different types of spectral data to predict analyte concentrations (Rencher, 2002; Smith, 2002). They are particularly useful in spectral analysis because the concurrent inclusion of large spectral data for the analyte can greatly improve the precision and applicability of multivariate analysis. In this chapter, the most commonly used multivariate techniques for classification approach for unknown samples (qualitative analysis) as well as regression approach for unknown samples (quantitative analysis) are reviewed. In addition, the performance measurands for multivariate analysis are mentioned, because they will be used in the following chapters for evaluation.

3.1 Classification for qualitative analysis

In the classification approach, the sample properties that relate to spectral variations belong to several different groups or classes. Classes may represent the identity or

quality, e.g. normal or high, good or bad, yes or no. In this regard, a variety of methods have been developed for classifying samples based on measured responses, and these methods can be subdivided in non-supervised and supervised methods. Non-supervised methods do not require any pre-established class memberships but instead produce the grouping or clustering themselves. Supervised methods, also known as discriminant analysis, use known class memberships so that qualitative data are added to the quantitative spectral data. As the following study focuses on the use of supervised methods, supervised methods such as linear discriminant analysis, artificial neural network, and support vector machine will therefore be discussed.

3.1.1 Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is used to classify analyte concentration based on the spectral data (Huberty & Olejnik, 2006). A discriminant function f(x) that is a linear combination of the components **x** can be written as

$$\mathrm{f}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = a^{\mathrm{T}}y \tag{3.1}$$

where **w** is the weight vector, w_0 is the threshold weight, and $a = [w_0, \mathbf{w}]^{\mathrm{T}}$, $y = [1, \mathbf{x}]^{\mathrm{T}}$ is the augmented weight and feature vector, respectively. The superscript $^{\mathrm{T}}$ denotes the matrix transpose. The goal here is to find the weight vector **w**. The expression can be simplified by using matrix notation as follows,

$$\mathbf{Ya} = \mathbf{b} \tag{3.2}$$

where **Y** is the *n*-by-*d* matrix whose *i*-th row is the vector $\boldsymbol{y}_i^{\mathrm{T}}$ and **b** is the output column vector. Now, the purpose is transformed to the problem of finding the weight vector **a**.

One approach is to minimize the square length of the error vector. Hence, the minimum sum of squared error criterion function is used

$$J_s(\mathbf{a}) = \|\mathbf{Ya} - \mathbf{b}\|^2 = \sum_{i=1}^{n} (\mathbf{a}^T\mathbf{y}_i - b_i)^2 \tag{3.3}$$

which can be then solved by a gradient search procedure. A simple closed form solution can also be found by forming the gradient

$$\nabla J_s = \sum_{i=1}^{n} 2(\mathbf{a}^T\mathbf{y}_i - b_i)\mathbf{y}_i = 2\mathbf{Y}^T(\mathbf{Ya} - \mathbf{b}) \tag{3.4}$$

and setting it to zero. This yields the necessary condition

$$\mathbf{Y}^T\mathbf{Ya} = \mathbf{Y}^T\mathbf{b}, \text{ and} \tag{3.5}$$

$$\mathbf{a} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{b} \tag{3.6}$$

which are desirable because $\mathbf{Y}^T\mathbf{Y}$ is often nonsingular, meaning that a minimum square error solution always exists.

3.1.2 Artificial neural network (ANN)

An artificial neural network (ANN) is a statistical modelling inspired in the natural neurons and is commonly used in modelling complex relationships among independent variables and dependent variables. ANN is a self adaptive and data driven modelling technique although it resembles regression analysis, but has much more flexibility because it is not restricted by any statistical assumptions or pre-specified algorithms. It basically consists of inputs, which are multiplied by weights and then computed by a mathematical function that determines the activation of the neuron. Figure 3.1 shows the

basic architecture of an artificial neuron. The higher a weight of an artificial neuron is, the stronger the input which is multiplied by it will be. Depending on the weights, the computation of the neuron will be different. By adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. The presence of artificial neurons in the network greatly increases its capacity to deal with various complicated relationships. However, when an ANN of hundreds or thousands of neurons exists, it would be quite complicated to find all the necessary weights.

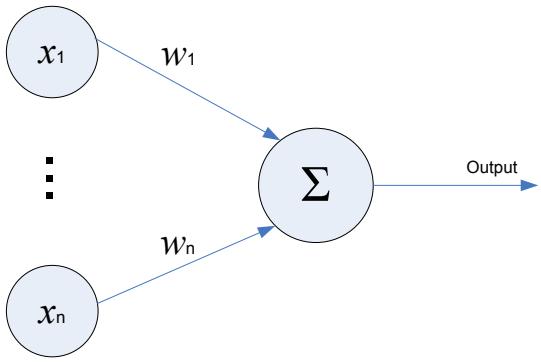


Figure 3.1 The basic architecture of an artificial neuron

A fundamental architecture of an ANN is shown in Figure 3.2 according to the idea of artificial neurons. This consists of an input layer, a hidden layer and an output layer. The distinguishing characteristic of an ANN is the hidden layer that contains different numbers of hidden neurons so that the weights of the ANN can be adjusted in order to obtain the desired output from the network. This process of adjusting the weights is called learning or training.

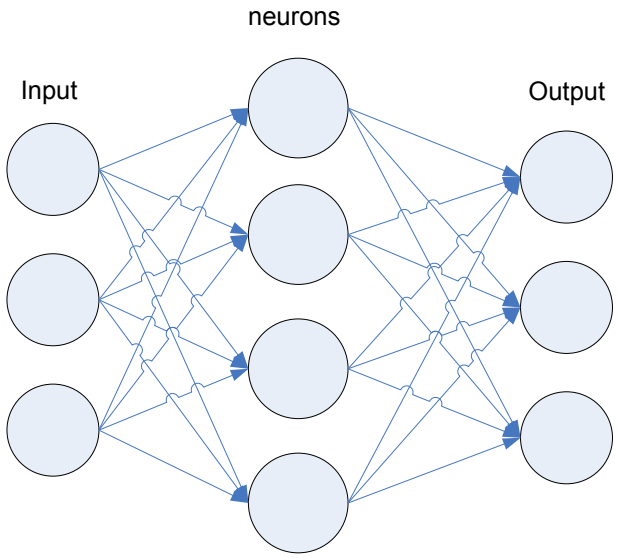


Figure 3.2 A fundamental architecture of an ANN

There are many different types of learning algorithms for ANNs, and the most popular training algorithm is backpropagation since many other learning algorithms are based on it. The backpropagation algorithm (Rumelhart & Mcclelland, 1986) is used in layered feed forward ANN, which means the artificial neurons are organized in layers, send their signals forward, and the errors are propagated backwards. The backpropagation algorithm uses supervised learning. This utilizes the steepest gradient descent in a multilayer perceptron to minimize the sum squared error. The steepest gradient descent is a mathematical algorithm that locates the local minimum of a function by taking steps proportional to the negative of the gradient of the function at the current point. The weights are updated according to the sum squared error feedback from the output neurons until the mean squared error is minimized in backpropagation training.

There may be one or more intermediate hidden layers. However, it is worth noting that for practical reasons, ANNs implementing the backpropagation algorithm do not have too many layers (normally one hidden layer) because the computation time for training the networks increases exponentially.

3.1.3 Support vector machine (SVM)

Support vector machine (SVM) is a supervised learning system that uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that has originally been used for classification analysis (Cristianini & Shawe-Tayor, 2002; Kecman, 2001). Nowadays, SVM can be applied to a number of applications such as system identification, pattern recognition, bioinformatics, finance, and marketing. In a supervised learning machine, it is given a training set (inputs) with associated variables (outputs). A number of sets of hypotheses can be chosen if a set of training data points are present. To tackle the problem, linear function is the best understood and simplest to apply. Since SVM is a linear learning machine, this means that a linear function is used to solve a system. The problem of linear regression consists in finding a linear function as listed below that best interpolates a given set of training data points

$$f(\mathbf{x}) = y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \tag{3.7}$$

where $\mathbf{w}$ is the weight vector to the hyperplane, b is the offset, and the superscript $^{\mathrm{T}}$ denotes the matrix transpose.

In a classification process, the output is either 1 or -1 which indicates the class to which the input belongs. The optimization problem is to find the pair of hyperplanes that give the maximum margin that depends on the norm of **w** such that minimize ($\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$) subject to $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x_i} - b) \geq 1$ where $i = 1, \cdots, n$.

Consider how to transform this optimization problem into its corresponding dual problem by introducing the Lagrange multipliers. The Lagrangian function for the constrained problem can be expressed as

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b) - 1] \tag{3.8}$$

where $\alpha_i \geq 0$

The Lagrangian function in Equation (3.8) must be minimized with respect to **w** and *b* with the necessary condition that the derivatives of *L* with respect to all the α_i vanish. This problem can now be solved by standard quadratic programming techniques. It is found by differentiating with respect to **w** and *b*, imposing stationarity.

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathrm{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \tag{3.9}$$

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.10}$$

Substituting (3.9) and (3.10) into (3.8) to obtain

$$
\begin{aligned}
L(\mathbf{w}, b, \alpha) &= \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b) - 1] \\
&= \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j
\end{aligned}
\tag{3.11}
$$

The corresponding $\mathbf{x}_i$ are exactly the support vectors (critical elements of the training set) that lie on the margin and satisfy the condition. Because of the specific formulation of the cost function and the use of the Lagrangian theory, it can be proven that the solution found is always global because the problem formulation is convex (Burges, 1998). In addition, the solution is unique due to the convex property.

If the surface separating the two classes is not linear, this approach has to be extended (Boser, Guyon, & Vapnik, 1992). This can be done by replacing $\mathbf{x}_i$ by a mapping into feature space with a kernel function $k(\mathbf{x}_i)$ that linearizes the relation between the inputs and outputs variables. In the feature space, Equation (3.11) can be expressed as

$$
L(\mathbf{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{3.12}
$$

Many kernel functions can be used and the most widely used is the Gaussian radial basis function, which is shown as

$$
k(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, \qquad \text{for } \sigma \geq 0 \tag{3.13}
$$

With a suitable kernel, SVM can separate in the feature space the data that were non-separable in the original input space. This property means that non-linear algorithms can

be obtained by using proven methods to handle linearly separable data sets, and they usually exhibit good generalization performance.

3.2 Regression for quantitative analysis

The regression approach is focused on the relationship between a dependent or criterion variable and one or more independent or predictor variables. It is indirect and relies on the ability to develop a model that relates the set of measured variables to the property of the system. It is widely used for prediction and forecasting (Escandar, Damiani, Goicoechea, & Olivieri, 2006; Eriksson, Gottfries, Johansson, & Wold, 2004; Wold, Cheney, Kettaneh, & McCready, 2006). Most importantly, the regression approach is used to explore the forms of these relationships so as to understand which among the independent variables are related to the dependent variable. Multiple linear regression (MLR) (Ostrom, 1990; Weisberg, 2005) and principal component regression (PCR) (Martens & Naes, 1989; Naes, Isaksson, Fearn, & Davies, 2002) are widely used methods to construct the quantitative model in spectral analysis. Furthermore, the key component of quantitative analysis is dependent on multivariate training methods that require sufficient and useful training data to allow analytical information to be extracted accurately. Undoubtedly, partial least squares regression (PLS) (Hoskuldsson, 1988; Paul & Bruce, 1986) is the most commonly used multivariate training procedure to build the quantitative model, especially in absorption spectroscopy, because PLS attempts to maximize covariance between the response and independent variables. Therefore, the

PLS algorithm is introduced here, and which will focus on utilization for the prediction process in later chapters.

3.2.1 Multiple linear regression (MLR)

Multiple linear regression (MLR) is based on the least squares approach in which the sum of squares differences of observed and predicted values is minimized to construct the model (Wise & Kowalski, 1995). It assumes that a regression vector **b** can be used to determine a property of the system **y** from the measured variables **A**. Therefore, the model is

$$\mathbf{y} = \mathbf{Ab} \tag{3.14}$$

where **y** is the measured response vector, **A** is the matrix of component responses and **b** is the vector containing the weights of the analytes

The regression vector **b** should be determined by using the matrix **A** and the known values of the property of system **y**. Therefore, **b** can be estimated as

$$\mathbf{b} = \mathbf{A}^{+}\mathbf{y} \tag{3.15}$$

where $\mathbf{A}^{+}$ is the pseudoinverse of **A**

Now, $\mathbf{A}^{+}$ is defined as

$$\mathbf{A}^{+} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}} \tag{3.16}$$

The regression vector **b** can be determined based on (3.16) and the known values **y**. However, MLR is not always effective because of collinearity of some or all of the response variables in **A** (Naes, Isaksson, Fearn, & Davies, 2002); for example, some columns of variables **A** are linear combinations of other columns, or variables **A** contain fewer samples than measured variables. In either case, the $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ would not exist. Moreover, the use of highly collinear variables in MLR increases the chance of overfitting for the resulting model, which is not practical when used to predict the properties of new samples. Therefore, the selection of response variables to be included in variables **A** should be carefully considered so that collinearity can be avoided.

3.2.2 Principal components regression (PCR)

Principal components regression (PCR) is based on the basic concept of principal component analysis (PCA). PCA is a method of data reduction or data compression that constructs a set of regression factors, also known as principal component (PC), such that the PCs are linear uncorrelated combinations of the original ones, with weight vectors that are orthogonal to each other. The first PC has the maximum variance which captures as much of the variability as possible in the entire original ones (Naes, Isaksson, Fearn, & Davies, 2002). The concept of PCR is that the properties of interest are regressed onto the PC scores of the measured variables instead of regressing the properties onto the original response variables. Similar to MLR, the least squares approach appearing in Equation (3.14) can be used to solve the regression vector, **b**, by defining the pseudoinverse of the response data matrix **A** as

$$\mathbf{A}^{+} = \mathbf{P}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}} \tag{3.17}$$

where **T** is the principal component scores and **P** is the loadings or the corresponding matrix.

Comparing the pseudoinverse for PCR in (3.17) to the pseudoinverse for MLR in (3.16), this operation is much more stable because of the orthogonality of the principal component scores, **T**, and loadings, **P** with respect to PCA. Also, PCR can avoid the underdetermined problem because the maximum possible numbers of PCs is less than or equal to the number of response variables and the number of calibration samples. PCR is less susceptible to overfitting than MLR because it directly addresses the collinearity problem that happens to MLR. However, if too many PCs are preserved in the PCR model, the overfitting problem still exists. Thus, determining the optimal number of PCs to retain in the model is an important part of PCR.

3.2.3 Partial least squares regression (PLS)

Partial least squares (PLS) regression is a well-established tool in chemometric analysis. This technique is commonly used in spectral quantitative analysis. Scientific research often involves using variables that are easily (or cheaply) measured to explain or predict the behavior of response variables that are often much more difficult (or expensive) to acquire. When the factors are many in number and are highly collinear such as in spectroscopy, PLS is a robust method used to construct predictive models. The advantage of the PLS is that the spectral and concentration information are included in

the calculation of the factors and the scores. It has the ability to maximize the spectral signals by its generated latent variables while minimizing less important variables and eliminating them from analysis Therefore, the resulting spectral vectors are directly related to the concentrations (Hoskuldsson, 1988).

In this study, the output data **y** is a one-dimensional vector. Therefore, the linear regression model is usually included in the relation

$$\mathbf{y} = \mathbf{Xb} + \varepsilon \tag{3.18}$$

where **X** is a *n* x *p* matrix containing *p* dependent variables of *n* samples, **b** is a *p* x 1 vector of regression coefficients obtained from PLS analysis, and **ε** is the model offset.

The solution of this linear regression model is firstly to compute the singular value decomposition (SVD) of **X**

$$\mathbf{X} = \mathbf{U}\,\Sigma\,\mathbf{V}^{\mathrm{T}} \tag{3.19}$$

where $\mathbf{U} = \left[u_1, u_2, \cdots, u_{N-p}\right]$ and is a (N-p) x 1 vector, and $\mathbf{V} = \left[v_1, v_2, \cdots, v_p\right]$ and is a p x 1 vector, with **U** and **V** are orthonormal matrices. The columns of **U** and **V** are called the left singular vector and right singular vector respectively. Σ is a diagonal matrix and its dimension is equal to the rank of **X**. Its diagonal elements $\{\lambda_i\}$ are known as singular values of **X** and are arranged in descending order of magnitudes

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_{\min(N-p,p)} \geq 0 \tag{3.20}$$

Thus SVD of **X** can be written as

$$\mathbf{X} = \sum_{i=1}^{p} \lambda_i u_i v_i^{T} \tag{3.21}$$

Multiplying $\mathbf{X}^T$ on both sides of (3.18) and neglecting the offset, it becomes

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T\mathbf{b} = \mathbf{V}\Sigma^2\mathbf{V}^T\mathbf{b} \tag{3.22}$$

Set $\mathbf{c} = \mathbf{X}^T\mathbf{y}$ and $\mathbf{A} = \mathbf{V}\Sigma^2\mathbf{V}^T$, this problem is equivalent to computing the least squares solution of the normal equation

$$\mathbf{A}\mathbf{b} = \mathbf{c} \tag{3.23}$$

Finally, the vector of regression coefficients, **b**, can be found by using the pseudoinverse of **A**

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{c} = \sum_{i=1}^{rk(\mathbf{A})} \frac{v_i u_i^T}{\lambda_i}\mathbf{c} \tag{3.24}$$

where rk(**A**) denotes the rank of a matrix **A.**

3.3 Performance evaluation

After a model of the above mentioned algorithms is computed, it is essential to determine its ability to predict unknown values. The performance of multivariate analysis for the prediction models is assessed by the root mean square error and the correlation coefficient, and they are illustrated in this section.

3.3.1 Room mean square error (RMSE)

In all of the measures considered, researchers attempt to estimate the average deviation of the model from the data. The root mean square error (RMSE) helps describe the fit of the model to the training data. It is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}{n}} \tag{3.25}$$

where $\tilde{y}_i$ are the values of the predicted result and n is the number of training samples.

RMSE is a measure of how well the model fits the data to the training data, and Equation (3.25) can also be applied to describe the root mean square error of prediction ($RMSE_p$) which is a measure of a model's ability to predict samples that are not used to build the model. In addition, the root mean square error of cross-validation ($RMSE_{cv}$) is used similar to prediction testing as it only tests predictors on data that are not used for training, but used for cross-validations that are done by successively omitting samples from the training set themselves. In the study, the "leave one out" cross-validation is used when performing the cross-validation test for the analysis.

3.3.2 Correlation coefficient (R)

Another commonly used objective measurement is the correlation coefficient (R) which is used for comparing the correlation between the predicted value and the actual value. It is a measure of how well the predicted values from a forecast model fit with the data.

The value is normally between magnitudes of 0 to ±1; zero means no relationship between the predicted values and the actual value, and one means perfect fit. As the strength of the relationship between the predicted values and actual values increases so does the correlation coefficient. Therefore, the higher the correlation coefficient the better the predicted model. The correlation coefficient is given by the formula:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.26}$$

where $\bar{x}$ and $\bar{y}$ are the samples means of variables *x* and *y*

Like the RMSE, Equation (3.26) can be applied to describe the correlation coefficient of prediction (R_p) and the correlation coefficient of cross-validation (R_{cv}), which is a measure of a model's ability to predict samples that are not used to build the model according to the same statement declared in previous section.

3.4 Summary

The most commonly used multivariate algorithms for the classification approach and regression approach were described in this chapter. In addition, the root mean square error and correlation coefficient, which were introduced to tell us how good the measurements are, are used in the following performance comparison. This study will present an improved algorithm by applying the Monte Carlo method to PLS algorithms that demonstrate better performance and stability than the algorithms discussed in this chapter.

CHAPTER 4

DATA ACQUISITION

This study was based on clinical validation of blood glucose levels between laboratory results and absorption spectroscopy analysis. In the in-vitro validation process, the human blood glucose level was employed as the gold standard, with the spectrometry method used for the testing. This chapter presents the description of data collection, which describes the blood glucose distribution among the subjects, and evaluates the reproducibility of blood glucose measurements using the data from two ISO (International Organization for Standardization) accredited laboratories.

4.1 Setting

The study was conducted in the Integrative Health Clinic (IHC) of the School of Nursing (SN) of The Hong Kong Polytechnic University (PolyU) from 8 December 2008 to 17 January 2009. This was approved by the Human Subjects Ethics Committee of The Hong Kong Polytechnic University (approval number: HSEARS20080108001).

4.2 Subjects

Subjects were recruited from the community by convenience. Several strategies were employed for recruitment, including the use of self-help groups, professional organizations, and the mass media. The poster for internal recruitment and newspaper advertisement is attached in Appendix B-1. All subjects who were interested in the study and able to provide a voluntary informed consent were recruited to the study. All known infectious disease subjects were excluded.

4.3 Sample size

The sample size (N) required was estimated at 385, based on the sample size formula shown in Equation (4.1), so as to achieve a precision of 95% in a wide range of blood glucose levels for the study.

$$\mathrm{N} = \frac{z^2 \mathrm{p}(1-\mathrm{p})}{e^2} \tag{4.1}$$

where z is the upper quantile order of the standard normal distribution, p is the estimated proportion of an attribute that is present in the population, and e is the desired level of precision. In this study, z is equal to 1.96 for 95% confidence interval, p is equal to 0.5 for maximum variability, and e is equal to 0.05 for ±5% precision.

4.4 Instruments

The major equipment in the experiment was the NIR spectrometer, the probe, and the tungsten halogen light source. A digital thermometer was used to measure skin temperature for every subject. The following equipment was used in the clinical trial.

4.4.1 Spectrometer

The NIR spectrometer from Control Development, Inc. (CDI, Foundation Drive, South Bend, IN) was used in the trial. The specification of the spectrometer is listed in Table 4.1 for easy reference. This model was selected due to its low cost and its wide range of coverage.

Brand	**CDI**
Model	NIR-256-L-2.2T2
Spectral Range	1,082 – 2,237 nm
Spectral Resolution	6 nm
Detectors	Dual stage TE-cooled 256 element array
Scanning Time	20 milliseconds
Wavelength Accuracy	± 0.5 nm
Photometric Repeatability	<0.001 AU
Max Integration time	20 ms
Noise Level	<50 counts p-p
Measurement	Absorbance

Table 4.1 Specification of the CDI spectrometer

4.4.2 Reflection probe

The reflection probe from Avantes BV (The Netherlands, Europe) was used as the connection between the spectrometer and the light source in the trial. A standard SMA905 connector light is coupled into a fiber bundle consisting of 6 fibers and carried to the probe end. The surface will reflect light back into the seventh fiber and data can be transferred to a spectrometer for analysis. The specification of the probe is listed in Table 4.2 for reference.

Brand	**Avantes**
Model	FCR-7IR400-2-ME
Fibers	7 fibers 400 μm core, 6 light-fiber, 1 read fiber
Wavelength Range	350 – 2,000 nm (VIS/NIR)
Probe End	Stainless steel cylinder, 100mm long x 2.5mm diameter
Tubing	The optical fibers are protected by a Kevlar reinforced PTFE tubing with PVC sleeving
Temperature	-20°C to 65°C
Bending	Minimum bend radius: long term: 60mm

Table 4.2 Specification of the Avantes reflection probe

4.4.3 Light source

An illumination light source is needed for absorption spectroscopic setup. A tungsten halogen lamp is the most typical NIR light generator because it generates continuous NIR spectral light with wavelengths ranging from 360 – 2,500nm. The Avantes BV tungsten halogen light source was used in the study. The specification of the light source is listed in Table 4.3 for reference.

Brand	**Avantes**
Model	AvaLight-HAL
Light Source	10W Tungsten Halogen Lamp, fan-cooled
Wavelength Range	360 – 2,500 nm
Stability	±0.1%
Time to Stabilize	15 minutes
Optical power 200µm fiber	0.5 mWatt
Optical power 600µm fiber	4.5 mWatt
Bulb Color Temperature	2850K

Table 4.3 Specification of the Avantes light source

By integrating the above equipment, the full set of assembled equipment is shown in Figure 4.1. The figure shows there is a finger clip at the end of the probe when measuring using fingers so as to provide a constant force and to fix the scanning position.

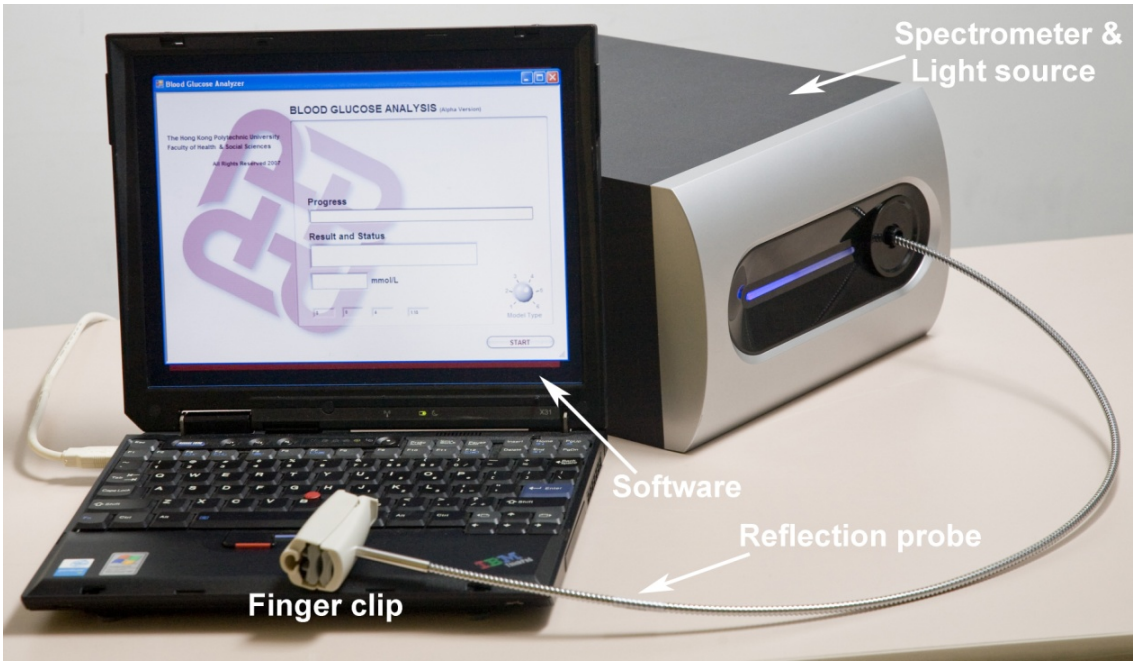


Figure 4.1 The assembled NIR spectrometer in the study

4.4.4 Thermometer

The Fluke 52 II dual input digital thermometer shown in Figure 4.2 was used for skin temperature measurement. A K-type probe was used for direct contact with the skin surface for measurement. The specification of the thermometer is listed in Table 4.4 for reference.

Brand	**Fluke**
Model	Fluke 52 Series II
Temperature Accuracy	Above -100°C: ±[0.05% + 0.3°C]
Temperature	K-type probe: -200°C to 1372°C
Temperature Scale	ITS-90
Display Resolution	0.1°C, 0.1K < 1,000 1°C, 1K ≥ 1,000

Table 4.4 Specification of the Fluke thermometer

Figure 4.2 Fluke 52 Series II

4.5 Procedure

Consent to conduct the study was sought prior to data collection (Appendix B-2). The purpose and procedure of the study were explained to all subjects and an information sheet (Appendix B-3) was given to ensure that there would not be any risk of harm to them. All subjects were asked to fast overnight before participating in the study.

Upon arrival at the IHC, subjects were asked to complete a demographic data sheet, health history, and medication history. The skin temperature of the right and left fingers (palmar surface of distal phalanx), right and left forearms (mid-way between the radial side of the first wrist crease and the medial side of the elbow), and right and left ear lobes were then measured. Spectra were obtained using the spectrometers at the sites mentioned. The assembled NIR spectrometer was used to scan the right and left finger tips (10 spectra were obtained in total), right and left wrist, right and left forearms (mid-way between the radial side of the first wrist crease and the medial side of the elbow), and right and left ear lobes (6 spectra were obtained in total). Simultaneous veni-punctures for blood glucose (12 ml, blood glucose testing in fluoride bottles) were performed. Since the aim was to collect high blood glucose levels as much as possible in this study, all subjects with diabetes mellitus (DM) history were required to have blood taken one more time, at least 30 minutes after they had had their breakfast. The non-DM subjects were asked to give their consent before being invited to have breakfast (with calories ranging from 233 to 298 Kcal), and then had their blood taken for glucose measurement. Unlike the DM subjects, this was optional for non-DM subjects.

In addition, a paired test was carried out to ensure the reproducibility of the lab testing method. Subjects were invited to participate in this procedure using separate consent because one more blood specimen was needed for this paired test. With their consent, two specimens were taken for the blood glucose measurement, i.e. double the amount of blood being drawn. The two specimens were sent to the two ISO-accredited laboratories for testing.

A meal coupon of HKD$20 was given to each of the subjects as a token of appreciation for their participation upon completion of the whole procedure.

4.6 Demographic profile

Five hundred and twelve subjects (225 male and 287 female) voluntarily participated in the study. The mean age was 52.33 (SD12.8). Among them, 219 (42.8%) suffered from DM. Of these, 44 (20.2%) did not have any co-existing disease. The ten drugs most commonly consumed by the subjects are listed in Table 4.5.

Drug	Frequency (by subject)
Metformin	131
Gliclazide	69
Nifedipine	39
Simvastatin	26
Glibenclamide	24
Aspirin	24
Metoprolol	19
Atenolol	18
Perindorpil	17
Lisinopril	17

Table 4.5 The ten drugs most commonly consumed by subjects

Ideally, there would be 1,024 (512x2) NIR spectra samples because blood taking was conducted twice (i.e. before and after breakfast) for all subjects. However, some of the normal subjects were not willing to have a post-breakfast blood sample taken, and some of the DM subjects also declined to have the post-breakfast blood sample taken for reasons of their own discretion; for example, feelings of inconvenience, change of personal schedule. As a result, a total of 850 specimens were collected. Table 4.6 shows the distribution of the blood specimens.

	Fasting blood glucose	Post-meal blood glucose
DM	219	217
Non-DM	293	121

Table 4.6 Distribution of the blood specimens (N=850)

The mean (SD) for HbA1c, fasting blood glucose and post-meal glucose from PHC Medical Diagnostic Centre Ltd (PHC) and PathLab Medical Laboratories Ltd (PathLab) are shown in Table 4.7.

		DM (n=219) (mean, SD)	Non-DM (n=293) (mean, SD)
PHC (219 specimens)	HbA1c	7.30 (1.26)	5.82 (2.88)
	Fasting blood glucose	7.55 (2.11)	4.93(0.68)
	Post-meal blood glucose	11.73 (3.15)	7.56 (1.77)
PathLab (60 specimens)	Fasting blood glucose	6.97 (1.91)	4.56 (0.60)
	Post-meal blood glucose	11.08 (2.92)	6.85 (1.74)

Table 4.7 The mean (SD) for HbA1c, fasting blood glucose and post-meal glucose by laboratory (N=512)

Samples with blood glucose levels (BGL) ranging from 3.4 to 24.9mmol/L were collected in this clinical trial. Figure 4.3 shows the histogram of the BGL and Table 4.8 shows the distribution of the BGL.

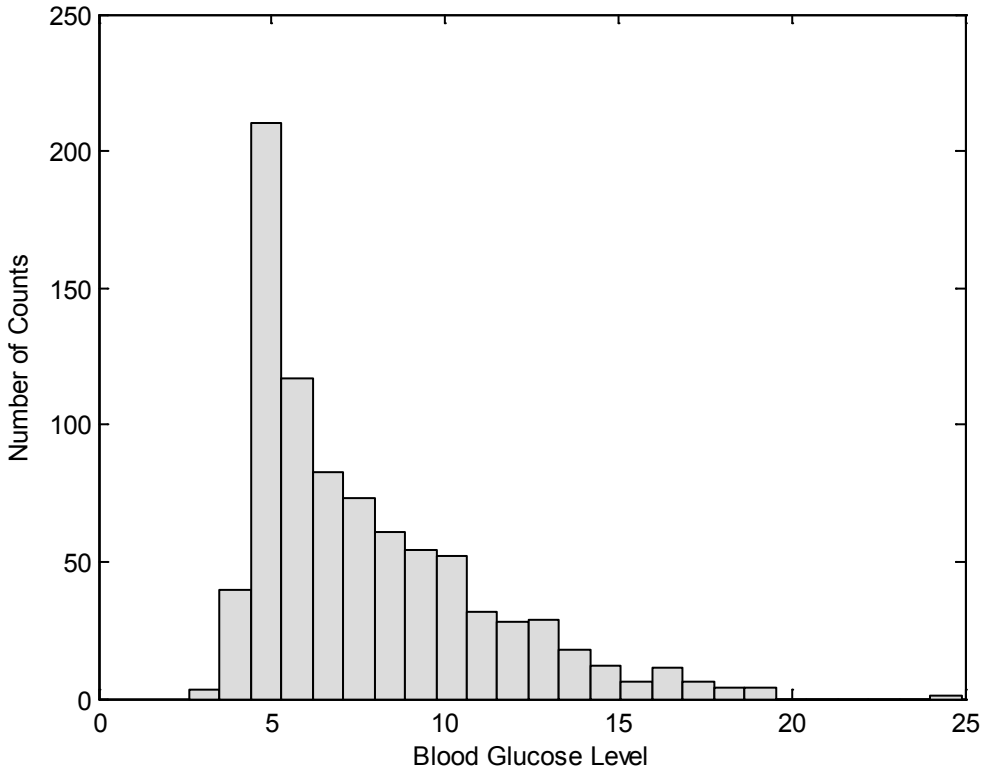


Figure 4.3 Histogram of BGL ranging from 3.4 – 24.9mmol/L (N=850)

Glucose range	Number of counts
2.5 – 3.4	3
3.5 – 4.4	65
4.5 – 5.4	218
4.5 – 6.4	114
6.5 – 7.4	96
7.5 – 8.4	62
8.5 – 9.4	67
9.5 – 10.4	64
10.5 – 11.4	38
11.5 – 12.4	31
12.5 – 13.4	32
13.5 – 14.4	19
14.5 – 15.4	11
15.5 – 16.4	14
16.5 – 17.4	7
17.5 – 18.4	4
18.5 – 19.4	4
19.5 – 20.4	0
20.5 – 21.4	0
21.5 – 22.4	0
22.5 – 23.4	0
23.5 – 24.4	0
24.5 – 25.4	1
Total:	850

Table 4.8 Distribution of BGL ranging from 3.4 – 24.9mmol/L (N=850)

As mentioned previously, a total of 16 locations of the body were scanned for each subject. However, due to complications arising from feelings of inconvenience, change of personal schedule, missing fingers, among other things, some positions may not have full records for all subjects and Table 4.9 shows the distribution of all spectra that were collected in this study. The locations of the body are abbreviated and shown in the table for easy reference. They will be used in later chapters.

Location of body	Abbreviation	Number of samples collected
Left thumb	FL1	846
Left index finger	FL2	846
Left middle finger	FL3	844
Left ring finger	FL4	844
Left little finger	FL5	844
Right thumb	FR1	846
Right index finger	FR2	844
Right middle finger	FR3	845
Right ring finger	FR4	846
Right little finger	FR5	846
Left wrist	LW	845
Right wrist	RW	846
Left forearm (mid-way between the radial side of the first crease and the medial side of the elbow)	LA	846
Right forearm (mid-way between the radial side of the first crease and the medial side of the elbow)	RA	846
Left ear lobe	LE	846
Right ear lobe	RE	845

Table 4.9 The distribution of all spectra in the clinical trial

4.7 Reproducibility of blood glucose measurement between the two laboratories

The reproducibility test of blood glucose measurement was carried out in accordance with ISO 15197 (International Organization for Standardization, 2003). Two ISO-accredited laboratories (ISO 15189) were engaged to conduct the glucose measurement. The two accredited laboratories concerned were PHC Medical Diagnostic Centre Ltd (PHC) and PathLab Medical Laboratories Ltd (PathLab).

One hundred and fifty-seven subjects participated in the paired test. The blood specimens were sent to two laboratories, PHC and PathLab. The mean blood glucose measured by PHC was 5.76 (SD 1.73), compared to 5.50 (SD 1.73) for PathLab. Figures 4.4 and 4.5 show the box plot of the fasting blood sugar from PHC and PathLab respectively.

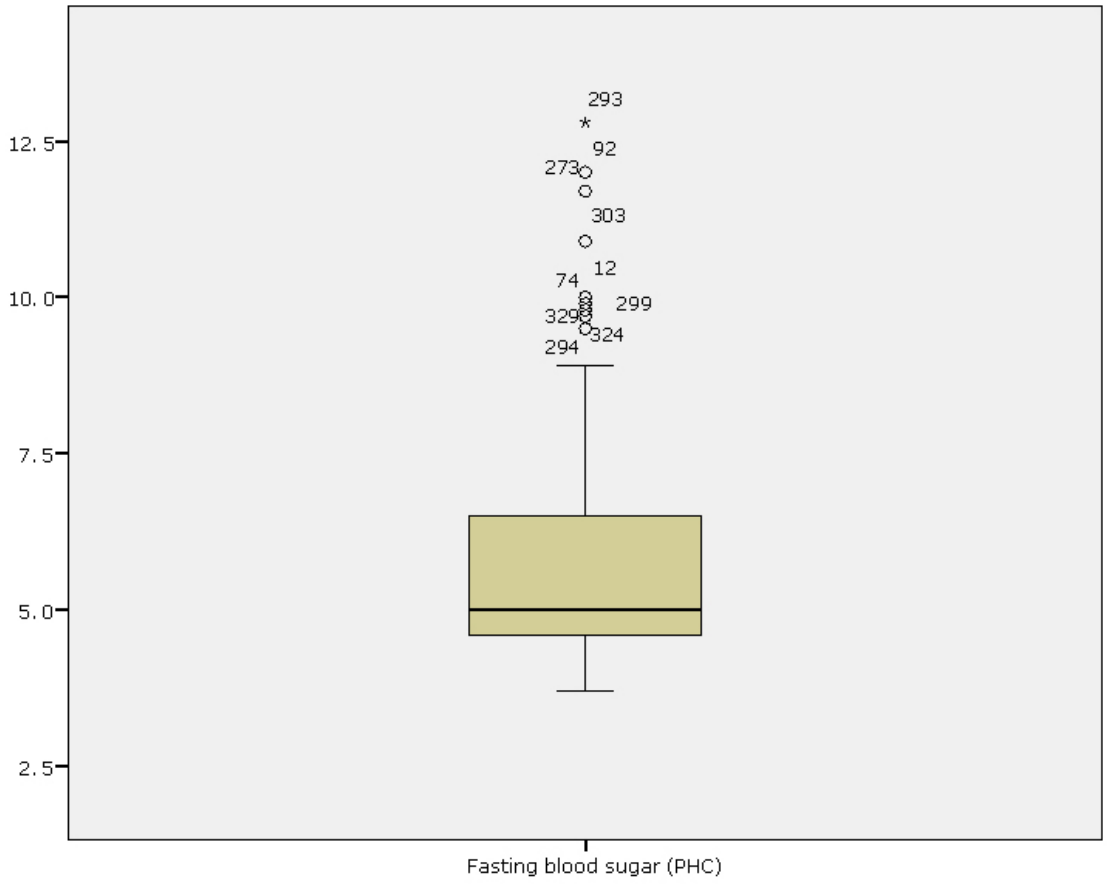


Figure 4.4 Box plot for fasting blood sugar from PHC

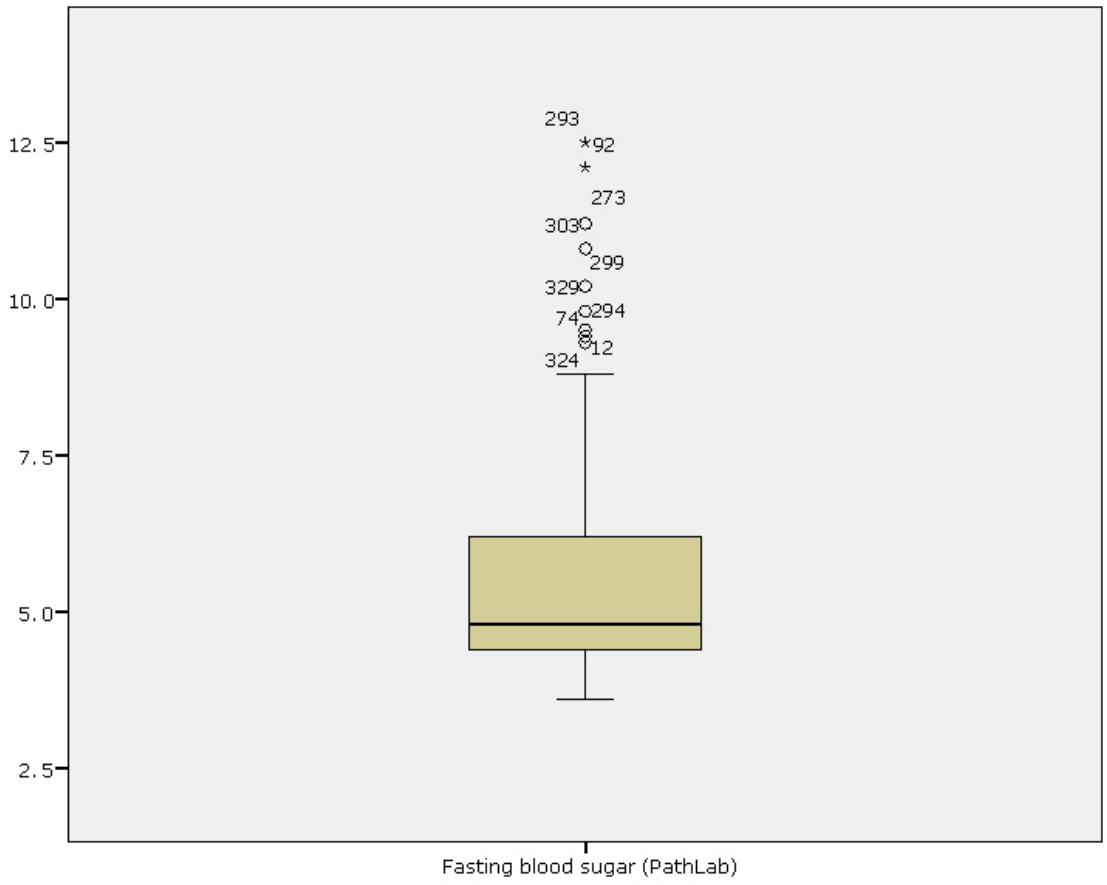


Figure 4.5 Box plot for fasting blood sugar from PathLab

The values of the fasting blood glucose level from PHC were plotted against PathLab in Figure 4.6. The results show that the data were centralized along a line with a slope of nearly 1.

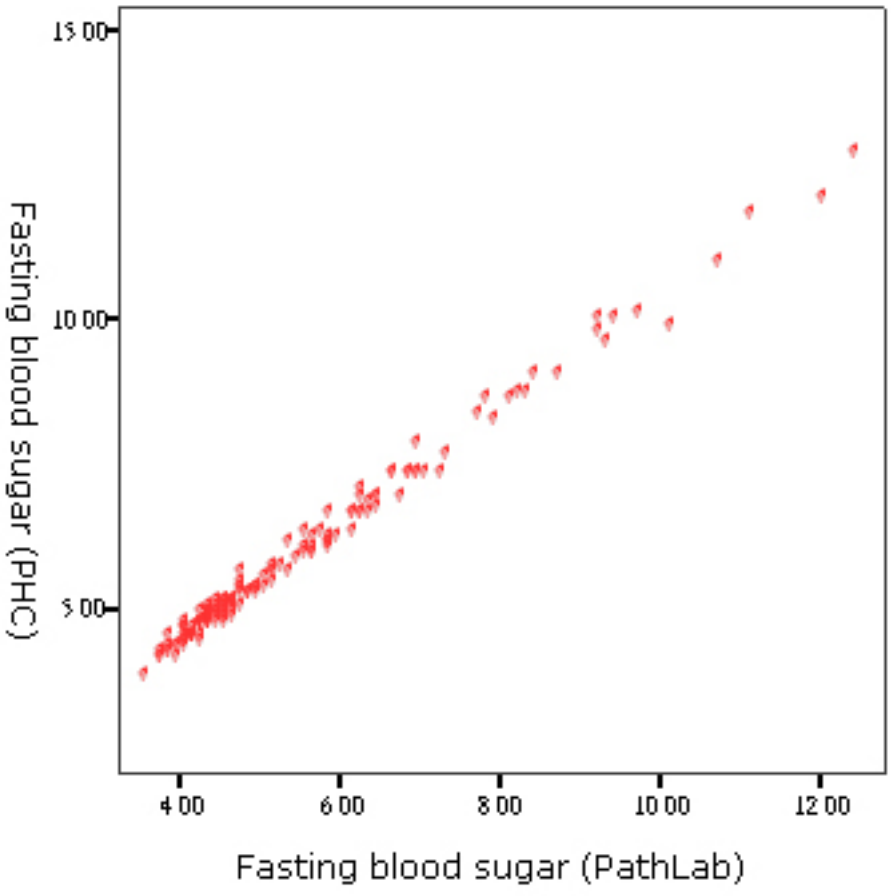


Figure 4.6 Scatterplot of fasting blood glucose results from the two laboratories

The difference in the fasting blood glucose levels from both laboratories was plotted in Figure 4.7 against their mean value (0.26). It was found that 95% of the data were clustered around the horizontal line of the mean value. The standard deviation of the difference was 0.16, meaning that 95% of the data of the fasting glucose level as measured by PHC was 0.58 higher or 0.06 lower than that measured by PathLab. The reference ranges of serum glucose level were 3.9 to 6 and 3.6 to 7.8. Thus, there is hardly any clinical significance between the results from the two laboratories.

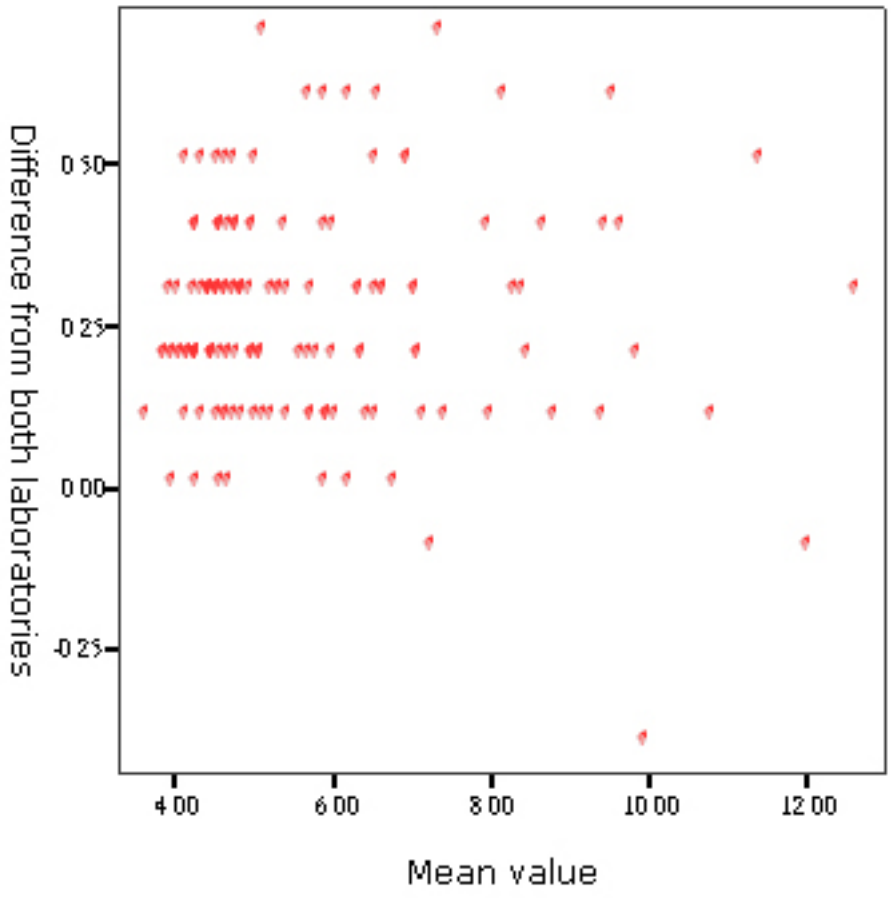


Figure 4.7 Scatterplot including the mean difference in the fasting blood glucose level from both laboratories

4.8 Summary

The data collection was successfully completed. This clinical trial managed to recruit subjects with a wide range of glucose levels (3.4 – 24.9mmol/L). The widely spread glucose readings facilitated accurate computation in the study.

The reproducibility of the blood glucose measurement by the two ISO-accredited laboratories was established. Thus, one of the criteria of the requirement for glucose meter production as stated in ISO 15197 had been met: ‘In vitro diagnostic test system – requirements for blood-glucose monitoring for self-testing in managing diabetes mellitus’. Also, this confirms the accuracy of our gold standard so that further analysis can be done to improve the measurement.

CHAPTER 5

DATA PREPROCESSING

There are many studies trying to apply NIR spectroscopy for non-invasive blood glucose monitoring. Most of the studies were either performed in experimental settings or under a strictly controlled environment: for example, the measurement of subjects' spectra under a fixed room temperature and moisture; a pre-defined constant force was applied to the location of tissue with particular location of body. These settings are barriers against the use of non-invasive technology in real applications as home-based medical devices because room temperature and moisture frequently vary in home environments.

Undoubtedly, the body locations used to measure the spectra and the temperature of the human tissue are critical parameters for NIR spectroscopic analysis of blood glucose monitoring. The body location strongly affects the measurement of absorption because it influences the overall output reading, which may possibly dictate the signal-to-noise (SNR) of the measurement. Moreover, changes in temperature alter the extent of hydrogen bonding, which may cause shifts in the band positions. These bands shift to higher frequencies at higher temperatures. Ideally, non-invasive clinical measurement by NIR spectroscopy is independent of sample temperature, which means the prediction

model should be insensitive to baseline variations that can be several orders of magnitude larger than the size of the analyte absorption bands.

This chapter evaluates measurements on 16 measurement sites for non-invasive blood glucose monitoring. The right and left fingers, right and left wrist, right and left forearms (mid-way between the radial side of the first wrist crease and the medial side of the elbow), and right and left ear lobes (16 measurement sites in total) were examined. These sites are evaluated on the basis of their chemical and physical properties that are relevant to the non-invasive measurement of glucose. The root mean square error of cross validation ($RMSE_{cv}$) and the correlation coefficient of cross validation (R_{cv}) are used to evaluate the measurement sites. In addition, the various temperature combination models for the prediction models are evaluated and the results of the $RMSE_{cv}$ and R_{cv} are used to identify the temperature range that is suitable for measurement.

Pre-processing of spectral data is an important procedure before chemometric analysis according to the four-stage framework of bio-signal processing because scaling differences arise from path-length effects, scattering effects, source variations, or other instrumental sensitivity effects in NIR spectroscopy that will influence the recorded sample. The objective of pre-processing is to remove unwanted spectral variation in order to improve the output results. Various pre-processing techniques such as normalization process and pre-filtering process are explored and evaluated.

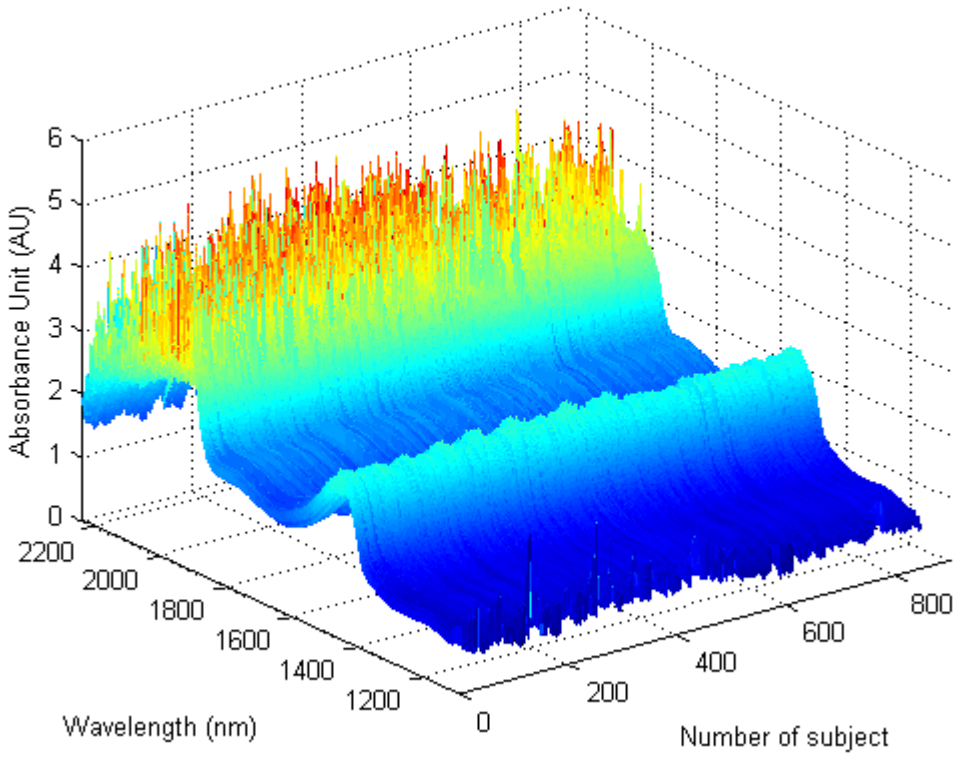
5.1 Wavelength consideration

The molecular formulation of glucose is $C_6H_{12}O_6$ and it can exist in several different forms in humans. Different isomers of glucose can inter-convert among themselves (Koolman & Roehm, 2005). NIR spectroscopy is based on molecular overtone and combination vibrations. The CH, OH and CH_2 groups can be stimulated by NIR spectroscopy causing vibrations in carbon bonds and subsequently release energy to form overtones (Arnold, Harvey, McNeil, & Hall, 2002). Table 5.1 shows the four bands of overtone regions that can indicate the existence of the above mentioned groups.

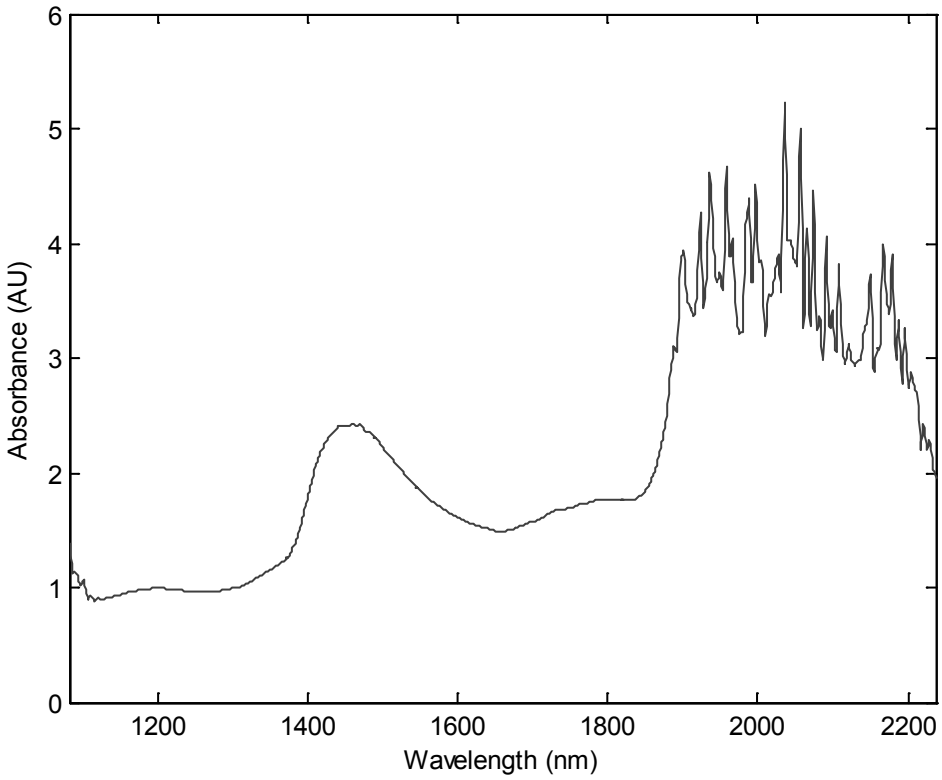
Band	Wavelength region
Combination band region	1,950 – 2,500 nm
First overtone region	1,475 – 2,050 nm
Second overtone region	1,050 – 1,650 nm
Third overtone region	700 – 1,150 nm

Table 5.1 Four overtone regions in NIR spectroscopy

In this study, the CDI spectrometer ranging from 1,082 – 2,237nm, which covers first and second overtone regions and nearly half of the combination band region, was used to measure the NIR spectra for all subjects. Figure 5.1a shows the full spectra of left middle finger (FL3) for all subjects (N=844) and Figure 5.1b shows the full spectra of FL3 for a subject, since all the spectra recorded are similar so the spectra of FL3 only is shown here.



(a) All subjects (N=844)

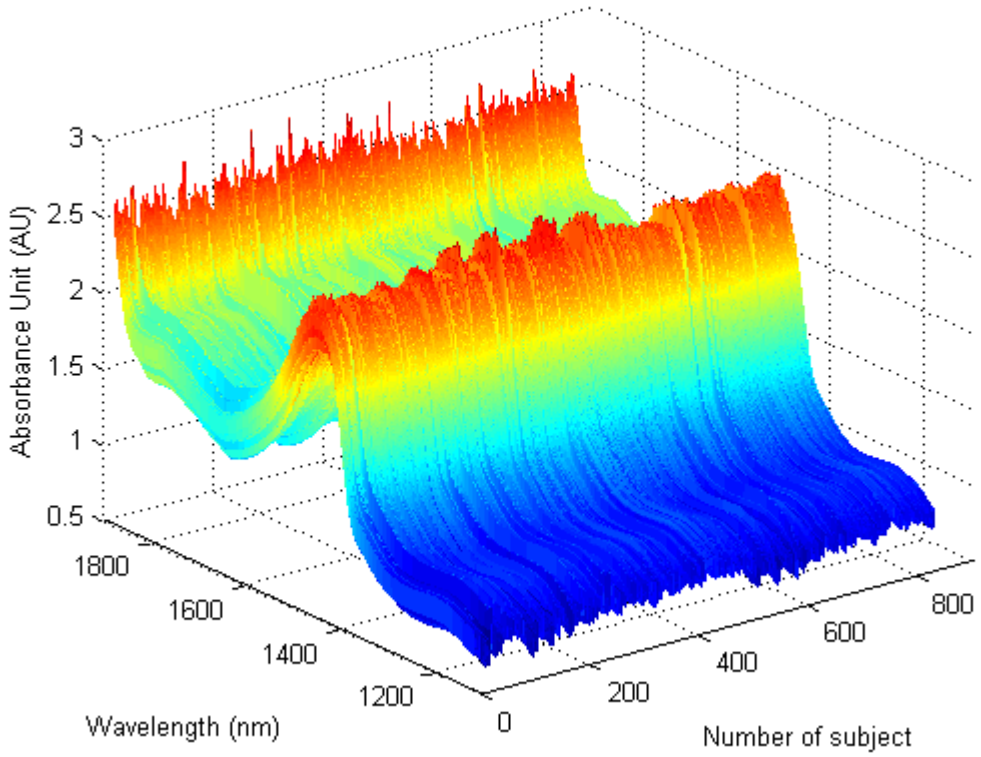


(b) Single subject

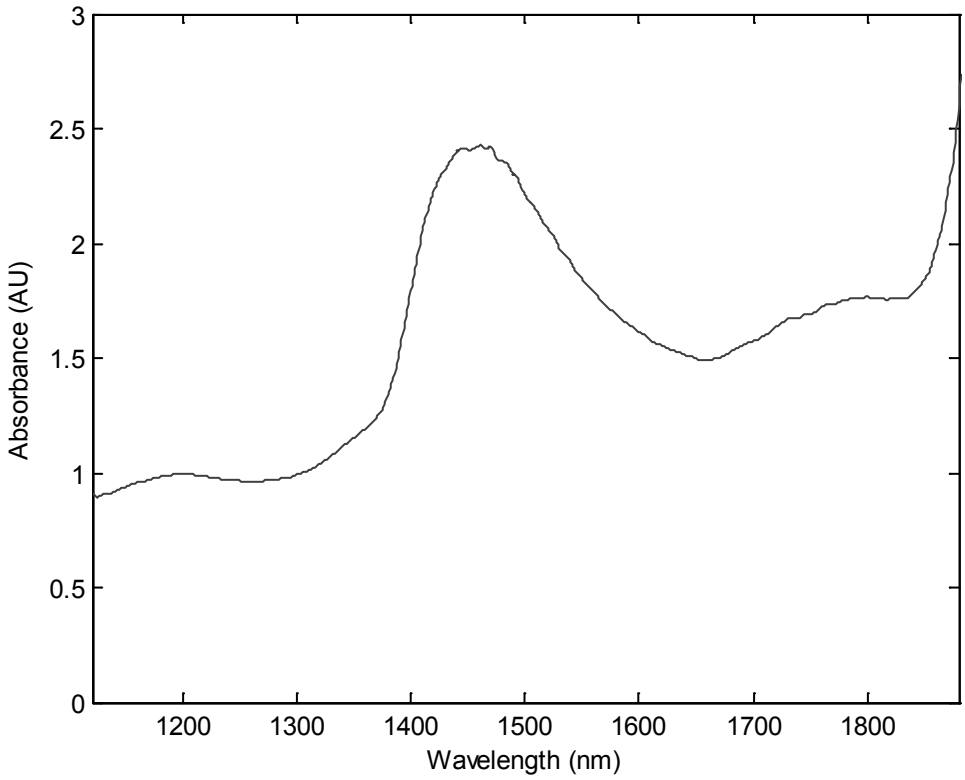
Figure 5.1 Full spectra of FL3 (1,082 – 2,237nm)

Obviously, noises exist in the measurements (the beginning and ending of the scanned spectra) as shown in Figure 5.1. Therefore, these data are not suitable for analytical use, particularly since the noise level is even higher than the gain of the measurand in the combination-bands region (i.e. 1,950 – 2,237nm). Although the combination bands region may contain some important data, these data should be ignored because of this limitation caused by the spectrometer. The samples, however, contain the first and second overtone NIR spectra that serve the goal of this study. Hence, this study considers the wavelength range between 1,121 – 1,880nm for analysis. Figure 5.2a

shows the selected wavelength region for FL3 for all subjects, and Figure 5.2b shows the selected wavelength region of FL3 for a single subject.



(a) All subjects (N=844)



(b) Single subject

Figure 5.2 Full spectra of selected wavelength region (1,121 – 1,880nm)

5.2 Normalization

In many analytical methods, variables measured for a given sample are subject to overall scaling or gain effects. The measured variables are increased or decreased from their true value by a multiplicative factor, and each sample can experience a different level of multiplicative scaling. In spectroscopic measurement, scaling differences arise from path length effects, scattering effects, source variations, or other instrumental sensitivity effects (Martens & Naes, Multivariate calibration, 1989). Under these circumstances, it

is often the relative value of variables that should be used during multivariate analysis rather than the absolute measured value.

A normalization process attempts to correct for these kinds of effects by identifying some aspect of each sample that should essentially be constant from one sample to the next, and correcting the scaling of all variables based on this characteristic. This process helps to provide all samples with an equal weighting in the model. Some samples may have severe scaling effects if they do not attempt the normalization process. Even worse, some samples will not be considered as important information that can contribute to properties of interest by many multivariate techniques. Therefore, the normalization process can be very useful for obtaining accurate and robust PLS models for the study (Wolthuis, Tjiang, Puppels, & Schut, 2006). This section briefly discusses four generally-used spectral normalization methods, and the one that provides the minimum $RMSE_{cv}$ and the highest R_{cv} will be used in the current study.

5.2.1 1-norm

The normalization process of each sample is a common approach to the multiplicative scaling problem. The norm of a matrix is a measure of its size. Several different types of norms exist. 1-norm is used to divide each variable by the sum of the absolute value of all variables for the given samples (Suli & Mayers, 2003). It returns a vector with unit area; that is, the area under the curve is equal to 1. It is defined as follows:

$$\|\mathrm{X}_i\| = \sum_{j=1}^{n} |x_{i,j}| \tag{5.1}$$

where $\|\mathrm{X}_i\|$ is the normalization weight for sample *i*, x_i is the vector of observed values for the given sample, *j* is the variable number, and *n* is the total number of variables.

5.2.2 Euclidean norm (2-norm)

The most commonly used norm in 2-dimensional Euclidean space is the Euclidean norm or 2-norm (Suli & Mayers, 2003). It normalizes to the sum of the square value of all variables for the given samples and returns a vector of unit length; that is, length is equal to 1. 2-norm is a form of weighted normalization where larger values are weighted more heavily in the scaling. It is defined as follows:

$$\|\mathrm{X}_i\| = \sum_{j=1}^{n} {x_{i,j}}^2 \tag{5.2}$$

where $\|\mathrm{X}_i\|$ is the normalization weight for sample *i*, x_i is the vector of observed values for the given sample, *j* is the variable number, and *n* is the total number of variables.

5.2.3 Standard Normal Variate (SNV)

The Standard Normal Variate (SNV) normalization process is a weighted normalization (Barnes, Dhanoa, & Lister, 1989). It is different from the 1-norm or 2-norm mentioned above because not all samples contribute to the normalization equally. SNV calculates

the standard deviation of all variables for the given sample and the entire sample is then normalized by this value, hence giving the sample a unit standard deviation; that is, the standard deviation is equal to 1. The SNV scaling is defined as follows:

$$\|\mathrm{X}_i\| = \frac{x_{i,j} - \bar{x}_i}{\sigma_{x_i}} \tag{5.3}$$

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{i,j}}{n} \tag{5.4}$$

$$\sigma_{x_i} = \sqrt{\frac{\sum_{j=1}^{n} \left(x_{i,j} - \bar{x}_i\right)^2}{n-1}} \tag{5.5}$$

where $\bar{x}_i$ is the mean of samples *i*, σ_{x_i} is the standard deviation for samples *i*, and *n* is the total number of variables in the samples.

However, care should be taken when using the SNV as the normalization process. This approach is an empirical normalization process that relies on their mean and standard deviation in the entire samples. This process can work well if most of the signal in a sample is the same in all samples. However, it may introduce a non-linearity in the relation between the sample and the analyte concentration if the measurement signal varies significantly from sample to sample (Fearn, Riccioli, Garrido-Varo, & Guerreo-Ginel, 2009).

5.2.4 Multiplicative Scatter Correction (MSC)

Multiplicative scatter correction (MSC) is another method for normalization (Geladi, MacDougall, & Martens, 1985). It is based on the idea of correcting the scatter level of all spectra of a group of samples to the level of an average spectrum (Naes, Isaksson, Fearn, & Davies, 2002). This is possible because the wavelength dependency of the light scatter is different from that of the light absorption of the analyte concentration. Each spectrum is fitted to the average spectrum as closely as possible by least squares equation:

$$x_i = a_i + b_i \bar{x}_j + e_i \tag{5.6}$$

where x_i is the vector of observed values for the given sample, $\bar{x}_j$ is the mean of samples j, and e_i is the residual spectrum

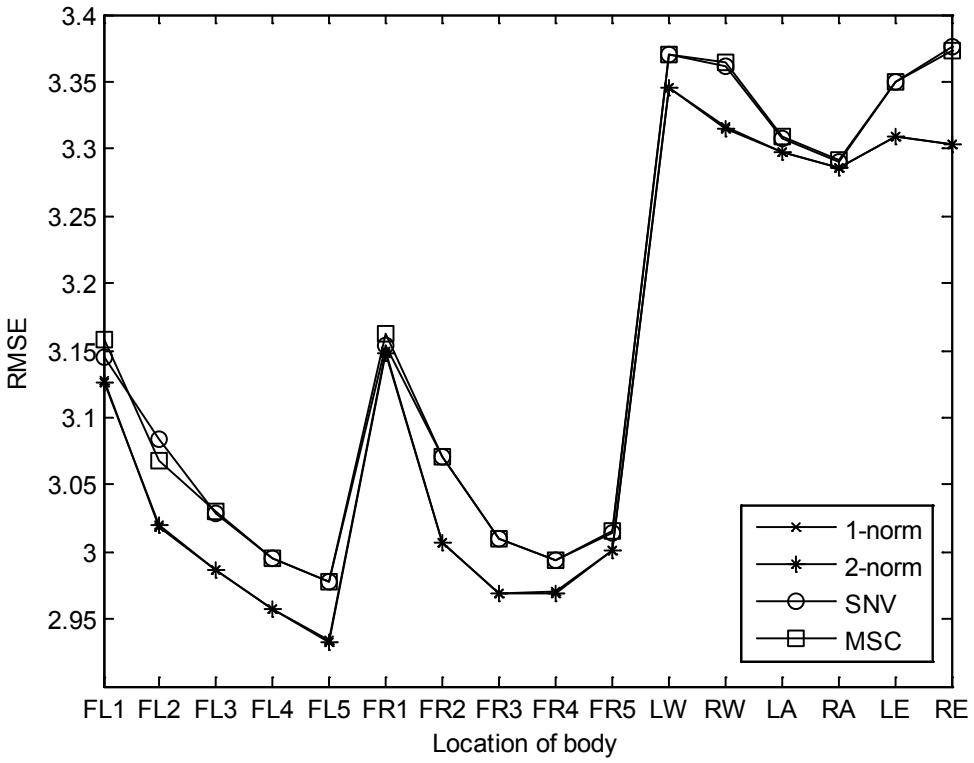
The corrected spectrum for the normalization process is calculated using the intercept, a_i, and the slope, b_i, as follows:

$$\|\mathrm{X}_i\| = \frac{x_i - a_i}{b_\mathrm{i}} \tag{5.7}$$

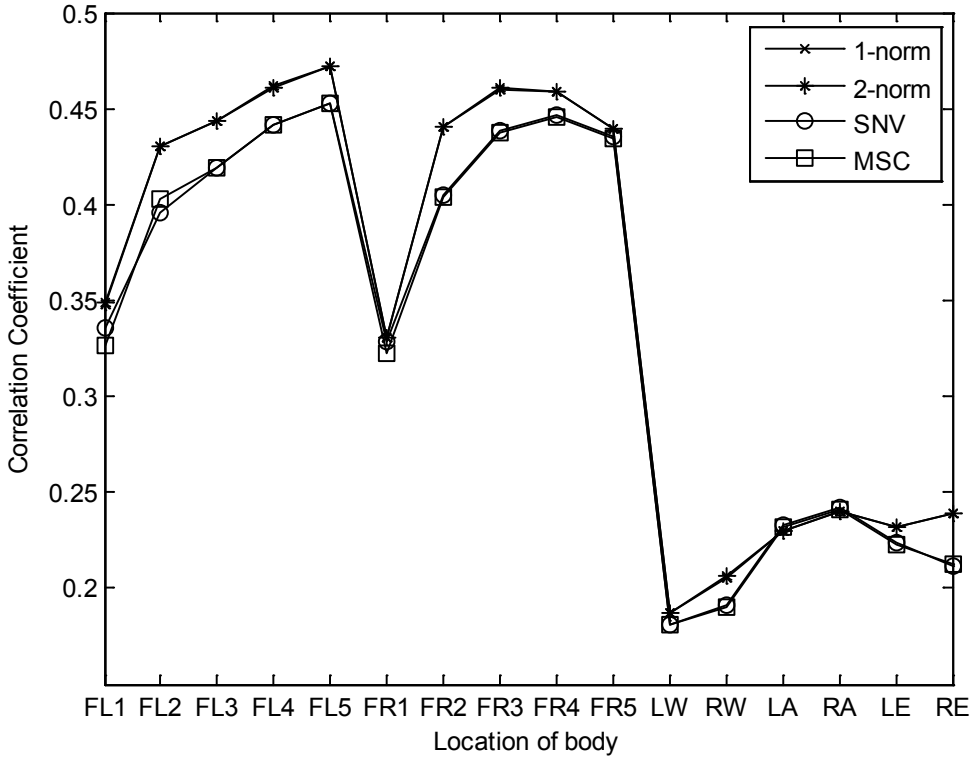
The resulting set of corrected spectra have a non-zero mean and a variance related to the set mean spectrum variance. The research group of Fearn have reported that MSC and SNV are linearly related, which therefore implies MSC has the non-linearity issue as well (Fearn, Riccioli, Garrido-Varo, & Guerreo-Ginel, 2009). Moreover, MSC is a set dependent transformation. If the raw data set is modified, the mean spectrum is likely to change and the MSC corrected spectra will therefore need to be recalculated.

5.2.5 Performance comparison

The PLS regression model mentioned in section 3.2.3 to evaluate the RMSE and the correlation coefficient, R, discussed in section 3.3, was used to study the performance of various normalization processes. The PLS regression model was constructed with the latent variables set equal to 10 in this simulation. The $RMSE_{cv}$ and R_{cv} of the 16 locations of body mentioned in Table 4.9 are plotted in Figure 5.3 to study the overall performance of various normalization processes.



(a) $RMSE_{cv}$



(b) R_{cv}

Figure 5.3 Performance of various normalization processes in the 16 locations of body

The results in Figure 5.3 show that the 1-norm and 2-norm approaches provide comparable outcome while SNV and MSC provide similar outcome. Furthermore, the 1-norm and 2-norm approaches had smaller RMSE and higher R than SNV and MSC. Therefore, based on these findings, the body location FL3 was chosen specifically for data acquisition, where the spectral datasets were obtained for analysis. The "leave one out" cross-validation was performed to test the robustness of the normalization process. The RMSE of the training sets and $RMSE_{cv}$ of cross-validation are shown to determine which process provides the minimum RMSE and $RMSE_{cv}$. In addition, the R for the

training set and R_{cv} for cross-validation show how well the prediction model can perform. Table 5.2 shows the results for the four normalization processes.

Normalization process	RMSE	$RMSE_{cv}$	R	R_{cv}
1-norm	2.8208	2.9868	0.5266	0.4435
2-norm	2.8206	2.9865	0.5267	0.4436
SNV	2.8455	3.0291	0.5381	0.4190
MSC	2.8460	3.0297	0.5144	0.4187

Table 5.2 The results for the four normalization processes

The results shows that 2-norm performs best, and that 1-norm and 2-norm provide very similar RMSE and R both for training set and cross-validation, while SNV and MSC deliver relatively higher RMSE than 1-norm and 2-norm and the correlation coefficients are relatively smaller than they are. Analysis of variance (ANOVA) was applied to the data obtained at FL3 in order to evaluate the performance of the normalization methods. The result shows that the p-value is equal to 1.000 which denotes that there is not significant between these normalization methods. Table 5.3 shows the ANOVA results for the four normalization processes.

	SS	***df***	***p-value***
Between groups	0.000	3	1.000
Within groups	10066.529	3372	
Total	10066.529	3375	

Table 5.3 The ANOVA results for the four normalization processes

Nevertheless, since 2-norm is a form of weighted normalization where larger values are weighted more heavily in the scaling and it is also the most commonly used

normalization process in 2-dimensional Euclidean space, 2-norm was used in the following analysis to take advantage of the uneven weighting mechanism.

5.3 Location of the body

Sixteen measurement sites were evaluated for non-invasive blood glucose monitoring as shown in Table 4.9. They included ten fingers, left and right wrists, left and right forearms, and the left and right ear lobes. Fingers are the most convenient location for glucose measurement, and this experiment applied the idea to measure blood oxygen saturation using a finger clip to provide constant force and relatively fixed positions on the fingers when performing measurements. Figure 5.4 showed the setup of the finger clip to the equipment. With the finger clip, the measurement is comparatively easy to handle and can reduce human errors.

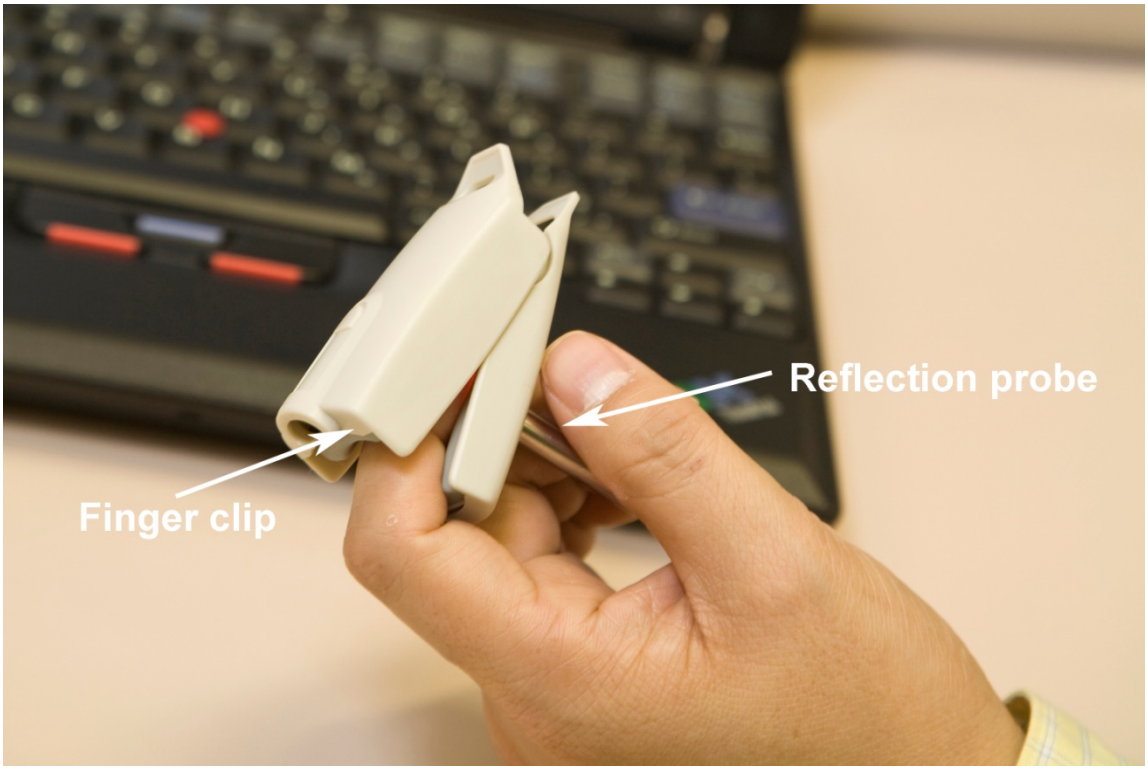


Figure 5.4 Finger clip used in measuring NIR spectra

The forearm, the wrist, and the ear lobes are also easy locations for glucose measurement though the force applied to these locations may vary and the position may not be exactly the same for all subjects. It is agreed that locating the same position for all subjects are difficult and therefore, we clearly defined the position of these six measurement sites before we start collecting the data in the trials. In addition, we designated a technician who assisted in this experiment tried to locate the same position on forearms and ear lobes to avoid inconsistency.

5.3.1 Performance comparison

The same setting in section 5.2.5 of PLS prediction models together with 2-norm pre-processing were used to evaluate the results. Again, the RMSE and R of the training set and the $RMSE_{cv}$ and R_{cv} of cross-validation are shown to determine which process provides the minimum RMSE and $RMSE_{cv}$ and R for training set and R_{cv} for cross-validation are shown to determine how well the prediction model can perform. Table 5.4 shows the results for 16 measurement sites for comparison.

Location of body	Abbreviation	RMSE	$RMSE_{cv}$	R	R_{cv}
Left thumb	FL1	3.0261	3.1265	0.4158	0.3493
Left index finger	FL2	2.8433	3.0200	0.5199	0.4302
Left middle finger	FL3	2.8206	2.9868	0.5267	0.4436
Left ring finger	FL4	2.7784	2.9581	0.5467	0.4612
Left little finger	FL5	2.7869	2.9334	0.5427	0.4724
Right thumb	FR1	3.0576	3.1477	0.3948	0.3310
Right index finger	FR2	2.8161	3.0063	0.5341	0.4409
Right middle finger	FR3	2.7940	2.9689	0.5439	0.4604
Right ring finger	FR4	2.8037	2.9693	0.5389	0.4589
Right little finger	FR5	2.8475	3.0011	0.5178	0.4391
Left wrist	LW	3.0497	3.3453	0.4007	0.1872
Right wrist	RW	3.0528	3.3156	0.3977	0.2060
Left forearm (mid-way between the radial side of the first crease and the medial side of the elbow)	LA	3.0164	3.2977	0.4221	0.2297
Right forearm (mid-way between the radial side of the first crease and the medial side of the elbow)	RA	3.0062	3.2860	0.4286	0.2399
Left ear lobe	LE	2.9920	3.3089	0.4375	0.2315
Right ear lobe	RE	2.9736	3.3032	0.4497	0.2391

Table 5.4 Results for sixteen measurement sites

The results illustrated that ten fingers generally provide smaller RMSE and higher R than wrists, forearms and ear lobes. Moreover, independent samples test for equality of means was used to test whether the means of two groups, namely, the ten fingers group and the other body locations group (i.e., wrists, forearms and ear lobes), are statistically different from each other. Table 5.5 shows the results of the independent samples test for equality of means between two groups. The result showed that there was statistically significant difference in RMSE, $RMSE_{cv}$, R and R_{cv} between the two groups, with all the p-values less than 0.05. Therefore, we conclude that ten fingers generally provide smaller RMSE and higher R than wrists, forearms and ear lobes.

	mean different	***standard error different***	***p-value***
RMSE	-0.1577	0.0425	0.002
$RMSE_{cv}$	-0.2976	0.0300	0.000
R	0.0854	0.0238	0.003
R_{cv}	0.2065	0.0211	0.000

Table 5.5 Results of independent samples test between two groups

Besides, the finger clip adopted for measurement could provide a constant force and comparatively firmly fixed position when performing measurements. Among the ten fingers, the thumbs, FL1 and FR1, yielded poorer results than the other fingers did. For the rest of eight fingers, although the values were not varying so much in pairs, the results showed that FL5 provided the highest R_{cv} but a similar result was not found in FR5. This might be due to the different size of last finger for each subject, since the same person has different sizes of left and right little finger itself. Therefore, selecting the location on the finger for testing must be done with caution. It appeared that FL4 provided the second best result among the others, and FR4 can perform similar results as

FL4. Therefore, the results suggested using FL4, the left ring finger, for the measurement site of the body for the following experiment.

5.4 Temperature test

The following experiment was designed to evaluate the temperature effect of the measurement site. The previous section suggested using the left ring finger (FL4) as the measurement site of the body to evaluate the non-invasive blood glucose monitoring. Hence, in this experiment, FL4 was used as the body location to determine whether the temperature effect will cause the results to vary. Table 5.6 shows the distribution of the temperature range of the FL4 and Figure 5.5 plots the histogram for the distribution.

Temperature range °C	Number of counts
16 - 16.9	1
17 - 17.9	2
18 - 18.9	2
19 - 19.9	11
20 - 20.9	24
21 - 21.9	49
22 - 22.9	54
23 - 23.9	78
24 - 24.9	74
25 - 25.9	54
26 - 26.9	46
27 - 27.9	46
28 - 28.9	35
29 - 29.9	47
30 - 30.9	63
31 - 31.9	79
32 - 32.9	86
33 - 33.9	50
34 - 34.9	32
35 - 35.9	10
36 - 36.9	1
Total:	844

Table 5.6 Distribution of the temperature range from 16.8 – 36°C

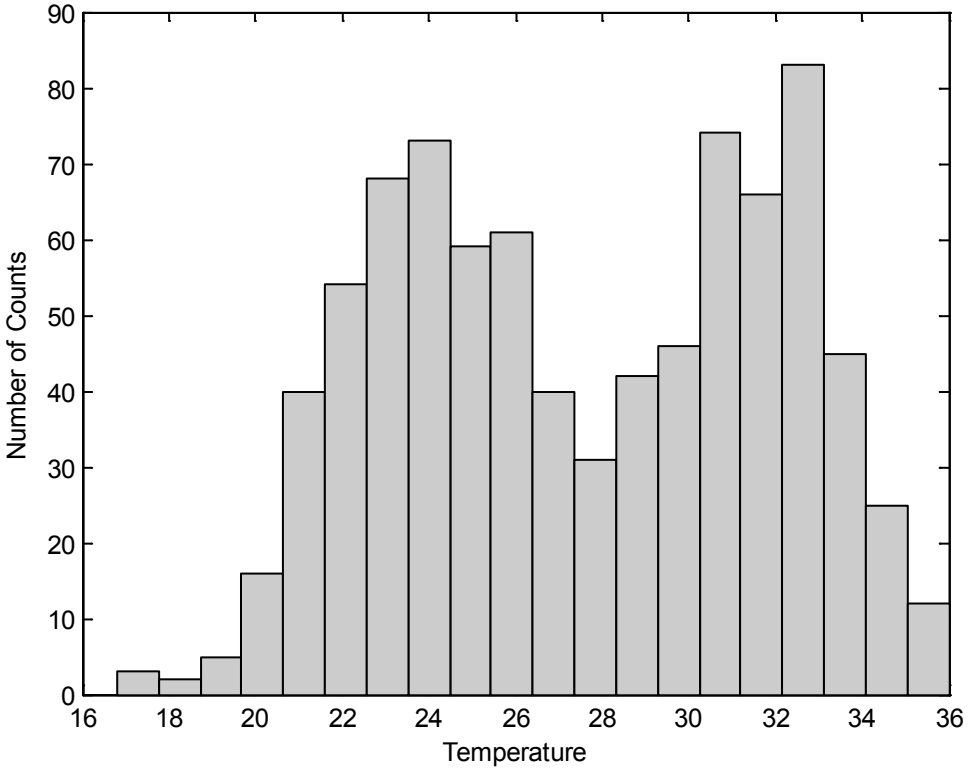


Figure 5.5 Histogram of the temperature range of the FL4

5.4.1 Performance comparison

Recorded temperatures were grouped into ranges to perform the PLS regression method; the range of 19 – 27.9°C, the range of 28 – 34.9°C, the range of 21 – 30.9°C, the range of 19 – 32.9°C, and the full range 16.8 – 36°C were selected for this test. The five temperature ranges were selected based on the availability number of counts (measurements) that we are able to record in the respective ranges. The results of RMSE, $RMSE_{cv}$, R and R_{cv} are shown in Table 5.7.

Temperature range	RMSE	$RMSE_{cv}$	R	R_{cv}
19 - 27.9 °C	2.5754	2.8900	0.5434	0.3769
28 - 34.9 °C	2.8038	3.1910	0.6008	0.4439
21 - 30.9 °C	2.6163	2.8701	0.5671	0.4457
19 - 32.9 °C	2.7640	2.9518	0.5496	0.4605
16.8 - 36 °C	2.7784	2.9581	0.5467	0.4612

Table 5.7 Results of various temperature range of the FL4

As the results show, the temperature effect relative to the various temperature range was not noteworthy and it appeared that the R_{cv} is the highest amongst others groups even if it contains the full dataset for the modelling. Therefore, it is suggested maintaining the full temperature range for the following study.

5.5 Pre-filter analysis

The normalization of spectra is done (in section 5.2) to deal with scaling differences arising from path length effects, scattering effects, source variations, or other instrumental sensitivity effects. Before training the quantitative model, it is possible to manipulate the spectra using various pre-filter methods to try to improve the performance of the quantitative model. Many of the interferences, often described as noise (low frequency or high frequency), are caused by known background signals and artefacts. These interferences are common in many measurements but can be corrected by taking advantage of the relationship between variables in a dataset. In many situations, the adjacent columns in the data matrix are related to each other and contain similar information. Pre-filtering the signals to remove this noise or the effects of signal

variance can be useful for obtaining a more accurate quantitative model (Martens, Hoy, Wise, R., & Brockhoff, 2003; Skibsted, Boelens, Westerhuis, Witte, & Smilde, 2004). The idea is to reduce spectral problems such as noise and baseline drift, or to eliminate unwanted spectral features.

A variety of pre-filter methods exist to remove the interferences (Rinnan, Berg, & Engelsen, 2009). The signal being removed from each sample is assumed to be only interference and is generally not useful for quantitative analysis. This section evaluates various kinds of pre-filter methods for the NIR spectra to enhance performance of the PLS prediction model.

5.5.1 Savitzky-Golay derivation

Derivatives are a common method used to remove unimportant baseline signals from samples by taking the derivative of the measured responses with respect to the wavelength in spectroscopy. The simplest form of derivative is a point difference first derivative in which each variable in a sample is subtracted from its neighbor variable (Eigenvector Research Inc., 2011; Orfanidis, 1996). This subtraction removes the signal that is the same between the two variables and retains the informative signal that is different. When performed on an entire sample, a first derivative effectively removes any offset from the sample and a second derivative can be calculated by repeating the same subtraction process to further accentuate the informative signals. The Savitzky-Golay (SG) derivation is commonly used for numerical derivation because it includes a

smoothing step when it takes the derivative, in which it can improve the utility of data (Savitzky & Golay, 1964). The algorithm basically fits individual polynomials to windows around each point in the spectra. After that, these polynomials are used to smooth the data. This method requires selection of the window size (filter width) and the order of the polynomial. In general, the larger the window and the lower the polynomial order, the more smoothing that occurs for the data. Since this operation is applied to all variables in the spectra sequentially, the number of the filter width and the degree of the fitted polynomial are both decisions that need to be made.

5.5.2 Detrending

In traditional time series analysis, a time series is decomposed into trend or periodic components. Trend is a gradual change of the time series over the whole period of time and it is sometimes loosely defined as a long-term change in the mean. Detrending is the statistical or mathematical operation of removing trend from the series. Detrending can be used to remove a constant, curved offset if present or linear trends from regular data signal (Barnes, Dhanoa, & Lister, 1989). This method fits a polynomial of a given order to the entire sample and simply subtracts this polynomial. The method is very simple to use because it fits the polynomial to all points, baseline and signal, while it tends to work only when the largest source of signal in each sample is background interference (Eigenvector Research Inc., 2011). It tends to remove variations that are useful to modelling and may even create nonlinear responses from other linear relationships; also an individual polynomial fitted to each spectrum may increase the amount of interfering

variance in a data set. Therefore, the use of a detrending method is recommended only when the overall signal is dominated by backgrounds that generally have the same shape and are not greatly influenced by other processes.

5.5.3 Generalized Least Squares (GLS) Weighting

In the case of a calibration transfer problem, especially in this study, similar samples would be measured on the same instrument for different people at different points in time (Leung, Xiong, Lau, & So, 1999). The goal of a generalized least squares (GLS) weighting pre-filter is to down-weight the differences between similar samples and thus make them appear more similar (Martens, Hoy, Wise, R., & Brockhoff, 2003). A GLS weighting pre-filter can be used to remove variance from the spectral responses that are mostly orthogonal to the concentration information. It uses samples with similar concentration values to identify the sources of variance to down weight. The GLS weighting pre-filter method involves the calculation of a covariance matrix from the differences between similar samples. Given two sample matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$, the difference equation for them can be calculated as

$$\mathbf{X}_d = \mathbf{X}_2 - \mathbf{X}_1 \tag{5.8}$$

The covariance matrix, $\mathbf{C}$, of $\mathbf{X}_d$, then needs to be calculated, which is

$$\mathbf{C} = \mathbf{X}_d^{\mathrm{T}}\mathbf{X}_d \tag{5.9}$$

By using the SVD of the matrix shown in (3.19) into (5.9), it becomes

$$\mathbf{C} = \mathbf{V}\Sigma^2\mathbf{V}^{\mathrm{T}} \tag{5.10}$$

Since the collinearity problem may exist when $\mathbf{X}_d^{\mathrm{T}}\mathbf{X}_d$ is singular, the ridge estimator is used for the sake of stability to calculate the singular values and it becomes

$$\Sigma_{\mathrm{ridge}} = \sqrt{\Sigma^2 + \alpha\mathbf{I}} \tag{5.11}$$

where $\mathbf{I}$ is a diagonal matrix of ones and α is the weighting parameter that defines how strong effect of GLS weighting should be.

Finally, the inverse of the weighted eigenvalues in (5.11) is used to calculate the filtering matrix as

$$\mathbf{X}_{\mathrm{GLS}} = \mathbf{V}\Sigma_{\mathrm{ridge}}^{-1}\mathbf{V}^{\mathrm{T}} \tag{5.12}$$

In general, α is normally set equal to 0.001, while the larger the α value, the decrease in the effect of the GLS weighting. The choice of α value depends on the scale of the original values. If the interferences are similar to the variance necessary to the analytical measurement, then α value will need to be larger in order to keep from removing analytically-useful variance. However, a larger α value will decrease the extent to which interferences are down weighted. Therefore, an α value between 1 and 0.0001 is the range often used in practice.

An alternative multivariate pre-filter called External Parameter Orthogonalization (EPO) uses the same process as GLS weighting pre-filter except that only a certain number of eigenvectors calculated in (5.10) are kept, and the ridge estimator in (5.11) is a diagonal vector of ones. The principle of EPO pre-filter method is that the spectral responses should be orthogonal to the concentration output in which the directions are completely removed, whereas GLS weighting pre-filter is trying to shrink the directions. More

details of the EPO pre-filter method is described in Roger's research team (Roger, Chauchard, & Bellon-Maurel, 2003).

5.5.4 Orthogonal Signal Correction (OSC)

Orthogonal Signal Correction (OSC) was introduced to remove systematic variation from the spectral responses that is unrelated or orthogonal to concentration information (Wold, Antti, Lindgren, & Ohman, 1998). Such variance is identified as some number of components of the spectral responses that have been made orthogonal to the concentration. Therefore, one can be certain that important information regarding the analyte is retained. The OSC model can be expressed by

$$\mathbf{X} = \mathbf{t}_{osc}\mathbf{p}_{osc}^{T} + \mathbf{E} \tag{5.13}$$

where $\mathbf{t}_{osc} = \mathbf{X}\mathbf{w}_{osc}$ and $\|\mathbf{w}_{osc}\| = 1$ and the $\mathbf{t}_{osc}$, $\mathbf{p}_{osc}$ and $\mathbf{w}_{osc}$ represent the OSC component.

The OSC model is similar to the standard PLS regression model because it has two sets of loading vectors. The main difference is that the score vectors are orthogonal to the output vector.

The working principle of the OSC pre-filter method proposed by Sjoblom's team starts by identifying the first principal component (PC) of the spectral responses and then the loading is rotated to make the scores become orthogonal to the concentration (Sjoblom, Sevensson, Josefson, Kullberg, & Wold, 1998). This loading represents a feature that is not influenced by changes in the property of interest described in the concentration.

Once the rotation is completed, a model is created that can predict these orthogonal scores from the spectral responses. The number of components in the model is adjusted to achieve a given level of captured variance for the orthogonal scores. Finally, the weights, loadings, and predicted scores are used to remove the given orthogonal component.

There are three settings for the OSC pre-filter method, which include number of components, number of iterations, and tolerance level. The number of components is used to set how many times the entire process will be performed. The number of iterations is used to set how many cycles will be used to rotate the initial PC loading to be as orthogonal to the output concentration as possible. The tolerance level is used to set the percent variance by the model.

5.5.5 Performance comparison

Various pre-filter methods mentioned previously were compared; each method had its own parameter settings and the results are listed in Table 5.8. The first SG derivative and second SG derivative methods were implemented. The linear, second order, third order, and the fourth order of detrending methods are listed. The GLS weighting pre-filter method with various α values were evaluated, and the EPO pre-filter method with different PCs was investigated. Moreover, the OSC pre-filter with various PCs and the tolerance values is listed as well. The results with and without normalization are shown together for comparison. This experiment considered performing the pre-filter with and

without normalization for their own parameter settings and that result was shown in section 5.2.5. Comparing the results with and without normalization showed that the normalization process can improve the RMSE and R for all pre-filter methods, particularly for the GLS weighting pre-filter method. Therefore the results further revealed the benefit of implementing the normalization process and the pre-filter process thereafter so as to provide better performance for analysis.

Methods		**$RMSE_c$**	**$RMSE_{cv}$**	**R_c**	**R_{cv}**
With 2-norm applied					
1st SG	(order 2, window 15pt)	2.7735	3.0523	0.5282	0.4141
Derivative	(order 6, window 7pt)	2.7322	3.0263	0.5675	0.4343
2nd SG	(order 2, window 15pt)	2.7449	3.0597	0.5620	0.4171
Derivative	(order 3, window 9pt)	2.8432	3.1043	0.5166	0.3887
	(order 5, window 9pt)	2.8724	3.1160	0.5008	0.3720
Detrending	(linear)	2.8304	3.0607	0.5223	0.4093
	(2nd order)	2.8019	3.0654	0.5358	0.4083
	(3rd order)	2.7988	3.0652	0.5373	0.4091
	(4th order)	2.7949	3.0811	0.5392	0.4015
GLS weighting	(α=0.001)	2.4477	3.0359	0.6752	0.4668
	(α=0.002)	2.4828	2.9892	0.6634	0.4784
	(α=0.005)	2.5803	2.9539	0.6287	0.4807
	(α=0.01)	2.6637	2.9487	0.5963	0.4749
	(α=0.02)	2.7476	2.9620	0.5606	0.4617
EPO pre-filter	(PC=1)	2.7785	2.9598	0.5466	0.4605
	(PC=5)	2.7083	2.9723	0.5784	0.4606
	(PC=10)	2.6363	2.9855	0.6073	0.4630
OSC pre-filter	(PC=1, tol=99.9%)	2.5770	3.1551	0.6300	0.4007
	(PC=8, tol=99.9%)	3.3172	3.3747	0.1798	0.0202
	(PC=10, tol=95%)	3.1029	3.3405	0.3543	0.1715

(a) With 2-norm applied

Methods		$RMSE_c$	$RMSE_{cv}$	R_c	R_{cv}
Without 2-norm applied					
1st SG	(order 2, window 15pt)	2.7771	3.0266	0.5473	0.4280
Derivative	(order 6, window 7pt)	2.7300	3.0259	0.5684	0.4341
2nd SG	(order 2, window 15pt)	2.7442	3.0627	0.5622	0.4158
Derivative	(order 3, window 9pt)	2.8411	3.1068	0.5169	0.3834
	(order 5, window 9pt)	2.8735	3.1212	0.5006	0.3728
Detrending	(linear)	2.8246	3.0614	0.5249	0.4088
	(2nd order)	2.7956	3.0650	0.5387	0.4084
	(3rd order)	2.7923	3.0639	0.5402	0.4098
	(4th order)	2.7887	3.0811	0.5419	0.4012
GLS weighting	(α=0.001)	7.7865	7.8822	0.1635	0.0890
	(α=0.002)	6.7393	7.0674	0.1374	0.0952
	(α=0.005)	3.9973	4.2710	0.0403	0.0185
	(α=0.01)	2.6580	3.1751	0.6035	0.4160
	(α=0.02)	2.4598	3.0300	0.6711	0.4659
EPO pre-filter	(PC=1)	20.7718	18.3372	0.3112	0.1262
	(PC=5)	15.7205	15.3530	0.2759	0.1265
	(PC=10)	23.2130	19.0799	0.2813	0.0672
OSC pre-filter	(PC=1, tol=99.9%)	2.7625	3.0239	0.5539	0.4304
	(PC=8, tol=99.9%)	3.1174	3.3713	0.4151	0.1575
	(PC=10, tol=95%)	3.0105	3.3298	0.4205	0.2116

(b) Without 2-norm applied

Table 5.8 Results for various pre-filter methods with and without 2-norm applied

According to the simulation results, GLS weighting pre-filter performed the best amongst all methods. It yielded the smallest $RMSE_c$ and the highest R_c when α=0.001, while it provided the smallest $RMSE_{cv}$ and highest R_{cv} when α=0.005. Considering the difference between the two α values, the author suggests using the GLS weighting pre-filter with α=0.001 as it provides the best $RMSE_c$ and R_c and can also perform $RMSE_{cv}$ and R_{cv} reasonably well, in which the difference between the results of α=0.001 and α=0.005 were 0.082 in $RMSE_{cv}$ and 0.0139 in R_{cv}.

5.6 Features selection

The high dimensionality of spectral data increases the difficulty of using quantitative regression models. Reducing this high dimensionality helps reduce variable numbers, potentially improves the accuracy by removing irrelevant spectral information, and reduces computation time and cost. In this section, two heuristic optimization algorithms, sequential floating selection (SFS) and genetic algorithm (GA), are applied to variable selection for the spectral data. The study provides a robust and efficient optimization approach to reduce high data dimensionality of the spectral data, also improves the performance of the regression model, and potentially reduces the overall cost for future applications.

5.6.1 Sequential floating selection (SFS)

Sequential floating selection (SFS) is a well-known suboptimal search algorithm that is very efficient and effective even for problems of high dimensionality involving non-monotonic feature selection (Pudil, Novovicova, & Kittler, 1994). The feature selection approach is to select a subset of n features from a given set of N measurements, where n<N, without significantly degrading the performance of the whole system. Sequential floating forward selection (SFFS) and sequential floating backward selection (SFBS) are two similar but reverse process approaches to search the optimal settings by sequentially adding (forward) or removing (backward) variables. SFFS starts from an empty space and at each step the variable that provides the best improvement of the objective

function is added to the feature space. This process is repeated until all variables are selected. SFBS works in reverse, beginning with the whole feature space and at each step the sensor that provides the least reduction of the objective function is removed from the selection. This process is repeated until all variables are removed from the array. These approaches are effective for reducing variables while they explore only a fraction of the whole search space since the order of removing or adding variables influences the search results. However, SFFS/SFBS can result in nested feature subsets from previous steps that do not allow corrections of decisions in later steps, leading to a performance that is often far from optimal. In view of this, an improved version of SFF has been developed by combining the (SFFS) and (SFBS) to avoid the nesting effect (Pudil, Ferri, Novovicova, & Kittler, 1994).

The idea of this approach simply considers the conditional inclusion and exclusion of features controlled by the value of the criterion itself. SFFS is first applied for the forward steps so long as the resulting subsets are better than the ones in the previous stage. If the performance cannot be improved at a certain point, SFBS will be applied in the backward steps. Further forward steps will then be applied in attempt to improve the performance. The process repeats iteratively in this way until reaching the number of features required. This approach provides a reverse tracking mechanism that achieves optimal solutions by dynamically adjusting the trade-off between forward and backward steps to keep away from the nesting effect. The combined SFS algorithm is described in Figure 5.6.

SFS algorithm

Input:
$\mathrm{X} = \{x_1, \cdots, x_\mathrm{N}\}$
Output:
$\mathrm{Y} = \{y_1, \cdots, y_n\}$ where $y_\mathrm{i} \in \mathrm{X}, n = 0, 1, \cdots, \mathrm{N}$
Initialization:
$\mathrm{Y} = 0; \mathrm{i} = 0$
Termination:
i = *n*

Forward Step:
$y_\mathrm{F} = \max_{y \in \mathrm{X}-\mathrm{Y_i}} \mathrm{J}(\mathrm{Y_i} + y_\mathrm{F})$
$\mathrm{Y_{i+1}} = \mathrm{Y_i} + y_\mathrm{F}$
$\mathrm{i} = \mathrm{i} + 1$

Backward Step:
$y_\mathrm{B} = \max_{y \in \mathrm{Y_i}} \mathrm{J}(\mathrm{Y_i} - y_\mathrm{B})$
if $\mathrm{J}(\mathrm{Y_i} - y_\mathrm{B}) > J(\mathrm{Y_{i-1}})$
$\mathrm{Y_{i-1}} = \mathrm{Y_i} - y_\mathrm{B}$
$\mathrm{i} = \mathrm{i} - 1$
go to **Backward Step**
else
go to **Forward Step**

Figure 5.6 The SFS algorithm

5.6.2 Genetic algorithm (GA)

In the engineering field, genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution (Goldberg, 1989; Falkenauer, 1998). This heuristic is routinely used to generate useful solutions to optimization and search problems. GA is a randomized search algorithm that belongs to the class of evolutionary algorithms. It is inspired by the process of natural selection and performs a global random search on a population of solutions. It allows a population composed of many chromosomes to

evolve to reach a point where the fitness is maximized or the cost function is minimized. Once the genetic representation and the fitness function are defined, a GA proceeds to initialize a population of solutions and then to improve it through repetitive application of the selection, crossover, inversion, and mutation operators (Sivanandam & Deepa, 2008). Selection operator gives the chromosome with the lowest cost as measured by a fitness function in order to converge towards the best solutions. Crossover is a genetic operator used to generate a second generation population of solutions from those selected chromosomes. Inversion operator has the opportunity to place steps in consecutive order or any other order in favour of survival or efficiency. Mutation operator is to maintain genetic diversity from a certain percentage of the genes that compose population members.

The operation of the GA is that, given a predictor datum, X, and output value, Y, one can choose a random subset of variables from X through the use of fitness function; for example, multivariate regression methods in this study, then to determine the cost function; for example, RMSE or R, obtained when using the subset of variables in a fitness function. GA uses this approach iteratively to locate the variable subset that performs the lowest cost function, RMSE in this case. The first step of the GA is to generate a large number of random selections of the variables and calculate the RMSE for each of the given subsets by using PLS regression method. Each subset of variables is called an individual and the variables that are used by that individual are indicated by '0/1' status, in which 0 is equivalent to the variable that is not selected and 1 is equivalent to the variable that is selected. The pool of all tested individuals is the population. The RMSE values, described as the fitness of the individual, indicate how

predictive each individual's selection of variables is for the output value Y. The individuals with fitness greater than the median fitness (equivalent to providing a higher RMSE) are discarded in the selection operator.

The remaining individuals, which used variables that provide a lower RMSE, are considered as a better fit to the data. At this stage, the population has been reduced and the GA crosses over the retained individuals (parents) to form a new offspring (children) to replace the discarded individuals. If no crossover is performed, offspring is an exact copy of parents. After adding the new offspring to the population, all the individuals' genes are given a chance for random mutation. This allows for a finite chance of adding or removing the use of variables that might be over-represented or under-represented in the population. Finally, after all the individuals have been paired and bred, the population returns to the original size and the process can continue again at the fitness evaluation step (Ghasemi & Saaidpour, 2007). The outline of the basic genetic algorithm can be summarized as follows (Majumder, Roy, & Mazumdar, 2010):

1. [Start] Generate random population of chromosomes (variable subsets)
2. [Fitness] Evaluate the fitness function of each chromosome in the population
3. [New population] Create a new population by repeating the following steps until the new population is complete
    i. [Selection] Select two parent chromosomes from a population according to their fitness
    ii. [Crossover] With a crossover probability cross over the parents to form a new offspring

   iii. [Mutation] With a mutation probability mutate new offspring at each position in chromosome
   iv. [Accepting] Place new offspring in a new population
4. [Replace] Use new generated population for a further run of algorithm
5. [Test] If the end condition is satisfied, the GA process is stopped and it returns the best solution in the current population
6. [Loop] Repeat steps 2-5 until ending criteria are met

The ending criteria of the GA can be either by a number of iterations or by a percentage of the individuals in the population that are using identical variable subsets. With the GA, individuals containing less noisy information will tend to be selected and therefore the variables used by those individuals will become more and more representative in the overall gene population. Conversely, individuals containing noisy information will become less and less represented. Eventually, many of the individuals will contain the same genes if the rate of mutation is sufficient.

However, a practical consideration in GA variable selection is over fitting. It is possible that the variables selected may be particularly good for generating a model for the given data yet may not be useful for future data, especially when analyzing data with small samples size. Therefore, it is generally recommended to follow the guidelines as listed below:

1. Keep the ending criteria sensitive. The more iterations that occur, the more feedback the GA will have from the cost function, and therefore the more likely the over fitting.

2. Use random cost function and multiple iterations if practical.
3. Repeat the GA multiple times and observe the general trends if possible.
4. Keep the maximum number of latent variables low.

5.6.3 Performance comparison

The same setting in section 5.2.5 and the pre-filter result as found in section 5.5.5 were applied to perform these features selection methods. Considering the cross validation results of SFS and GA feature selections in Table 5.9, the study found that the GA method that selected 120 features (wavelength) provided the smallest $RMSE_{cv}$ and the highest R_{cv} amongst other results. In addition, the runtime of SFS and GA were also noted. The time elapsed to acquire the final dataset by SFS was much longer than that by GA, where GA normally took around 30 minutes up to 2 hours while SFS took more than 4 days to complete. The simulation was conducted with a PC containing an Intel Xeon 5150 @2.66GHz with 2G RAM, running Matlab version 2008a on the Windows XP SP3 operating system.

The analysis results must be interpreted with caution, especially for the GA method; it usually considered the entire set of selected features (120 wavelengths) to be used as a whole in the experiment because a feature may only be helpful to prediction when used in conjunction with other features included in an individual feature subset.

Methods	Variable used	$RMSE_c$	$RMSE_{cv}$	R_c	R_{cv}
None	760	2.4477	3.0359	0.6752	0.4668
SFS	96	2.6198	2.9068	0.6137	0.4979
SFS	709	2.4488	3.0189	0.6748	0.4722
GA	273	2.5802	2.8553	0.6266	0.5189
GA	178	2.4977	2.8900	0.6564	0.5126
GA	120	2.6453	2.7748	0.6014	0.5483

Table 5.9 Results of various methods for features (wavelength) selection

The selected wavelengths are shown in Figure 5.7 and the wavelength ranges that the GA method selected were 1,128 – 1,133nm, 1,158 – 1,163nm, 1,230 – 1,235nm, 1,344 – 1,367nm, 1,398 – 1,403nm, 1,422 – 1,427nm, 1,446 – 1,457nm, 1,476 – 1,481nm, 1,602 – 1,607nm, 1,698 – 1,703nm, 1,722 – 1,727nm, 1,740 – 1,745nm, 1,806 – 1,817nm, and 1,836 – 1,847nm. This set of result was used in the following experiments.

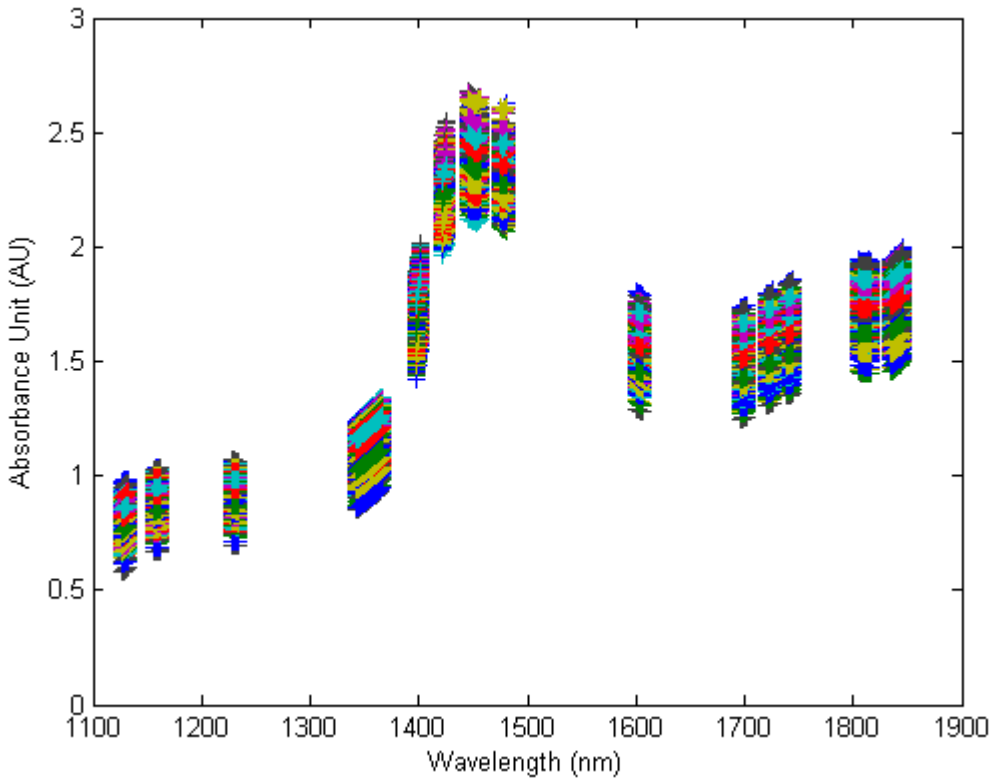


Figure 5.7 Selected features (wavelengths) based on GA (120 features selected)

5.7 Summary

This chapter determined the use of a normalization process together with GLS weighting pre-filter process to spectral data for multivariate analysis. It showed that the left ring finger, FL4, with the assistive finger clip was the best position for measurement. In addition, results showed that the effect of temperature was not significant, yet temperature provides a better R_{cv} for the range between 16.8 - 36°C. Finally, 120 features (wavelength) were selected according to the GA search. Further to these important findings in this chapter, an improved algorithm was presented for multivariate analysis to further enhance the performance.

CHAPTER 6

MONTE CARLO METHOD MULTIVARIATE ANALYSIS

As mentioned, NIR spectroscopy was developed to characterize the constituent components of a given material. This spectroscopic method has become a widely-used non-destructive measurement technique for agrochemical, pharmaceutical and medical applications (Burns & Ciurczak, 2008; Ciurczak & Drennen, 2002; Ozaki, McClure, & Christy, 2007). However, since the NIR spectrum is composed of overlapping peaks of the overtones (e.g. first, second or third overtones, and their combinations) due to various fundamental molecular vibrations (Siesler, Ozaki, Kawata, & Heise, 2002), it is difficult to identify the exact features of specific chemical components from the absorption spectrum. Hence, multivariate calibration methods, which are often used in quantitative spectral analysis, play a critical role in NIR spectral analysis where adequate calibration spectra are required for the analysis.

6.1 Classification analysis

Common multivariate calibration methods for classification include linear discriminant analysis (LDA) (Huberty & Olejnik, 2006; Kemsley, 1998; Klecka, 1980; McLachlan, 1992), principal component analysis (PCA) (Ferre, 1995; Jolliffe, 2002; Kettaneh, Berglund, & Wold, 2005; Wold, Esbensen, & Geladi, 1987), partial least squares discriminant analysis (PLS-DA) (Barker & Rayens, 2003; Chiang, Russell, & Braatz, 2000; Kemsley, 1996; Perez, Ferre, & Boque, 2009; Perez-Enciso & Tenenhaus, 2003; Sirven, Salle, Mauchien, Lacour, Maurice, & Manhes, 2007; Zhang, Ortiz, Xie, Davisson, & Ben-Amotz, 2005) and support vector machine analysis (SVM) (Cristianini & Shawe-Tayor, 2002; Ju, Shan, Yan, & Cheng, 2009; Lin & Yeh, 2009; Scholkopf, et al., 1997; Suykens & Vandewalle, 1999; Vapnik, 1998). All of these methods are widely used to build prediction models for classification in NIR spectral analysis. Among these multivariate calibration methods, PLS-DA is becoming more and more popular because of its ability to project the predicted and observable variables into a new space in order to maximize the covariance between the response and independent variables (Geladi & Kowalski, 1986). However, the performance of PLS-DA model can be severely deteriorated due to the presence of outliers (Chen, Shao, Hu, & Su, 2004; Lleti, Melendez, Ortiz, Sarabia, & Sanchez, 2005; Mittermayr, Tan, & Brwon, 2001). This is particularly unfavourable for NIR spectral analysis where outliers resulting from measurement errors or noise in sensors are quite common, especially for a large data matrix. Hence, a number of techniques have been developed to reduce the effect of outliers on PLS-DA (Kasemsumran, Du, Maruo, & Ozaki, 2006; Pierna, Jin,

Daszykowski, Wahl, & Massart, 2003). All of these methods attempt to reduce the noise and increase the effectiveness of data extraction for model construction.

Another issue of the PLS-DA method is that only one single quantitative model is constructed most of the time to predict the relationship between the NIR spectra and the concentrations of the samples. Obviously, this approach can easily misidentify or fail to identify important characteristic features contained in the NIR spectra, under or over estimating the number of NIR spectral features required for the analysis. The results obtained by a single prediction model are thus unstable or correlated to spurious spectral variance, particularly when the training set for PLS-DA is relatively small.

This section proposes a new algorithm to enhance both the stability and performance of the conventional PLS-DA prediction model. The method employs the Monte Carlo method (Chen, Cai, & Shao, 2008; Doucet, Freitas, & Gordon, 2001; Dunn & Shultis, 2012; Robert & Casella, 2004) to obtain a prediction model that is derived by averaging the results obtained from a number of PLS-DA models. The Monte Carlo method is a well-known computational algorithm that is often used for simulating the behaviour of a system when it is not feasible to use deterministic algorithms. The Monte Carlo method is used in this work to randomly select different subsets of training data from the whole training dataset to create a large number of PLS-DA models. As each of these models is created only with a subset of training data, the number of outliers within each subset is likely to be smaller than that in the whole training dataset, and the adverse effect on the individual PLS-DA models is therefore reduced. A prediction model is then obtained based on part of the models that exhibit superior classification performance. Compared

to conventional PLS-DA where only one single model is involved, the proposed approach can identify more characteristic features that are relevant to establish the relationship between the predicted and observable variables. It is able to decrease the risk of over fitting the prediction model and stabilize the performance, particularly for NIR spectra analysis involving a large data matrix.

6.1.1 PLS-DA with the Monte Carlo Method

PLS regression is a popular modelling method to establish the relationship between one or more dependent variables **Y** and a group of descriptors **X,** where **X** and **Y** are in matrix notation. It is a latent variable (LV) approach where the covariance structures between **X** and **Y** are modelled by making inferences from a mathematical model with directly measurable variables (Tabachnick & Fidell, 2001). The number of LVs of PLS-DA is an essential parameter to determine the dimensionality of data in the modelling. For the linear regression model, the prediction **Y** is computed by the equation

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{6.1}$$

where **X** is a n x p matrix containing p dependent variables of n samples, **B** is a p x 1 vector of regression coefficients obtained from PLS analysis, and **E** is the model offset. PLS-DA is a variant of PLS when the dependent variable is binary.

The idea of the proposed MC-PLS-DA is to create a large number of PLS-DA prediction models using different training subsets that are selected randomly from the whole

training data set by the Monte Carlo method. The stability of the corresponding coefficients is calculated by using the regression coefficients of these models. The MC-PLS-DA prediction model is then obtained by averaging the PLS-DA models so that the outliers can be removed. Figure 6.1 shows the MC-PLS-DA algorithm. In the proposed method, the raw data are randomly divided into two parts – the training set $\mathbf{X_t}$ and the prediction set $\mathbf{X_p}$. The prediction set $\mathbf{X_p}$ is only used to evaluate the robustness of the prediction model and not for training. Different subsets are randomly selected from the training set $\mathbf{X_t}$ to construct a large number of PLS-DA models. The size of each subset is 70% of that of the whole training set $\mathbf{X_t}$. These PLS-DA models are validated by the prediction set $\mathbf{X_p}$. The top 5% of the models are selected to create the final MC-PLS-DA model by averaging the prediction of these models.

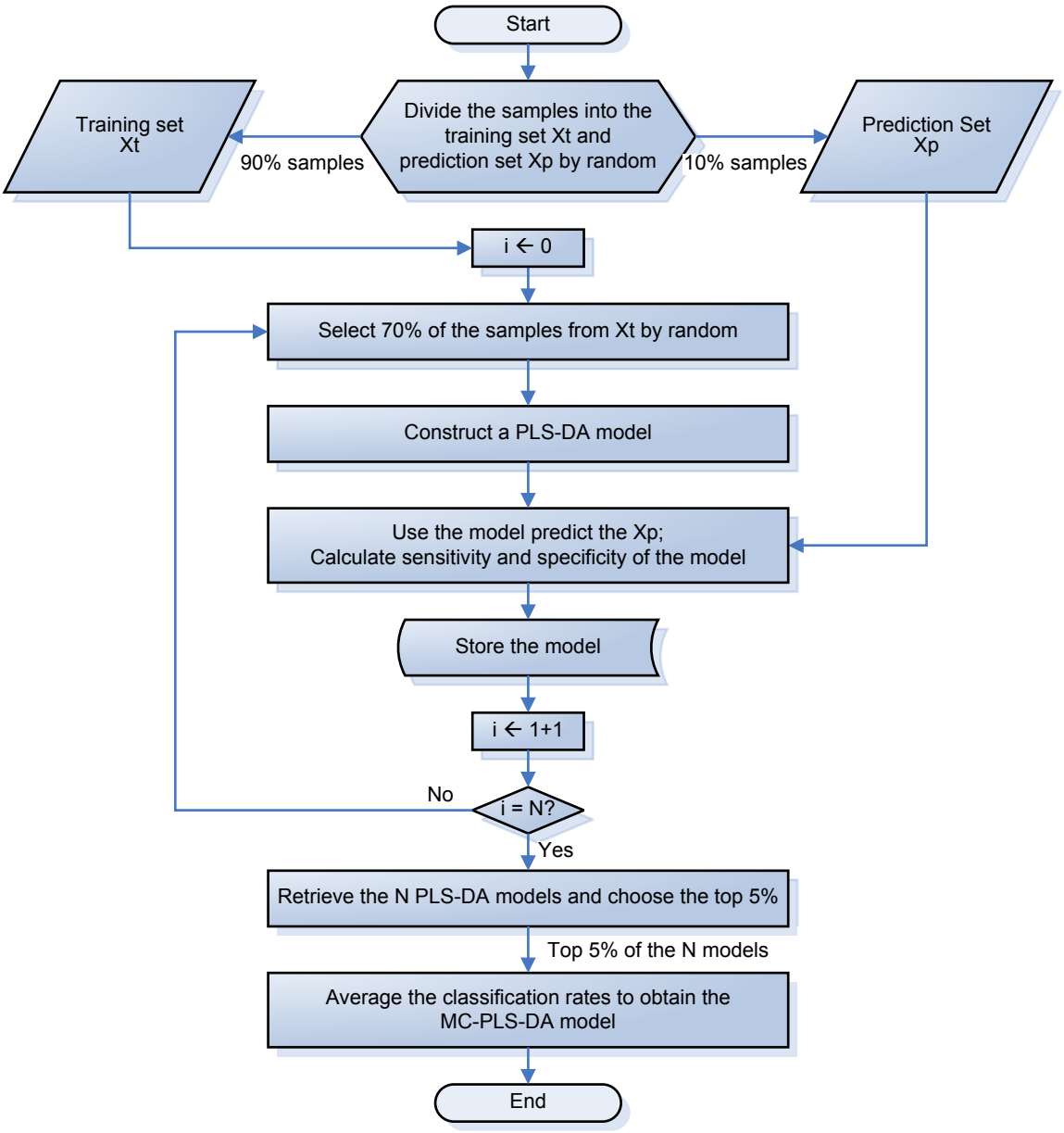


Figure 6.1 The MC-PLS-DA algorithm

In this study, the top performing model is defined as the one with the highest classification rate. The averaged PLS-DA prediction model is obtained by

$$\bar{\mathbf{B}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_i \qquad (6.2)$$

where n is the number of top 5% of the PLS-DA models, and $\mathbf{B}_i$ is the prediction model of the MC-PLS-DA model.

In the proposed method, it is necessary to determine the optimal number of PLS-DA models required to construct a stable and precise MC-PLS-DA model. The optimal number depends on the percentage of top performing PLD-SA models. On one hand, when the number of PLS-DA models used is too small, the resulting MC-PLS-DA model may overemphasize or underestimate the importance of certain features of the data, especially for a large data set like the NIR spectral data. The stability and precision are also adversely affected. On the other hand, using a large number of PLS-DA models is also undesirable because it is computationally intensive, and irrelevant features may be included to deteriorate the prediction result. While the percentage could be adjusted experimentally, it is pre-defined as being the top 5% of the PLS-DA modes in order to simplify and facilitate the determination of the optimal model number. The number was arrived at by drawing an analogy with the idea of confidence interval in probability and statistics, where a p-value of 5% (i.e. 5% of the total samples) is conventionally used to determine statistical significance.

6.1.2 Performance of MC-PLS-DA

In this study, the blood glucose level (BGL) was divided into two classes, with BGL falling between 4mmol/L and 7mmol/L considered as *normal BGL*, and BGL greater

than 7mmol/L defined as *high BGL*; that is, suffering from DM. The prediction **y** is thus classified by

$$\mathbf{y} = \begin{cases} 0, & 4mmol/L \leq BGL \leq 7mmol/L \\ 1, & BGL > 7mmol/L \end{cases} \tag{6.3}$$

The performance of the MC-PLS-DA model can be evaluated by the sensitivity, specificity, and overall accuracy of the model. Sensitivity is a measure of the ability of the classifier to identify normal BGL. Specificity is a measure of the ability to identify high BGL. Accuracy is a measure of the ability to identify the correct BGL. They are defined by the following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6.4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{6.5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6.6}$$

where TP, TN, FP and FN denote the number of *true positives*, *true negatives*, *false positives* and *false negatives* respectively. Here, TP and TN refer to the correct classification of normal and high BGL respectively, whereas FP and FN refer to incorrect classification of normal and high BGL respectively.

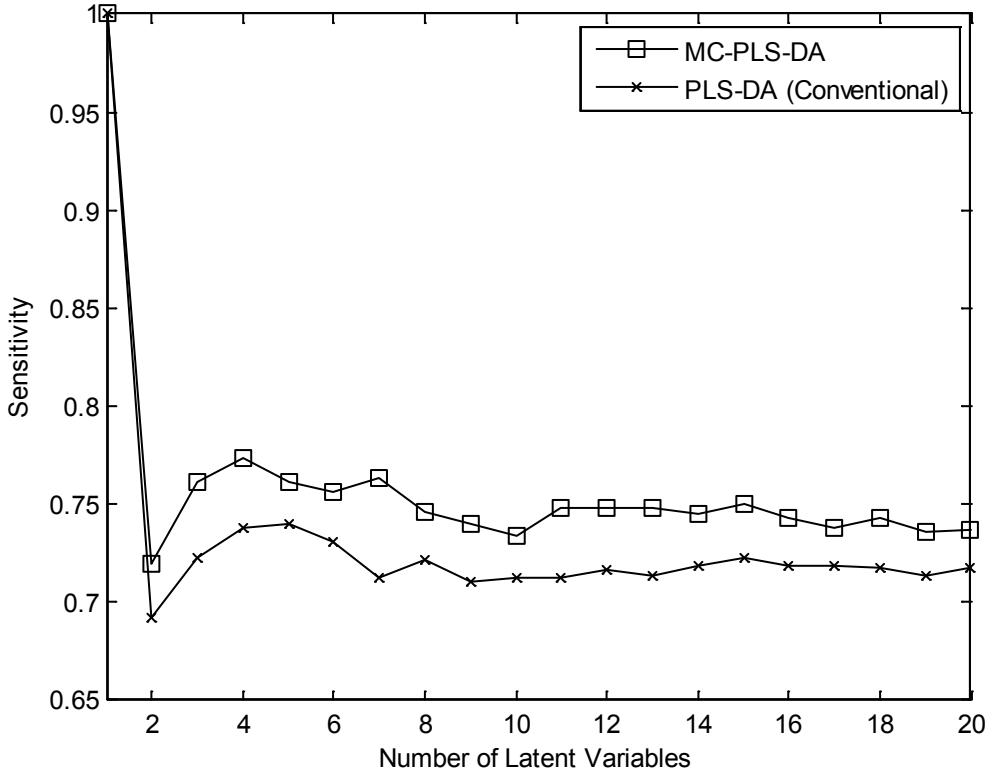
Since there were four samples in the left ring finger (FL4) where the blood glucose level of collected samples was below 4mmol/L, hence there were 840 samples of FL4 that were randomly divided into the training data set and the prediction data set. The former contained 756 samples (90% of the total number of samples) while the latter contained 84 samples. 530 samples were then randomly taken from the training data set to create a

PLS-DA model, where the correlation between the predicted and reference values was maximized. All the samples in the prediction data set were used to fit the training model and evaluate the classification rate. By the same token, another 530 samples were drawn the same subset of training data to create a new PLS-DA model. The procedure was the carried out repeatedly until 1,000 PLS-DA models were built. The top 5% (50) of the models, in terms of classification rate, were selected to create the MC-PLS-DA model with the same training and prediction data sets. The entire algorithm shown in Figure 6.1 was then executed repeatedly for a total of 20 runs. At each run, the 840 samples were randomly divided to produce different training and prediction data sets, so that a total of 20 MC-PLS-DA models were obtained. The sensitivity, specificity, and the accuracy of these 20 MC-PLS-DA models were averaged to evaluate the performance of the algorithm.

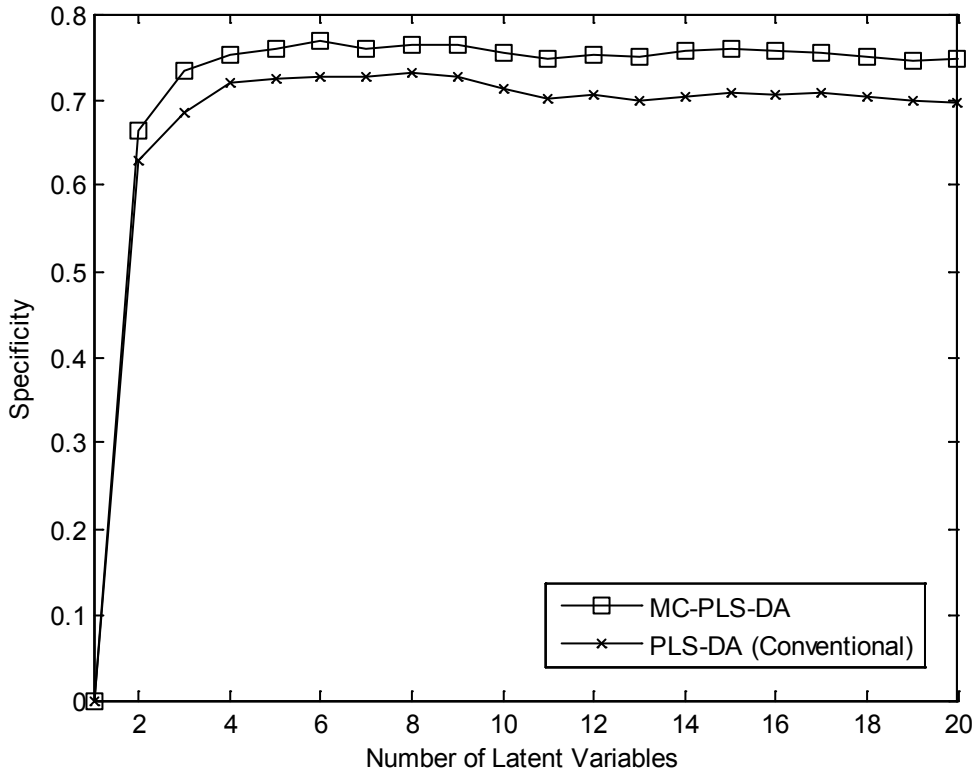
Optimal number of LVs for PLS-DA model

The number of the LVs of PLS-DA model was determined first because it is an important parameter in the modelling. The performance of the MC-PLS-DA and conventional PLS-DA model are compared in Figure 6.2, in terms of the variation of sensitivity in (6.4), specificity in (6.5), and accuracy in (6.6) with the number of LVs. The performance of the MC-PLS-DA model was found to be better than the conventional PLS-DA model for all cases. The sensitivity, specificity, and accuracy of the MC-PLS-DA model appeared to remain at the same level when the number of LVs was increased, and remained stable and constant after six LVs. The results indicated that

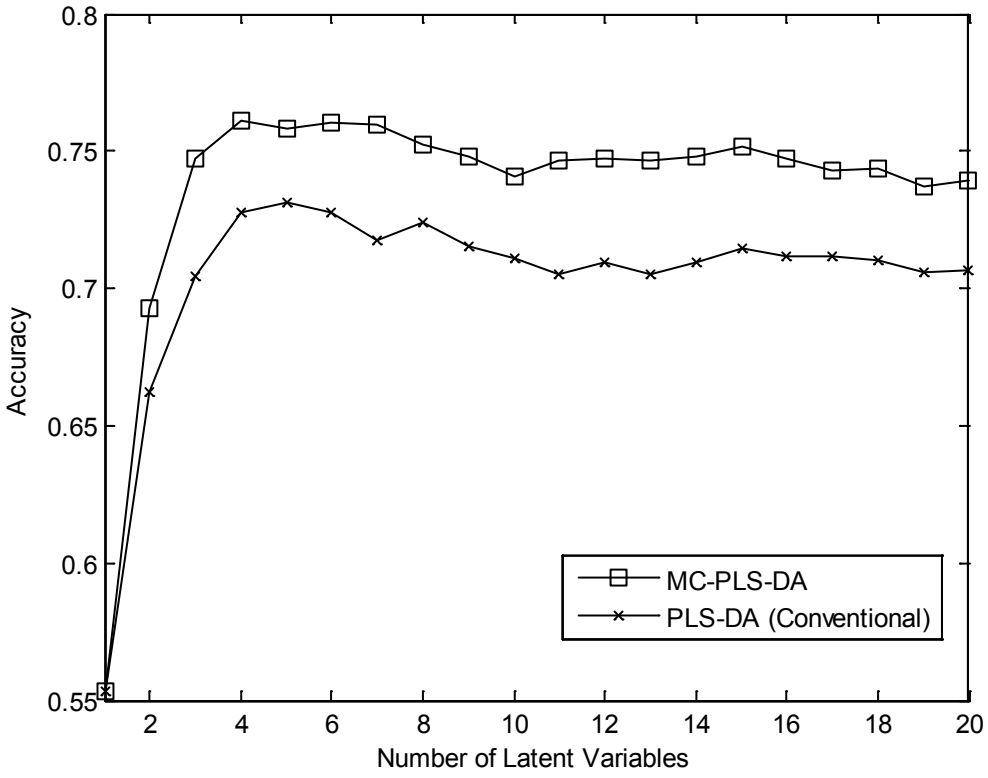
the application of the Monte Carlo method can stabilize and improve the performance of conventional PLS-DA.



(a)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Specificity
2
4
6
8
10
12
14
16
18
20
Number of Latent Variables
MC-PLS-DA
PLS-DA (Conventional)

(b)



(c)

Figure 6.2 Variation of sensitivity (a), specificity (b) and accuracy (c) with the number of LVs.

Optimal PLS-DA model number

The second experiment was to determine the number of PLS-DA models required to build a MC-PLS-DA model. The procedure defined by the MC-PLS-DA algorithm shown in Figure 6.1 was run repeatedly here where, at each run, a different number of top PLS-DA models was created to obtain a MC-PLS-DA model. Based on the results shown in Figure 6.2, the number of LVs was set at 6 in this experiment since the performance of MC-PLS-DA did not make significant improvement beyond this value.

The number of PLS-DA models generated at each individual run ranged from 20 to 50,000, and the top 5% of the models at each run (i.e. ranging from 1 to 2,500) were selected and averaged to create the corresponding MC-PLS-DA model.

The effects of the number of PLS-DA models on the sensitivity, specificity and the accuracy of the resulting MC-PLS-DA models were investigated. The results shown in Figure 6.3 demonstrate that the variation range of sensitivity and specificity were within approximately 0.015, and that of the overall accuracy was about 0.01.The performance was relatively low, however, when the number of PLS-DA models used was below 200; that is, using only one to ten top performing models. The overall accuracy appeared to saturate when over 5,000 PLS-DA models were used to create the MC-PLS-DA model. It did not change remarkably even when the number was further increased to 50,000. Therefore, using 5,000 PLS-DA models to build the MC-PLS-DA model appears reasonable to save computation time.

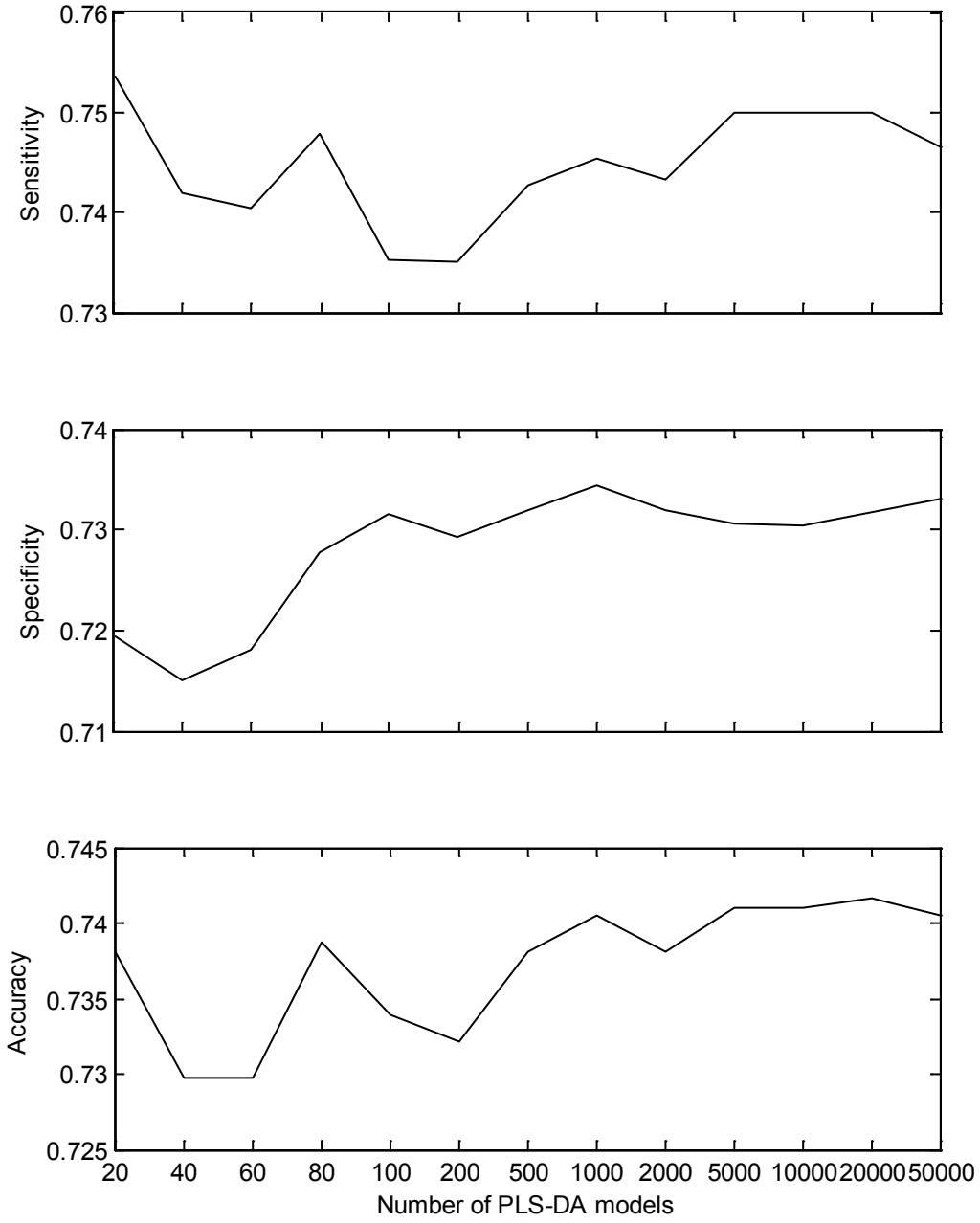


Figure 6.3. The sensitivity, specificity and accuracy obtained with various numbers of PLS-DA models

6.1.3 Performance comparison

Based on the findings obtained from the previous experiments, a MC-PLS-DA model was created to predict the BGL for NIR spectral data collected. The number of LVs and PLS-DA models was six and 5,000 respectively. The predictor matrix was obtained by averaging the training results. The performance of the MC-PLS-DA model was compared to that of conventional PLS-DA, LDA, NN and SVM. The same training and prediction sets were used for the comparison.

Table 6.1 shows the sensitivity, specificity, and the overall accuracy of the five methods. The results indicate that MC-PLS-DA outperforms the other methods in all three aspects. The overall classification rate of the MC-PLS-DA model reached 75.2%, which was the highest accuracy among all methods.

Method	Sensitivity	Specificity	Accuracy
LDA	64.6 %	65.3 %	64.8 %
NN	72.0 %	66.5 %	69.3 %
SVM (RBF)	72.9 %	68.6 %	70.7 %
PLS-DA	72.5 %	68.0 %	70.2 %
MC-PLS-DA	78.0 %	72.5 %	75.2 %

Table 6.1 Sensitivity, specificity and the accuracy of the five methods

The findings in these experiments suggest that the proposed MC-PLS-DA is a feasible method for classifying NIR spectral data and can be used to monitor the BGL of DM patients through non-invasive measurement. With the enhancement achieved by the Monte Carlo method, the MC-PLS-DA method is more stable and accurate than the conventional PLS-DA method.

6.1.4 Discussion

The MC-PLS-DA method is proposed to tackle the problems of the PLS-DA approach in which accuracy is limited by the use of one single prediction model. The method integrates the Monte Carlo method into the conventional PLS-DA to improve performance. The results show that the MC-PLS-DA provides higher sensitivity than the other methods — more than 20% — when compared with LDA and around 8% more than NN, SVM and conventional PLS-DA. In addition, MC-PLS-DA performs better in specificity than the other methods as none of them could achieve higher than 70% while MC-PLS-DA could reach 72.5%. The accuracy rate of MC-PLS-DA is 11% higher than LDA and around 6% higher than NN, SVM and conventional PLS-DA. These advantages make MC-PLS-DA a promising approach for NIR spectra analysis.

6.2 Regression analysis

Multivariate calibration methods for regression analysis are commonly used to extract relevant information from different types of spectral data to predict analyte concentrations (Eriksson, Gottfries, Johansson, & Wold, 2004; Escandar, Damiani, Goicoechea, & Olivieri, 2006; Smith, 2002; Wold, Cheney, Kettaneh, & McCready, 2006). They are particularly useful in spectral analysis because the concurrent inclusion of large spectral data for the analyte can greatly improve the precision and applicability

of quantitative analysis. Multiple linear regression (MLR) (Ostrom, 1990; Weisberg, 2005), principal component regression (PCR) (Martens & Naes, 1989; Naes, Isaksson, Fearn, & Davies, 2002; Naes & Martens, 1998) and partial least squares regression (PLS) (Hoskuldsson, 1988; Paul & Bruce, 1986) are most common methods used to construct the quantitative model in spectral analysis.

As mentioned in Chapter 2, many researchers are developing a variety of non-invasive methods that monitor blood glucose (Fine & Shvartsman, 2003; Heise, Bittner, & Marbach, 1998; Maruo, Tsurugi, Chin, Ota, Arimoto, & Yamada, 2003; Robinson, Eaton, Haaland, Koepp, Thomas, & Robinson, 1992). NIR spectroscopy has become a promising technique for blood glucose monitoring among those potential non-invasive approaches. However, an appropriate model of spectral response in humans is yet to be determined. The key component of NIR spectral quantitative analysis is dependent on multivariate calibration methods that require sufficiently useful calibration spectra and sufficient spectral data points to allow analytical information to be accurately extracted from spectra. The quality of a multivariate calibration model mainly depends on the quality of both response and independent variables. As mentioned, PLS is the most common multivariate calibration procedure to build the quantitative model in NIR spectroscopy because PLS attempts to maximize covariance between the response and independent variables. However, the performance of the PLS model will be severely degraded in the presence of outliers (Chen, Shao, Hu, & Su, 2004; Mittermayr, Tan, & Brwon, 2001). In practice, an analytical data matrix generally includes unexpected experimental errors or measurement noise and normally contains hundreds of samples in the NIR spectra. Hence, outliers possibly exist in such a large data matrix.

The fact that outliers can deteriorate the quantitative model derived from a normal PLS analysis has motivated the development of algorithms that are less affected by outliers (Pell, 2000; Pierna, Wahl, de Noord, & Massart, 2002). These algorithms should be effective for extracting data for use in process modelling. Although various PLS methods (Griep, Wakeling, Vankeerberghen, & Massart, 1995; Kasemsumran, Du, Maruo, & Ozaki, 2006) have been developed for the problem of outliers, most of these methods use a single quantitative model to predict the relationship between NIR spectra and blood glucose concentration. A single prediction model from PLS analysis may underestimate or overemphasize some features, and even worse ignore some important characteristics contained in the largely complex NIR spectra. The prediction results obtained by a single prediction model may be unstable or correlated with spurious spectral variance particularly when the training set is comparatively small. In this regard, a new approach to enhance both stability and performance of PLS prediction models is required.

Like the method presented in section 6.1.1, a PLS model with improved stability and performance derived from Monte Carlo simulation (Doucet, Freitas, & Gordon, 2001; Robert & Casella, 2004) has been presented. The Monte Carlo method is used to rate and analyze NIR spectra by simulating a large number of PLS models from calibration subsets that are randomly selected from the whole calibration set. The reason is that since the number of outliers will be significantly less than that of normal samples; the PLS models can minimize the adverse effect caused by the outliers. It is then determining the mean value over the models with small prediction errors. The advantage

of the Monte Carlo method over other techniques becomes more prominent where the sources of uncertainty increase. It is therefore suitable for predicting blood glucose concentration since glucose measurement in humans is subject to a number of confounders. The results show that the Monte Carlo method can improve both stability and performance of PLS prediction models of the NIR spectrometry.

6.2.1 Algorithm of MC-PLS

As mentioned in section 3.2.3, in the linear least squares model, the prediction $\mathbf{y}$ is computed by the equation

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon} \tag{6.7}$$

where $\mathbf{X}$ is a n x p matrix containing p dependent variables of n samples, $\mathbf{b}$ is a p x 1 vector of regression coefficients obtained from PLS analysis, and $\boldsymbol{\varepsilon}$ is the model offset.

The mechanism in this algorithm is similar to section 6.1.1 although some of the setting is modified. The raw spectra are randomly divided into two parts - training set $\mathbf{X_t}$ and prediction set $\mathbf{X_p}$. The prediction set is used to evaluate the robustness of the prediction model because it will not be involved in training the model. A large number of subsets i.e. 60% of the samples in the training set, will be randomly selected for training, and therefore a large number of PLS models are constructed by using these subsets from the whole training set. After that, these PLS models will be validated by the prediction set, and those providing the smallest RMSE of prediction as shown in (3.25); that is, 5% of

the overall models, will be selected and the averaged prediction of these models taken as the final model. After that, the averaged PLS prediction model is calculated as

$$\bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i \tag{6.8}$$

where N is the number of 5% of the overall models, and $\mathbf{b}_i$ is the prediction model of the smallest 5% RMSE of prediction of the overall models.

Figure 6.4 shows the MC-PLS algorithm with the number of samples mentioned. The main advantage of this approach is to obtain sufficient PLS models without outliers for the purpose of outlier detection. Finally, the performance of this averaged PLS model is assessed by means of the root mean square error of prediction ($RMSE_p$). All the results are averaged over 20 runs.

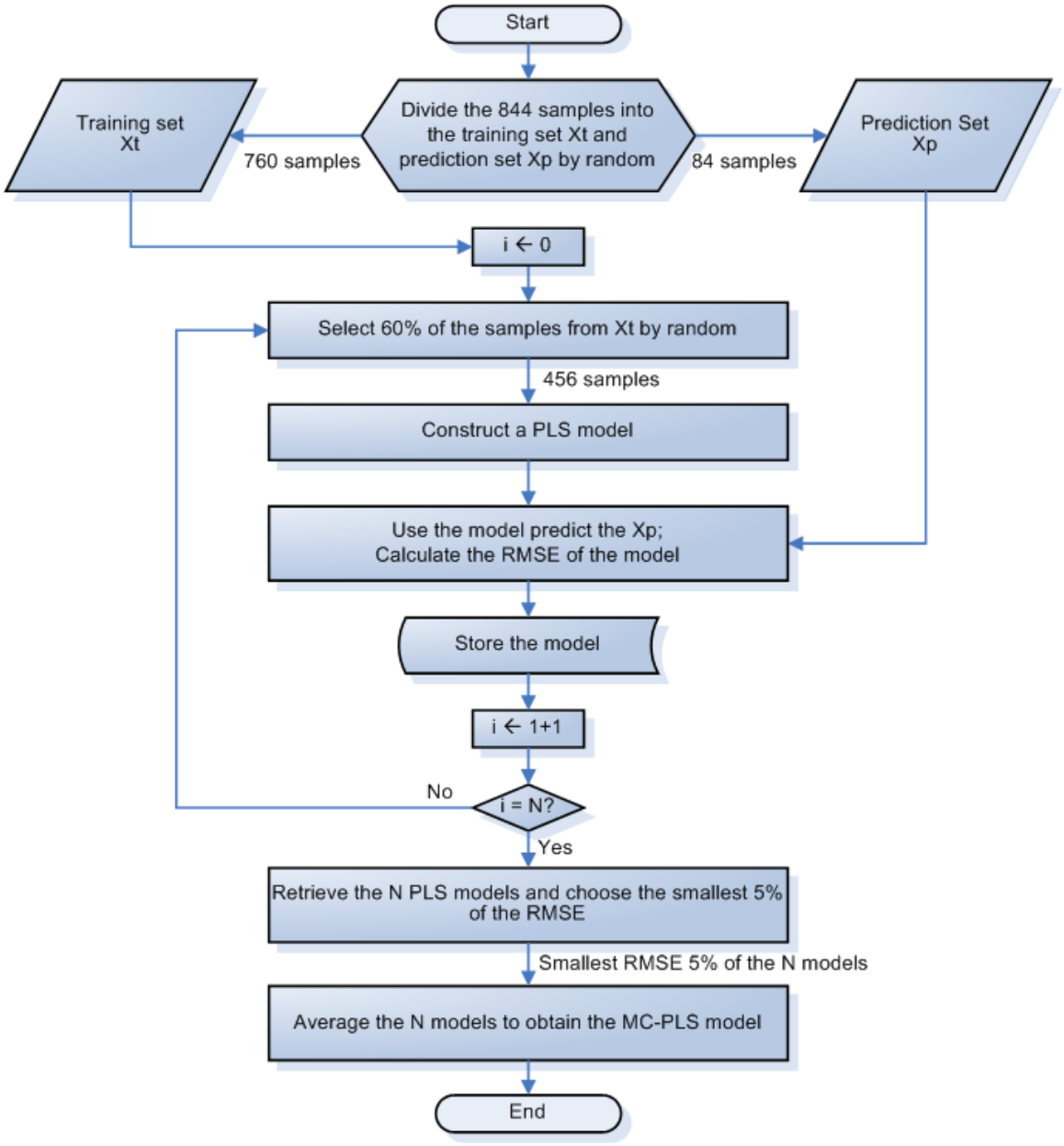


Figure 6.4 The MC-PLS algorithm

6.2.2 Performance of MC-PLS

Optimal number of LVs for PLS model

Firstly, the number of latent variables for the PLS model was determined because it was an important parameter in the modeling. Figure 6.5 shows the variation of $RMSE_p$ with the number of LVs of MC-PLS and conventional PLS. In this simulation, 1,000 PLS models were built and the optimal 5% of the overall models (50 models) were selected for the MC-PLS. The same training sets and prediction sets were used in both models for the sake of fairness. The figure shows that the MC-PLS provides smaller $RMSE_p$ compared with conventional PLS and has a descending trend with the increase of the number of LV while stabilizing after LV is larger than six. The correlation coefficient in Figure 6.6 shows that the performance of MC-PLS is higher than the conventional PLS and it stabilizes after LV is larger than five. The results confirm that using the Monte Carlo approach for PLS can perform better than the conventional PLS. The number of LV was set equal to six for the following simulations according to those results.

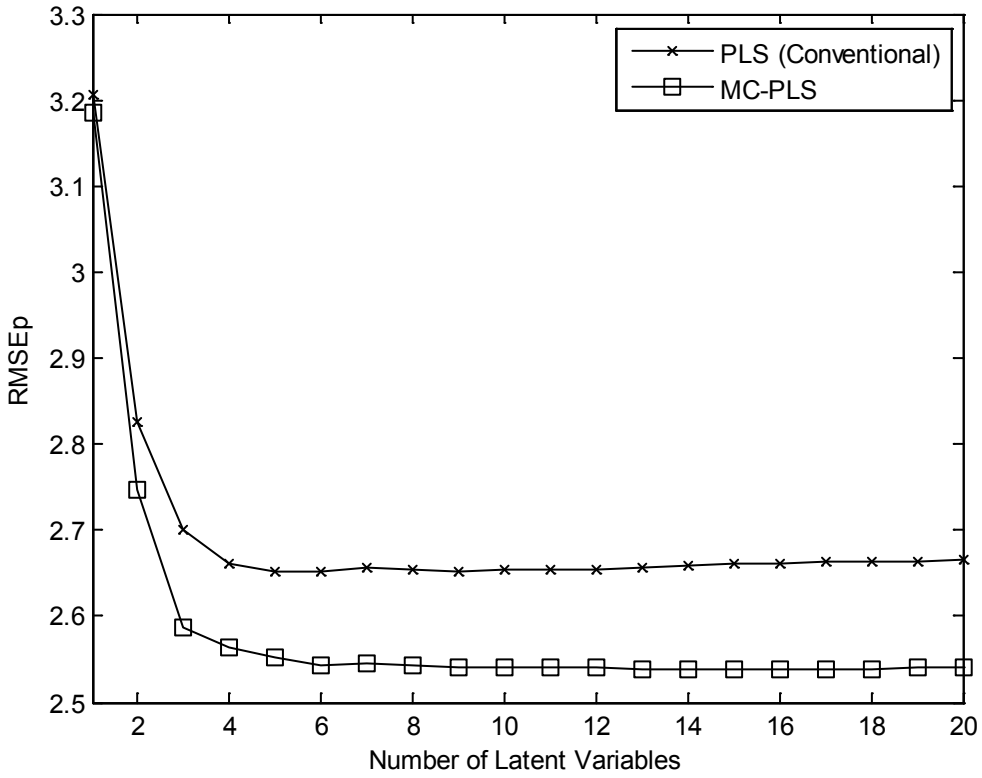


Figure 6.5 Variation of RMSEp with the number of LVs

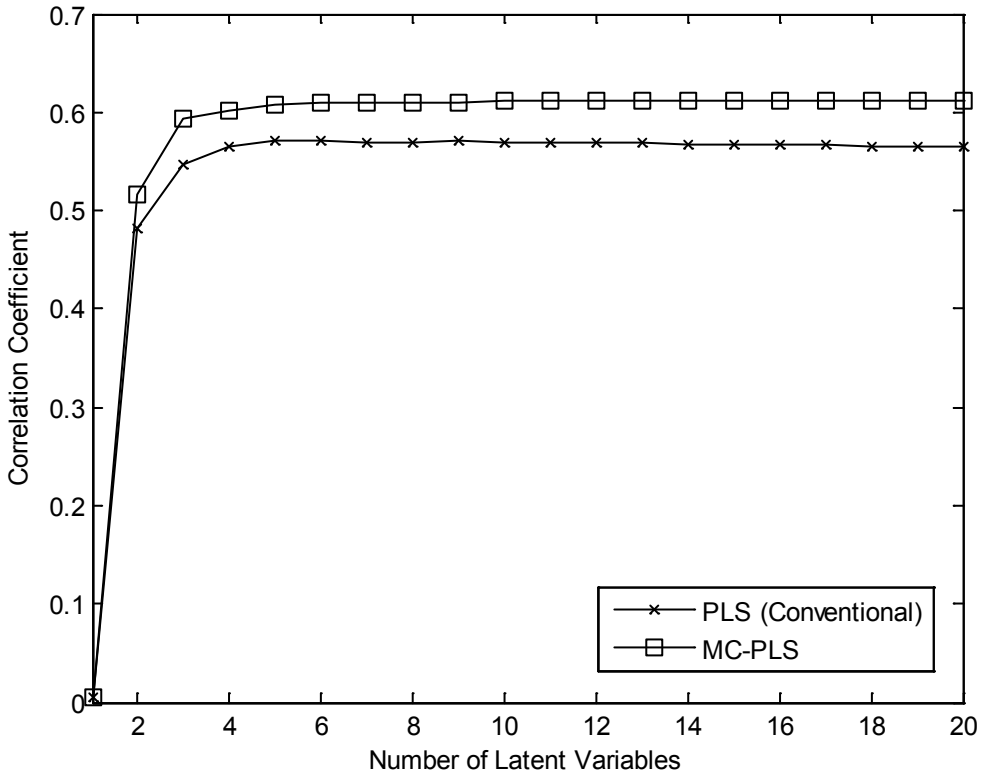


Figure 6.6 Variation of correlation coefficient with the number of LVs

Optimal PLS model number

Secondly, the number of PLS models for MC-PLS was determined; this decides the stability and accuracy of the model. When the number of PLS models is too small, the stability and accuracy of the model may be affected because it may either overemphasize or underestimate some aspects, and ignore many important features contained in such a large NIR data set. However, if the number of PLS models is too large, uninformative features may be contained in the model and, even worse, in the prediction result. Hence, the variation of the $RMSE_p$ for the number of PLS models is investigated. Figure 6.7 and

Figure 6.8 show the $RMSE_p$ and R_p obtained with various numbers of PLS models from 20 to 80,000. For each simulation, 5% of overall models that provided the smallest $RMSE_p$ were selected and averaged to develop the MC-PLS model. The number of models selected and the standard error of the $RMSE_p$ for the simulation trails are shown in Table 6.2. Simulation results show that the $RMSE_p$ was large for the case of one model selected out of 20 PLS models then, with the number of PLS models increased, the $RMSE_p$ becomes smaller. The $RMSE_p$ was steadily saturated at 1,000 PLS models even for the large number of PLS models (that is 80,000) were calculated for use in the number of models selection. It can be seen on one hand that the fewer the PLS models, the more significant the effect of outliers and thus a more deteriorated performance. On the other hand, when too many PLS models were used, the performance was not significantly improved. Therefore, 1,000 PLS models were used for MC-PLS.

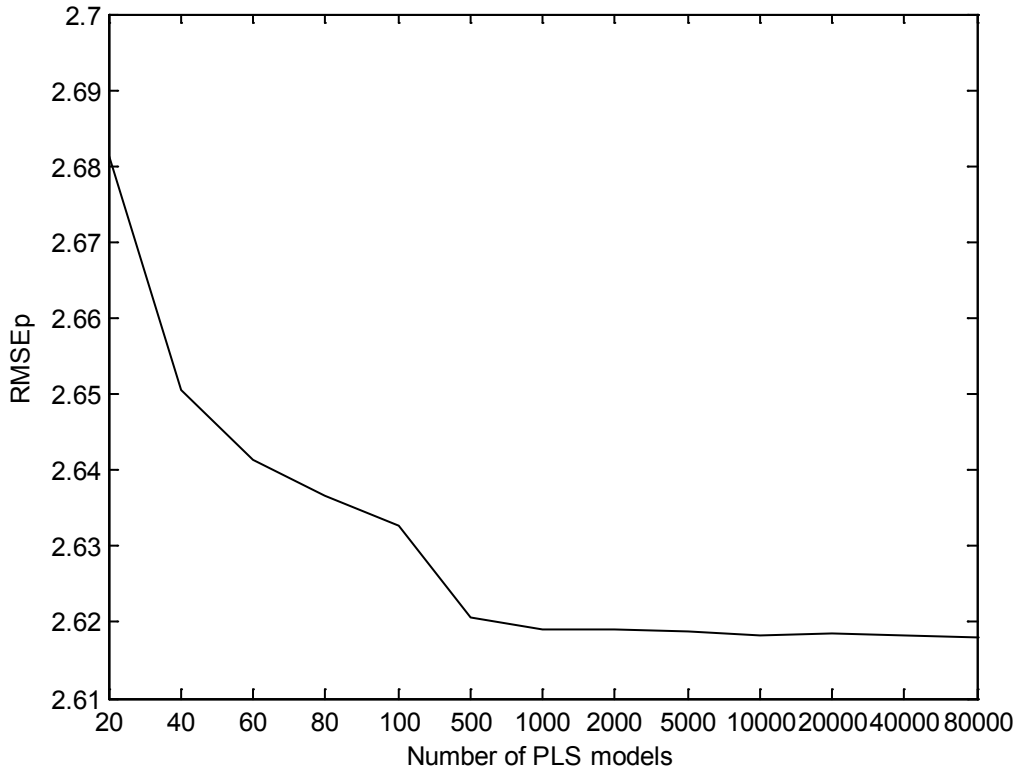


Figure 6.7 Variation of $RMSE_p$ with the number of PLS models

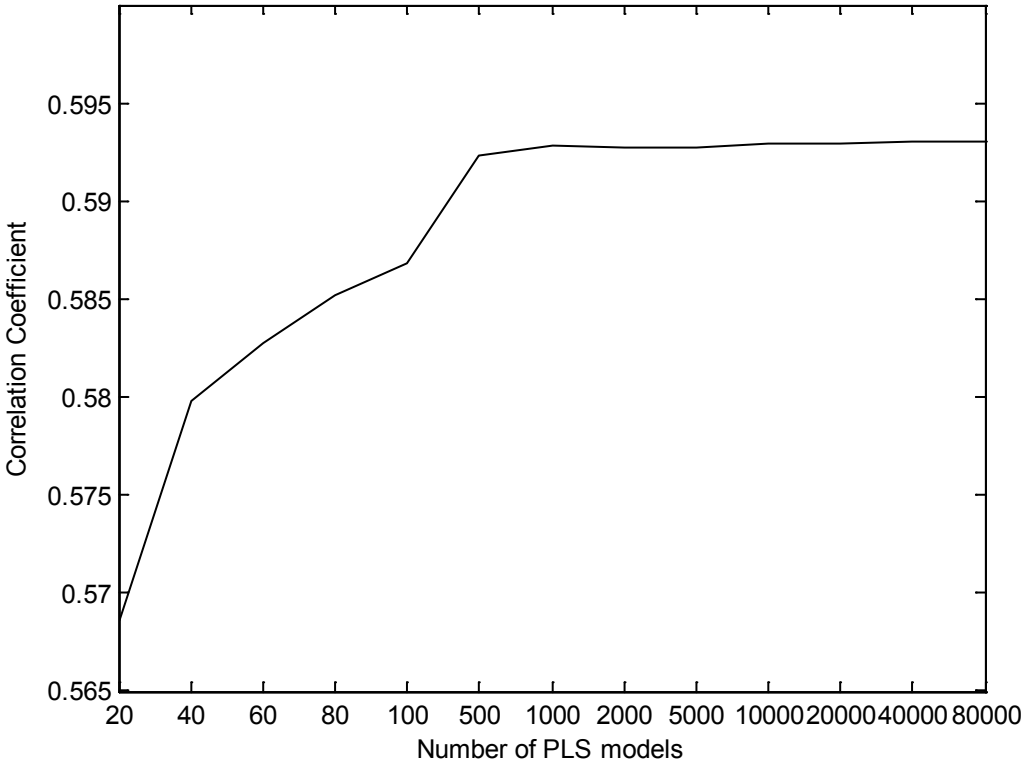


Figure 6.8 Variation of correlation coefficient with the number of PLS models

No. of PLS models	No. of model(s) selected	RMSEp (σ)
20	1	2.6812 (0.2610)
40	2	2.6506 (0.2581)
60	3	2.6412 (0.2597)
80	4	2.6366 (0.2555)
100	5	2.6328 (0.2567)
500	25	2.6206 (0.2597)
1000	50	2.6190 (0.2607)
2000	100	2.6190 (0.2608)
5000	250	2.6188 (0.2600)
10000	500	2.6183 (0.2595)
20000	1000	2.6185 (0.2596)
40000	2000	2.6183 (0.2594)
80000	4000	2.6181 (0.2595)

Table 6.2 $RMSE_p$ with the mean value obtained by various PLS models

6.2.3 Performance comparison

Finally, a MC-PLS model was constructed based on the previous findings. The training procedure was repeated 1,000 times, the number of LVs was set to six, and the predictor matrix was obtained with the mean of the training results. Based on the findings, a model was developed to predict the glucose value using PLS analysis. The performance of the MC-PLS model was compared to that of conventional PLS, MLR and PCR. The same training and prediction sets were used for the comparison. Table 6.3 shows the $RMSE_p$ and R_p for the four methods. The results indicate that MC-PLS provides the smallest $RMSE_p$ and the highest R_p among all methods.

Method	$RMSE_p$	R_p
MLR	2.8107	0.5186
PCR	2.9693	0.3851
PLS	2.5453	0.6143
MC-PLS	2.4392	0.6535

Table 6.3 Performance comparison of the four methods

To further determine the accuracy of estimating blood glucose level using this model, a Clarke Error Grid Analysis (EGA) was employed (Clarke, Cox, Gonder-Frederick, Carter, & Pohl, 1987). The EGA breaks down a scatterplot of a reference glucose measurement and a predicted glucose measurement into five zones as shown in Table 6.4. The more the values appear in A and B, the more accurate the device in terms of clinical utility.

Zone	Distribution	Clinical utility
A	Within +/-20% of the reference glucose value	Accurate (acceptable) glucose results
B	Outside +/-20%, within 95% CI	Accurate (acceptable) glucose results
C	Beyond 95% CI	Unnecessary corrections that could lead to a poor outcome
D	Beyond 95% CI	A dangerous failure to detect and treat
E	Beyond 95% CI	Erroneous treatment

Table 6.4 Description of zones in terms of clinical utility for EGA

Table 6.5 shows the comparison among MLR, PCR, conventional PLS and MC-PLS and the scatterplot of MC-PLS is shown in Figure 6.9. The estimated glucose values are plotted against the referenced glucose values. The results show that 100% of the data points fell within Zone A and Zone B for MC-PLS while other methods do have some predicted values falling into Zone D, which is a dangerous failure to detect and treat in clinical utility. Based on the results mentioned, it is proven that the Monte Carlo approach can not only stabilize the conventional PLS model, but also improve the performance of the prediction model.

	MLR	**PCR**	**PLS**	**MC-PLS**
$RMSE_p$	2.7633	2.9747	2.5635	2.4763
R_p	0.5372	0.3743	0.6020	0.6348
EGA				
A	33	29	38	38
B	48	50	45	47
C	0	0	0	0
D	4	6	2	0
E	0	0	0	0

Table 6.5 Comparison of results obtained by the four methods

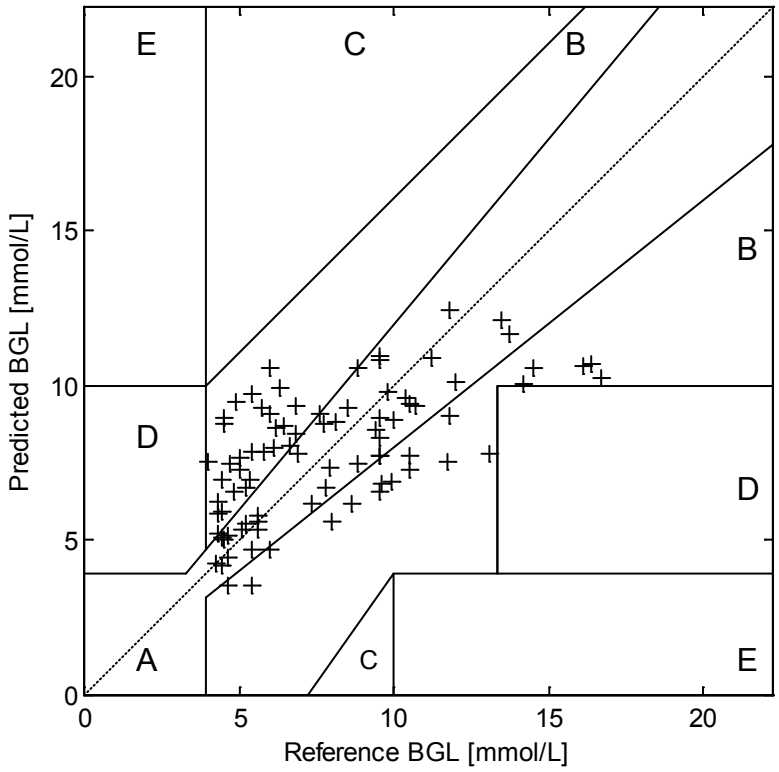


Figure 6.9 EGA prediction results of MC-PLS

6.2.4 Discussion

This study has presented an improved algorithm based on the Monte Carlo method to average the PLS model results, and the model provides stable, accurate and non-invasive blood glucose measurement. The results show that the MC-PLS provides smaller $RMSE_p$ than MLR and PCR, which have more than a 13% improvement, and further advance the conventional PLS with more than 4% improvement. The correlation coefficient is also greatly enhanced with nearly 70% enhancement compared with PCR, more than 18% improvement compared with MLR, and around 6% advancement over

conventional PLS. The results of these experiments and field trials support the feasibility of measuring blood glucose level using NIR spectroscopy with the proposed algorithm.

6.3 Summary

This study proposed the MC-PLS-DA method for the classification approach and the MC-PLS method for the regression approach to tackle the problems in which the performance is limited by the use of one single prediction model of PLS. These methods integrate the Monte Carlo method into the conventional PLS-DA and conventional PLS to improve their overall performance. The MC-PLS-DA algorithm exhibits better sensitivity, specificity and overall classification rates when compared with LDA, NN, SVM and conventional PLS-DA, as evident from the BGL classification results on the NIR spectral data. In addition, the MC-PLS algorithm reveals smaller RMSE and higher R when compared with MLR, PCR and conventional PLS, which provides more stable and accurate results for blood glucose measurement.

CHAPTER 7

DISCUSSION AND CONCLUSION

7.1 Discussion

This study successfully adopted the four-stage framework of bio-signal processing to handle the NIR spectroscopy that is used to measure blood glucose level. A vigorous systematic approach from signal acquisition to signal interpretation has been established. These include, in the first stage, the assembly of the specific equipment for NIR spectroscopy, which includes a finger clip at the end of the probe to provide a constant force and firmly fix the scanning position during measurement. The study also identified the left ring finger as probably the best measurement site to acquire NIR spectral data. In addition, the temperature effect relative to body location was shown to be not significant across different temperature ranges. Moreover, the study proposed using a 2-norm normalization process, together with the GLS weighting pre-filtering process, during the signal transformation stage so as to minimize the disturbance noise caused by the previous stage. After that, the study suggested using the genetic algorithm for feature (wavelength) selection in the parameter selection stage. Finally, in the signal interpretation stage, improved methods based on the Monte Carlo approach for the PLS

multivariate analysis were introduced, which were called MC-PLS-DA for classification analysis and MC-PLS for regression analysis to predict the result. Figure 7.1 shows the proposed methods together with the four-stage framework for easy reference.

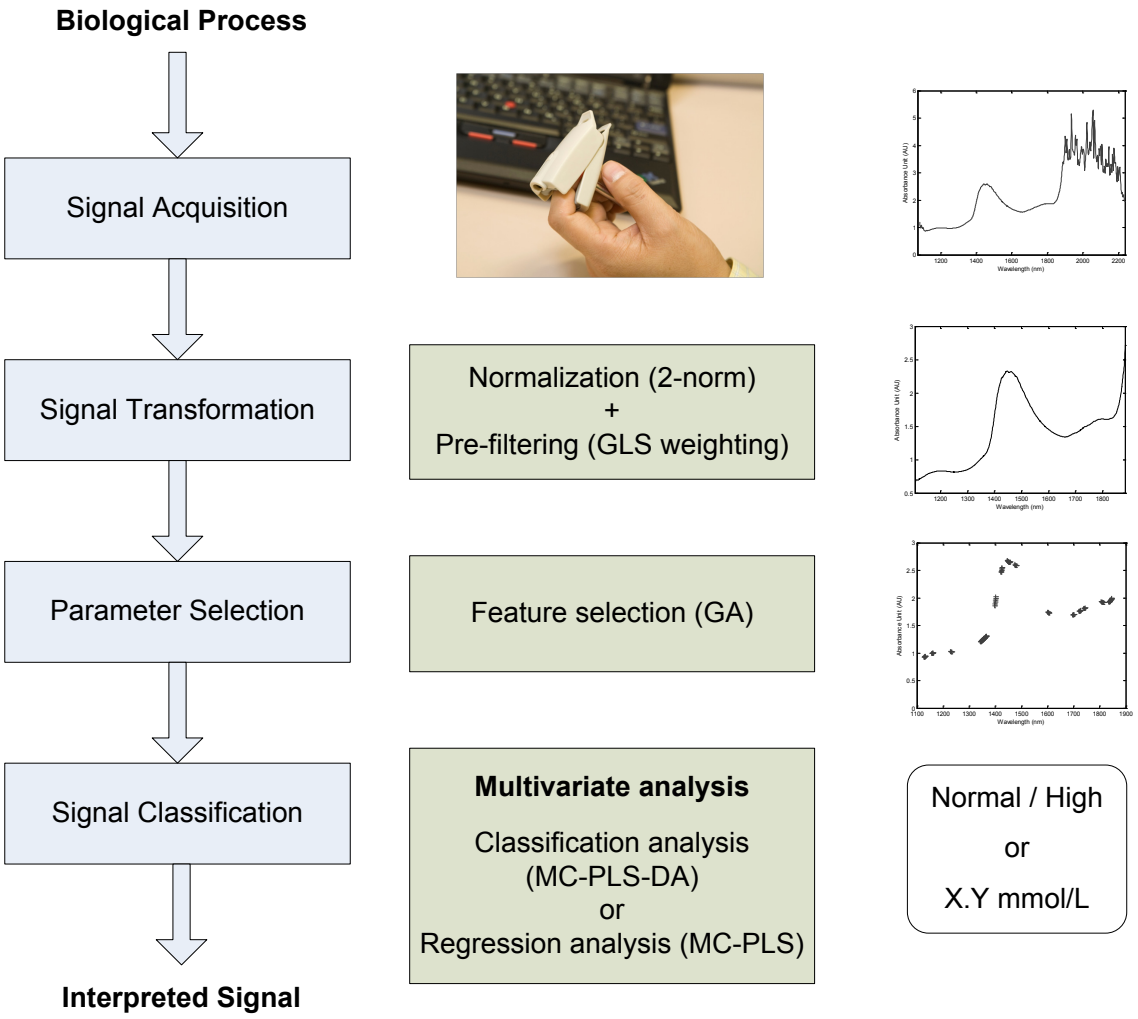


Figure 7.1 The four-stage framework of the proposed methods

The proposed multivariate analysis by making use of classification analysis for non-invasive blood glucose measurement is the new initiative for the purpose of DM

screening. The blood glucose level was divided into two classes that were listed in Equation (6.3), the normal BGL and the high BGL. The prediction output in this method is not represented in terms of exact value of the BGL, simply to classify it into normal or high BGL. The MC-PLS-DA integrates the Monte Carlo method into the conventional PLS-DA to enhance performance. It exhibits better sensitivity, specificity, and overall accuracy rate when compared to others. The classification output results for the relationship between the response and the independent variables is more accurate, hence enhancing the reliability of the classification model, especially for such a large training set in this study. These advantages make MC-PLS-DA a possible way for non-invasive blood glucose measurement using NIR spectroscopy.

Apart from the classification analysis method developed for the signal classification stage, the study also proposed another multivariate analysis based on regression analysis for prediction. Like the MC-PLS-DA, it also incorporates the Monte Carlo method with the conventional PLS algorithm, called MC-PLS, to provide more accurate prediction results. The results in section 6.2.3 showed that the MC-PLS can improve the performance of the conventional PLS algorithm and out-performance the other algorithms. In addition, the prediction results all fell within Zone A and Zone B of the EGA, which are accurate for clinical utility. However, it should also be noted that the correlation coefficient was not high enough to provide a good relationship between the response NIR spectra and the blood glucose levels. The reasons include the large sample size of this clinical study that contain many unexpected disturbances and noise from the dataset; for example, the measurement error and the physiological effect of the human

being are the greatest factors for generating such disturbances. Since the ultimate goal was targeted to develop a universal model, the study avoided the use of the user dependent or individual prediction models, which can easily achieve a high correlation coefficient and minimum RMSE.

Moreover, with the reproducibility test that was carried out in the clinical trial, it recognizes the use of blood tests from a laboratory as the gold standard for non-invasive blood glucose measurement. The large sample size of this clinical study revealed the output models, one for classification analysis and one for regression analysis, are attempting to adopt a universal model instead of single user model. Undoubtedly, it is easy to claim success in the development of non-invasive blood glucose measurement with a user dependent prediction model. While this study was trying to reduce the spurious correlation between blood glucose concentrations and spectral features within the data set from a single subject or few subjects.

The capability to determine blood glucose non-invasively in humans is a complex and difficult analytical problem. Many researchers and industry groups have attempted to measure blood glucose by various non-invasive methods (as mentioned in Chapter 2). Those early manuscripts published in this area were either measuring glucose solutions or measuring glucose with animal tests. It is particularly important to seek for publication with the test based on human. In this regard, Table 7.1 lists the manuscripts that fell between 2001 to 2011, and that the experimental setup used NIR spectroscopy with multivariate analysis in testing human subjects.

Method	Details	Results	Reference
Absorption of skin (wavelength range from 1,300 – 1,900nm)	Six subjects, OGTT, one calibration model per subject, forearm measurement	Overall correlation coefficient was 0.934, standard error of prediction was 1.32mmol/L, EGA: Zone A: 71.3%, Zone B: 21.3% and Zone D: 7.4%	(Maruo, Tsurugi, Tamura, & Ozaki, 2003)
Absorption of skin (wavelength range from 1,200 – 1,900nm)	One subject, OGTT, skin measurement	Correlation coefficient was 0.928, standard error of prediction was 1.79mmol/L, EGA: Zone A: 87.5%, Zone B: 8.3% and Zone D: 4.2%	(Maruo, Tsurugi, Chin, Ota, Arimoto, & Yamada, 2003)
Absorption of skin (wavelength range from 1,212 – 1,889nm)	One subject, OGTT, finger measurement	Correlation coefficient was 0.89, RMSEV was 1.12mmol/L, EGA: Zone A: 86%, Zone B: 7% and Zone D: 7%	(Kasemsumran, Du, Maruo, & Ozaki, 2006)
Absorption of skin (wavelength range from 900 – 1,700nm)	23 subjects, OGTT, fingertip measurement	EGA: Zone A: 93.7%, Zone B: 6.3% and Zone C-E: 0%	(Yamakoshi, et al., 2007)
Absorption of skin (wavelength range from 1,550 – 1,800nm)	One subject, 4 days finger measurement	Correlation coefficient was 0.8785, RMSEp was 0.8682mmol/L, EGA: 100% fall within Zone A & B	(Chuah, Paramesran, Thambiratnam, & Poh, 2010)

Table 7.1 List of NIR spectroscopy experiments in human tests

It is relatively simple to measure data and study the correlation with glucose solution tests or blood glucose levels under controlled conditions in research laboratories with only a few subjects/samples. The challenge here is to measure these variables in normal

and practical environments with large datasets of different subjects. This requires understanding the physical and physiological factors that may affect blood glucose measurement. It is important to notice that non-invasive monitoring will never be achieved without vigorous scientific and clinical evidence. At this stage, it is still far away from achieving the goal of non-invasive blood glucose monitoring, with many technical issues yet to be resolved.

7.2 Future work

The performance evaluation according to the four-stage framework in this study indicates promising directions of non-invasive blood glucose measurement using NIR spectroscopy. While advances have been made, there is still a great demand to further improve the accuracy and the reliability of non-invasive instruments so as to reach a level comparable to the gold standard. Some future work in the context of this thesis are suggested as follows.

1. The performance of the hardware for NIR technology must be sufficiently high to ensure the data acquired on blood glucose is reliable and the uninformative noise is minimal. Because the signal-to-noise ratio (SNR) of the instrumentation delineates the limit of detection for blood glucose in absorption spectroscopy, it is understood that an instrument with high SNR is highly recommended for spectral signal collection from humans. Mark's team suggested an instrument

with the SNR on the order of 50,000 is required in view of the relatively weak absorption property associated with blood glucose (David & Julie, 2010).

2. Thickness of the human tissue is a critical factor because it determines the absorbance according to Beer's Law in (1.1). Theoretically, the tissue layer must be thick enough to allow analyate molecules within the optical path to generate a signal. However, if the tissue layer is too thick or too thin, excessive scattering or inadequate scattering can result in loss of signal. According to the experimental results in this study, it was found that different measurement sites resulted in quite a different variation of prediction results, while the finger tip measurement sites (excluding the thumb) provided comparably similar prediction results. This may be due to the fact that the thicknesses of the tissues in different position of finger tip are similar. In view of this, it is suggested using the finger tip (the left ring finger, FL4) for the measurement site but it is worth paying attention that the thickness of the human tissue can influence the prediction results in the localized region of the measurement.

3. Beer's Law was applied in this study to determine the absolute concentration of glucose molecules present in blood. The law states that the quantity of light absorbed by a substance is directly proportional to the concentration of the substance and the path length of the light through the solution. However, this law does not take into consideration either reflection or scattering of incident light that accounts for the loss of transmitted light intensity because human tissue is

quite a high scattering medium. Since physiological differences would affect the reliability of different technologies, they are mainly due to individual metabolism, blood components, and other bodily fluid circulations for body regulation. Absorption spectroscopy mainly detects the glucose molecule, and glucose can be found everywhere in the human body. Maikala proposed modifying Beer's Law in relation to the optical properties of human tissue, which may help to explain more for this application (Maikala, 2010).

4. Since the colour of skin may be one factor that affects NIR spectroscopy measurement, this study mainly focused on the East Asia group of people — people of Asian descent (tan skin tone) to avoid different combinations of skin colour, thereby affecting the model's validation. It is undoubtedly worth carrying out other clinical trials to collect other skin tones so as to develop a universal model for people all over the world.

5. In addition, the force exerted on the measurement area may affect the deformation of the contact point of the tissue. This study involved applying a finger clip that provides constant force to measurement at the finger tip. However, it may likely produce a poor prediction result after a long period of time because different deformations of the tissues may cause diverse absorption or reflection properties and thus affect the resulting signal. Moreover, in view of the fact that the non-invasive technology is based on NIR optical sensing technique, a time lag may occur between measurements of blood glucose content from different

parts of body. This may introduce error in calibration, and hence this issue is worth investigating.

6. The Monte Carlo method applied to PLS-DA for classification approach was proposed to handle the two level classification: normal BGL and high BGL. It is noticeable that BGL can be further categorized into three or even four levels, such as the low BGL and very high BGL. Undoubtedly, PLS-DA may not be suitable for this multivariate analysis because of its two level classification in nature, yet this issue goes far beyond demonstrating the feasibility of classification approach for non-invasive blood glucose monitoring but pertains to its eventual practical implementation.

7. Although an improved method was investigated for quantitative analysis that can enhance the correlation of the spectroscopic properties of the glucose molecule with glucose concentration in blood, more effort should be taken to rigorously extend the technique to non-invasive blood glucose monitoring. In regression analysis, the correlation between the spectral data and the predicted concentration is not encouraging; the reason may be due to non-linearity of the spectral data for multivariate analysis. It is not easy to tackle the non-linear data for multivariate training method such as PLS and it is worth making an effort on other nonlinear multivariate methods, attempting to carry out the training method for a better prediction model.

8. Though this study developed a model with improved stability, the results need to be interpreted with caution. In this experiment, because of the limitation of the NIR spectrometer, many uninformative noises were generated in the wavelength range higher than 1,900nm; the constraint inherent in the study is the impossibility of using a higher range spectrometer (wavelength range 2,000 – 2,500nm). Thus, there is a possibility that the higher spectra portion was lost for analysis.

7.3 Conclusion

In this century, particularly in wealthy developed countries, diabetes is a complex group of syndromes that have a disturbance in the body's use of glucose. It can be controlled by an appropriate regimen that includes patient education, weight management, diet control, sensible exercise, oral medication, and insulin therapy. However, these diabetes management and medications rely heavily on blood glucose measurement. With ever-improving advances in diagnostic technology, the race for the next generation of bloodless, painless, accurate and consistent blood glucose measurement instruments has begun. Nevertheless, many hurdles remain before these products reach commercial markets.

This study identified a promising measurement site for non-invasive measurement, proved the temperature difference at the measurement site was insignificant to the prediction results, verified the advantage of the data pre-processing treatment of spectral

data, and proposed an improved method for multivariate analysis. Considerable progress has been made in the development of non-invasive blood glucose measurement devices. In principle, the approach can be used for screening purposes for DM prevention. However, for diabetes that needs frequent testing, using invasive blood glucose measurement via finger pricking remains a practical way to provide suitable information for diabetes management at this current stage.

REFERENCES

American Diabetes Assoication. (2012). Standards of medical care in diabetes. *Diabetes Care* (35), pp. S11-63.

Arnold, M., & Small, G. (2005). Noninvasive glucose sensing. *Analytical Chemistry* , pp. 5429-5439.

Arnold, S., Harvey, L., McNeil, B., & Hall, J. (2002). Employing near-infrared spectroscopic methods of analysis for fermentation monitoring and control. *BioPharm International* , pp. 26-34.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics , 17*, pp. 166-173.

Barnes, R., Dhanoa, M., & Lister, S. (1989). Standard normal variate transformation and detrending of NIR spectra. *Applied Spectroscopy , 43*, pp. 772-777.

Berger, A., Koo, T., Itzkan, I., Horowitz, G., & Feld, M. (1999). Multicomponent blood analysis by near-infrared Raman spectroscopy. *Applied Optics* (38), pp. 2916-2926.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational Learning Theory* (pp. 144-152). ACM Press.

Brancaleon, L., Bamberg, M., Sakamaki, T., & Kollias, N. (2001). Attenuated total reflection-Fourier transform infrared spectroscopy as a possible method to investigate biophysical parameters of stratum corneum in vivo. *Journal of Investigative Dermatology , 116*, pp. 380-386.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery , 2*, pp. 121-167.

Burns, D., & Ciurczak, E. (2008). *Handbook of near infrared analysis 3rd edition.* Boca Raton, FL: CRC Press/Taylor & Francis.

Caduff, A., Hirt, E., Feldman, Y., Ali, Z., & Heinemann, L. (2003). First human experiments with a novel non-invasive, non-optical continuous glucose monitoring system. *Biosensors and Bioelectronics , 19* (3), pp. 209-217.

Caduff, A., Talary, M., Mueller, M., Dewarrat, F., Klisic, J., Donath, M., et al. (2009). Non-invasive glucose monitoring in patients with Type 1 diabetes: a multisensor system

combining sensors for dielectric and optical characterisation of skin. *Biosensors and Bioelectronics* (24), pp. 2778-2784.

Chen, D., Cai, W., & Shao, X. (2008). A strategy for enhancing the reliability of near-infrared spectral analysis. *Vibrational Spectroscopy* , pp. 113-118.

Chen, D., Shao, X., Hu, B., & Su, Q. (2004). A background and noise elimination method for quantitative calibration of near infrared spectra. *Analytica Chimica Acta , 511*, pp. 37-45.

Chiang, L., Russell, E., & Braatz, R. (2000). Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and Intelligent Laboratory Systems , 50*, pp. 243-252.

Chuah, Z., Paramesran, R., Thambiratnam, K., & Poh, S. (2010). A two-level partial least squares system for non-invasive blood glucose. *Chemometrics and Intelligent Laboratory Systems , 104*, 347-351.

Ciurczak, E., & Drennen, I. J. (2002). *Pharmaceutical and medical applications of near-infrared spectroscopy.* New York: Marcel Dekker.

Clarke, W., Cox, D., Gonder-Frederick, L., Carter, W., & Pohl, S. (1987). Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care , 10*, pp. 622-628.

Cristianini, N., & Shawe-Tayor, J. (2002). *An introduction to Support Vector Machines : and other kernel-based learning methods.* Cambridge, England: Cambridge University Press.

David, D., & Julie, A. (2010). *In vivo glucose sensing.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Department of Health, The Government of the Hong Kong Special Administrative Region. (2012, 3 26). *Centre for Health Protection*. Retrieved from Diabetes Mellitus: http://www.chp.gov.hk/en/content/9/25/59.html

Doucet, A., Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods.* New York: Springler.

Dunn, W., & Shultis, J. (2012). *Exploring Monte Carlo methods.* Amsterdam, Boston: Elsevier.

Eigenvector Research Inc. (2011). *Advanced Preprocessing: Noise, Offset, and Baseline Filtering*. Retrieved 1 30, 2013, from Eigenvector Documentation Wiki: http://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Noise,_Offset,_and_Baseline_Filtering

Eriksson, L., Gottfries, J., Johansson, E., & Wold, S. (2004). Time-resolved QSAR: an approach to PLS modelling of three-way biological data. *Chemometrics and Intelligent Laboratory Systems , 73*, pp. 73-84.

Ermolina, I., Polevaya, Y., & Feldman, Y. (2000). Analysis of dielectric spectra of eukaryotic cells by computer modelling. *European Biophysics Journal , 29*, pp. 141-145. Escandar, G., Damiani, P., Goicoechea, H., & Olivieri, A. (2006). A review of multivariate calibration methods applied to biomedical analysis. *Microchemical Journal , 82*, pp. 29-42.

Falkenauer, E. (1998). *Genetic algorithms and grouping problems.* Chichester, New Work: Wiley.

Fearn, T., Riccioli, C., Garrido-Varo, A., & Guerreo-Ginel, J. (2009). On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems , 96*, pp. 22-26.

Ferre, L. (1995). Selection of components in principal component analysis. *Computational Statistics & Data Analysis* , pp. 669-682.

Fine, I., & Shvartsman, L. (2003). *Patent No. 658,7704.* United States.

Geladi, P., & Kowalski, B. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta , 185*, pp. 1-17.

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy , 39*, pp. 491-500.

Ghasemi, J., & Saaidpour, S. (2007). Quantitative structure–property relationship study of n-octanol–water partition coefficients of some of diverse drugs using multiple linear regression. *Analytica Chimica Acta , 604* (2), 99-106.

Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning.* Boston, MA, USA: Addison-Wesley Longman Publishing Co. Inc.

Gourzi, M., Rouane, A., Guelaz, R., Alavi, M., Mchugh, M., Nadi, M., et al. (2005). Non-invasive glycaemia blood measurements by electromagnetic sensor: Study in static and dynamic blood circulation. *Journal of Medical Engineering & Technology, , 29* (1), pp. 22-26.

Griep, M., Wakeling, I., Vankeerberghen, P., & Massart, D. (1995). Comparison of semirobust and robust partial least squares procedures. *Chemometrics and Intelligent Laboratory Systems , 29*, pp. 37-50.

Hanlon, E., Manoharan, R., Koo, T., Shafer, K., Motz, J., Fitzmaurice, M., et al. (2000). Prospects for in vivo Raman spectroscopy. *Physics in Medicine abd Biology* (45), pp. R1-R59.

Heise, H., & Marbach, R. (1998). Human oral mucosa studies with varying blood glucose concentration by non-invasive ATR-FT-IR-spectroscopy. *Cell. Mol. Biol. (Noisy-le-grand)* (44), pp. 899-912.

Heise, H., Bittner, A., & Marbach, R. (1998). Clinical chemistry and near infrared spectroscopy: technology for non-invasive glucose monitoring. *Journal of Near Infrared Spectroscopy , 6*, pp. 349-359.

Hillier, T., Abbott, R., & Barrett, E. (1999). Hyponatremia: evaluating the correction factor for hyperglycemia. *The American Journal of Medicine , 106* (4), pp. 399-403.

Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics* , pp. 211-228.

Huberty, C., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). Hoboken, NJ.: John Wiley & Sons.

International Diabetes Federation. (2011). *Guideline for the management of postmeal glucose in Diabetes.* International Diabetes Federation, Brussels, Belgium.

International Organization for Standardization. (2003). ISO-15197-2003 International Standard. *In vitro diagnostic test systems - requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus* . Switzerland: International Organization for Standardization.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Ju, W., Shan, J., Yan, C., & Cheng, H. (2009). Discrimination of disease-related non-synonymous single nucleotide polymorphisms using multi-scale RBF kernel fuzzy support vector machine. *Pattern Recognition Letters , 30*, pp. 391-396.

Kasemsumran, S., Du, Y., Maruo, K., & Ozaki, Y. (2006). Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares. *Chemometrics and Intelligent Laboratory Systems* , pp. 97-103.

Kecman, V. (2001). *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models.* Mass, Cambridge: MIT Press.

Kemsley, E. (1998). *Discriminant analysis and class modelling of spectroscopic data.* Chichester, New York: Wiley.

Kemsley, E. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems , 33*, pp. 47-61.

Kettaneh, N., Berglund, A., & Wold, S. (2005). PCA and PLS with very large data sets. *Computational Statistics & Data Analysis , 48*, pp. 69-85.

Khalil, O. (2004, 11 5). Non-invasive glucose measurement technologies: an update from 1999 to the dawn of the new millennium. *Diabetes Technology & Therapeutics* (6), pp. 660-697.

Khalil, O. (2004a). Noninvasive photonic-crystal material for sensing glucose in tears. *Clinical Chemistry , 50* (12), pp. 2236-2237.

Klecka, W. (1980). *Discriminant analysis.* Newbury Park, Calif: Sage Publication.

Klonoff, D. (2005). Continuous glucose monitoring - roadmap for 21st century diabetes therapy. *Diabetes Care , 28* (5), pp. 1231-1239.

Koolman, J., & Roehm, K. (2005). *Color atlas of biochemistry* (2nd ed.). Stuttgart, New York: Thieme.

Larin, K., Eledrisi, M., Motamedi, M., & Esenaliev, R. (2002). Noninvasive blood glucose monitoring with optical coherence tomography: a pilot study in human subjects. *Diabetes Care , 25* (12), pp. 2263-2267.

Lee, S., Nayak, V., Dodds, J., Pishko, M., & Smith, N. (2005). Glucose measurements with sensors and ultrasound. *Ultrasound in Medicine and Biology* (31(7)), pp. 971-977.

Leung, S., Xiong, Y., Lau, W., & So, C. (1999). Whitening prefiltered TLS linear predictor for frequency estimation. *IEE Electronic Letters , 35*, pp. 1232-1233.

Lilienfeld-Toal, H., Weidenmuller, M., Xhelaj, A., & Mantele, W. (2005). A novel approach to non-invasive glucose measurement by mid-infrared spectroscopy: The combination of quantum cascade lasers (QCL) and photoacoustic detection. *Vibrational Spectroscopy , 38*, pp. 209-215.

Lin, H., & Yeh, J. (2009). Optimal reduction of solutions for support vector machines. *Applied Mathematics and Computation , 214*, pp. 329-335.

Liu, L., & Arnold, M. (2009). Selectivity for glucose, glucose-6-phosphate, and pyruvate in ternary mixtures from the multivariate analysis of near infrared spectra. *Analytical and Bioanalytical Chemistry , 393*, 669-677.

Liu, R., Deng, B., Chen, W., & Xu, K. (2005). Next Step of Non-invasive Glucose Monitor by NIR Technique from the Well Controlled Measuring Condition and Results. *Optical and Quantum Electronics , 37* (13-15), 1305-1317.

Lleti, R., Melendez, E., Ortiz, M., Sarabia, L., & Sanchez, M. (2005). Outliers in partial least squares regression: application to calibration of wine grade with mean infrared data. *Analytica Chimica Acta , 544*, pp. 60-70.

MacKenzie, H., Ashton, H., Spiers, S., Shen, Y., Freeborn, S., Hannigan, J., et al. (1999). Advances in photoacoustic noninvasive glucose testing. *Clinical Chemistry* (45), pp. 1587-1595.

Maikala, R. (2010). Modified Beer's Law - historical perspectives and relevance in near-infrared monitoring of optical properties of human tissue. *International Journal of Industrial Ergonomics , 40* (2), pp. 125-134.

Majumder, M., Roy, P., & Mazumdar, A. (2010). A Generalized Overview of Artificial Neural Network and Genetic Algorithm. In B. Jana, & M. Majumder, *Impact of Climate Change on Natural Resource Management* (pp. 393-416). New York: Springer.

Malik, B., & Cote, G. (2010). Real-time, closed-loop dual wavelength optical polarimetry for glucose monitoring. *Journal of Biomedical Optics* (15(1)), p. 017002.

Martens, H., & Naes, T. (1989). *Multivariate calibration.* Chichester England, New York: Wiley.

Martens, H., Hoy, M., Wise, B., R., B., & Brockhoff, P. (2003). Pre-whitening of data by covariance-weighted pre-preocessing. *Journal of Chemometrics , 17*, pp. 153-165.

Maruo, K., Tsurugi, M., Chin, J., Ota, T., Arimoto, H., & Yamada, Y. (2003). Noninvasive blood glucose assay using a newly developed near-infrared system. *Journal of Selected Topics in Quantum Electronics* , pp. 322-330.

Maruo, K., Tsurugi, M., Tamura, M., & Ozaki, Y. (2003). In vivo noninvasive measurement of blood glucose by near-infrared diffuse-reflectance spectroscopy. *Applied Spectroscopy* (57), pp. 1236-1244.

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition.* New York: Wiley.

Melikyan, H., Danielyan, E., Kim, S., Kim, J., Babajanyan, A., Lee, J., et al. (2011). Non-invasive in vitro sensing of D-glucose in pig blood. *Medical Engineering & Physics* (in press), pp. 1-6.

Mendosa, D. (1995). *On-line diabetes resources: meters*. Retrieved 1 6, 2012, from David Mendosa: a writer about diabetes: http://www.mendosa.com/meters.htm

Mittermayr, C., Tan, H., & Brwon, S. (2001). Robust calibration with respect to background variation. *Applied Spectroscopy* , pp. 827-833.

Moran, G., Jeffrey, K., Thomas, J., & Stevens, J. (2000). A dielectric analysis of liquid and glassy solid glucose/water solutions. *Carbohydrate Research , 328* (4), pp. 573-584.

Moschou, E., Sharma, B., Deo, S., & Daunert, S. (2004). Fluorescence Glucose Detection: Advances Toward the Ideal In Vivo Biosensor. *Journal of Fluorescence , 14* (5), pp. 535-547.

Naes, T., & Martens, H. (1998). Principal component regression in NIR Analysis: viewpoints, backgorund details and selection of components. *Journal of Chemometrics , 2*, pp. 155-167.

Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification.* Chichester, UK: NIR Publications.

National Diabetes Information Clearinghouse. (2008). *Continuous glucose monitoring*. Retrieved from National Institute of Health: http://diabetes.niddk.nih.gov/dm/pubs/glucosemonitor/index.aspx#continue

Optical Science & Technology Center. (2010). *OSTC*. Retrieved 5 24, 2012, from OSTC - Biosciences: http://ostc.physics.uiowa.edu/research/bioscience.shtml

Orfanidis, S. (1996). *Introduction to signal processing.* Englewood Cliffs, N.J.: Prentice Hall.

Ostrom, C. (1990). *Time series analysis: regression techniques* (2nd ed.). Newbury Park, Calif: Sage Publications.

Ozaki, Y., McClure, W., & Christy, A. (2007). *Near-infrared spectroscopy in food science and technology.* Hoboken, NJ: Wiley.

Paul, G., & Bruce, R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta* , pp. 1-17.

Pell, R. (2000). Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems* , pp. 87-104.

Perez, N., Ferre, J., & Boque, R. (2009). Calculation of the reliability of classification in discriminant partial least squares binary classification. *Chemometrics and Intelligent Laboratory Systems , 95*, pp. 122-128.

Perez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics , 112*, pp. 581-592.

Pickup, J. C., Hussain, F., Evans, N., Rolinski, O., & Brich, D. (2005). Fluorescence-based glucose sensors. *Biosensors and Bioelectronics* (20), pp. 2555-2565.

Pierna, J., Jin, L., Daszykowski, M., Wahl, F., & Massart, D. (2003). A methodology to detect oultiers/inliers in prediction with PLS. *Chemometrics and Intelligent Laboratory Systems , 68*, pp. 17-28.

Pierna, J., Wahl, F., de Noord, O., & Massart, D. (2002). Methods for outlier detection in prediction. *Chemometrics and Intelligent Laboratory Systems , 63*, pp. 27-39.

Polevaya, Y., Ermolina, I., Schlesinger, M., & Ginzburg, B. (1999). Time domain dielectric spectroscopy study of human cells. II. Normal and malignant white blood cells. *Biochimica et Biophysica Acta , 1419* (2), pp. 257-271.

Pudil, P., Ferri, F., Novovicova, J., & Kittler, J. (1994). Floating search methods for feature selection with nonmonotonic criterion functions. *Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, *2*, pp. 279-283.

Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating Search Methods in Feature Selection. *Pattern Recognition Letters* (15), pp. 1119-1125.

Rabinovitch, B., March, W., & Adams, R. (1982). Noninvasive glucose monitoring of the aqueous humor of the eye: part I. measurement of very small optical rotations. *Diabetes Care , 5* (3), pp. 254-258.

Raghavachari, R. (2001). *Near-Infrared Applications in Biotechnology.* New York: Marcel Dekker.

Rencher, A. (2002). *Methods of multivariate analysis.* New York: J. Wiley.

Rinnan, A., Berg, F., & Engelsen, S. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry , 28*, pp. 1201-1222.

Robert, C., & Casella, G. (2004). *Monte Carlo Statistical Methods.* New York: Springler.
Robinson, M., Eaton, R., Haaland, D., Koepp, G., Thomas, E., & Robinson, P. (1992). Non-invasive glucose monitoring in diabetic patients: a preliminary evaluation. *Clinical Chemistry , 38*, pp. 1618-1622.

Roger, J., Chauchard, F., & Bellon-Maurel, V. (2003). EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar

content of intact fruits. *Chemometrics and Intelligent Laboratory Systems , 66*, pp. 191-204.

Rumelhart, D., & Mcclelland, J. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations.* Cambridge, MA: MIT Press.

Savitzky, A., & Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry , 36*, pp. 1627-1639.

Scholkopf, B., Sung, K., Burges, C., Firosi, F., Niyogi, P., Poggio, T., et al. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing , 45*, pp. 2758-2765.

Sieg, A., Guy, R., & Delgado-Charro, M. (2004). Noninvasive glucose monitoring by reverse iontophoresis in vivo: application of the internal standard concept. *Clinical Chemistry* (50), pp. 1383-1390.

Siesler, H., Ozaki, Y., Kawata, S., & Heise, H. (2002). *Near-infrared spectroscopy: principles, instruments, applications.* Weinheim: Wiley-VCH.

Sirven, J., Salle, B., Mauchien, P., Lacour, J., Maurice, S., & Manhes, G. (2007). Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods. *Journal of Analytical Atomic Spectrometery , 22*, pp. 1471-1480.

Sivanandam, S., & Deepa, S. (2008). *Introduction to genetic algorithms.* Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Sjoblom, J., Sevensson, O., Josefson, M., Kullberg, H., & Wold, S. (1998). An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemometrics and Intelligent Laboratory , 44*, pp. 229-244.

Skibsted, E., Boelens, H., Westerhuis, J., Witte, D., & Smilde, A. (2004). New indicator for optimal preprocessing and wavelength selection of near-infrared spectra. *Applied Spectroscopy , 58*, pp. 264-271.

Smith, B. (2002). *Quantitative Spectroscopy: Theory and Practice.* San Diego, California, USA: Academic Press.

Suli, E., & Mayers, D. (2003). *An introduction to numerical analysis.* Cambridge, New York: Cambridge University Press .

Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters , 9*, pp. 293-300.

Sverre, G., & Orjan Grottem, M. (2008). *Bioimpedance and bioelectricity basics (2nd edition).* Academic.

Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistics.* Mass, Boston: Allyn and Bacon.

Tao, D., & Adler, A. (2009). In Vivo Blood Characterization From Bioimpedance Spectroscopy of Blood Pooling. *IEEE Transactions on Instrumentation and Measurement , 58* (11), pp. 3831-3838.

Tarumi, T., Amerov, A., Arnold, M., & Small, G. (2009). Design considerations for near-infrared filter photometry: effects of noise sources and selectivity. *Applied Spectroscopy , 63*, 700-708.

The Diabetes Mall. (2010). *Diabetesnet.com*. Retrieved 5 24, 2012, from Future meters & monitors: http://www.diabetesnet.com/diabetes-technology/meters-monitors/future-meters-monitors

Tura, A., Sbrignadello, S., Cianciavicchia, D., Pacini, G., & Ravazzani, P. (2010). A Low Frequency Electromagnetic Sensor for Indirect Measurement of Glucose Concentration: In Vitro Experiments in Different Conductive Solutions. *Sensors , 10* (6), pp. 5346-5358.

van Bemmel, J., Musen, M., & Helder, J. (1997). *Handbook of Medical Informatics.* Germany: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley.

Waynant, R., & Chenault, V. (1998). *Overview of non-invasive optical glucose monitoring techniques*. Retrieved 5 24, 2012, from Overview of non-invasive optical glucose monitoring techniques: http://photonicssociety.org/newsletters/apr98/overview.htm

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). Hoboken, N.J.: Wiley-Interscience.

Wild, S., Roglic, G., Sicree, R., King, H., & Green, A. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2003. *Diabetes Care* , pp. 1047-1053.

Wise, B., & Kowalski, B. (1995). Process Chemometrics. In F. McLennan, & B. Kowalski, *Process Analytical Chemistry.* London: Chapman & Hall.

Wold, S., Antti, H., Lindgren, F., & Ohman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory , 44*, pp. 175-185.

Wold, S., Cheney, J., Kettaneh, N., & McCready, C. (2006). The chemometric analysis of point and dynamic data in pharmaceutical and biotech production (PAT) - some objectives and approaches. *Chemometrics and Intelligent Laboratory Systems , 84*, pp. 159-163.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems , 2*, pp. 37-52.

Wolthuis, R., Tjiang, G., Puppels, G., & Schut, T. (2006). Estimating the influence of experimental parameters on the prediction error of PLS calibration models based on Raman spectra. *Journal of Raman Spectroscopy , 37*, pp. 447-466.

World Health Organization. (2012). *About diabetes*. Retrieved 3 9, 2012, from Daibetes Programme: http://www.who.int/diabetes/action_online/basics/en/index.html

World Health Organization. (2013). *Diabetes* . Retrieved 1 22, 2013, from WHO: http://www.who.int/mediacentre/factsheets/fs312/en/index.html

Yamakoshi, Y., Ogawa, M., Tamakoshi, T., Satoh, M., Nogawa, M., Tanaka, S., et al. (2007). A new non-invasive method for measuring blood glucose using instantaneous differential near infrared spectrophotometry. *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, (pp. 2964-2967). France.

Yeh, S., Hanna, C., & Khalil, O. (2003). Monitoring blood glucose changes in cutaneous tissue by temperature-modulated localized reflectance measurements. *Clinical Chemistry , 49*, pp. 924-934.

Zhang, D., Ortiz, C., Xie, Y., Davisson, V., & Ben-Amotz, D. (2005). Detection of the site of phosphorylation in a peptide using Raman spectroscopy and partial least squares discriminant analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy , 61*, pp. 471-475.

APPENDIX A

A-1. BioSensors

A-2. ClearPath DS-120

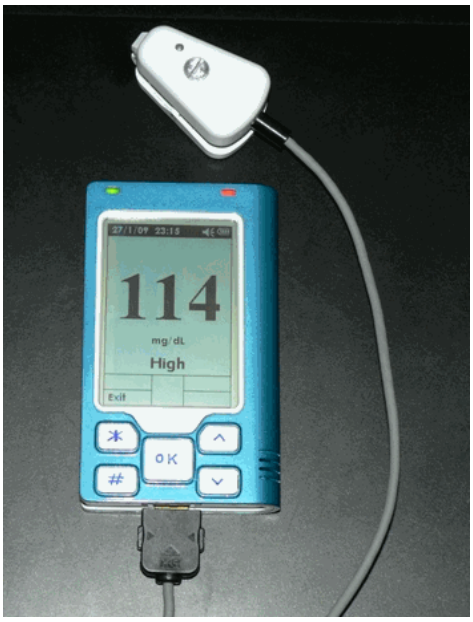
A-3. CNOGA

A-4. C8 MediSensors

A-5. Easy Check

A-6. EyeSense

A-7. Glucoband



A-8. GlucoTrack

A-9. Glove Instruments

A-10. OrSense

A-11. SCOUT DS

APPENDIX B-1

Poster for PolyU internal recruitment and newspaper advertisement

APPENDIX B-2

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

參與研究同意書

研究由新研創之無創性近紅外線血糖儀以單波長方法來量度血糖水平之準確性

本人 _____ 同意參加由鍾慧儀教授負責執行的研究項目。

我理解此研究所獲得的資料可用於未來的研究和學術交流。然而我有權保護自己的隱私,我的個人資料將不能洩漏。

我對所附資料的有關步驟已經得到充分的解釋。我理解可能會出現的風險。我是自願參與這項研究。

我理解我有權在研究過程中提出問題,并在 任何時候決定退出研究而不會受到任何不正常的待遇或責任追究。

參加者姓名 _____.

參加者簽名 _____.

父母姓名或監護人姓名 (如需要) _____.

父母或監護人簽名 (如需要) _____.

研究人員姓名 _____.

研究人員簽字 _____.

日期 _____.

APPENDIX B-3

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

School of Nursing 護理學院

有關資料(成人)

研究由新研創之無創性近紅外線血糖儀以單波長方法來量度血糖水平之準確性

誠邀閣下參加由香港理工大學護理學院鍾慧儀教授負責執行的研究計劃。這項研究目的旨在測試以單波長近紅外線來量度血糖的準確性。

閣下在參與研究前一晚，你需要於晚飯後開始禁止進食十二小時〔禁食期間，你只可飲用清水〕，翌日早上你會來到本診所，我們會為閣下抽取十二毫升空腹血液以作化驗。化驗包括全血像、腎功能測驗、肝臟功能測驗、血脂分析、空腹血糖及糖化血紅蛋白。我們會使用研究之血糖儀掃描閣下十隻手指、手前臂、手腕、口腔頰及耳垂。你會被要請提供性別、年齡及病史之人口統計資料。整個過程約半小時。為表謝意你會獲取你所有化驗報告及其解釋。

這項研究除在抽取血液樣本時有輕微針刺痛外，整個過程不會引起任何不適。閣下享有充分的權利在任何時間退出這項研究，而不會受到任何對閣下不正常的代遇或責任追究。凡有關閣下的資料均會保密。

如果閣下有任何對這項研究的不滿,請隨時與香港理工大學人事倫理委員會秘書親自或寫信聯絡(地址:香港理工大學人力資源辦公室 M1303 室轉交)。

如果閣下想獲得更多有關這項研究的資料,請與鍾慧儀教授，電話27666548聯系。

謝謝閣下有興趣參與這項研究。

鍾慧儀教授
首席研究員

APPENDIX B-3

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

School of Nursing 護理學院

INFORMATION SHEET (for Adult)

The precision of a single wavelength approach in measuring blood glucose level using the newly developed non-invasive near-infrared glucose meter

You are invited to participate on a study conducted by Professor Joanne Wai-yee CHUNG who is currently a Professor of Nursing in the School of Nursing in The Hong Kong Polytechnic University and the Principal Investigator of the study.

The aim of this study is to examine the precision in blood glucose measurement using a single wavelength approach.

The night before study, you are required to fast for twelve hours after your dinner (during the fasting period, you are only allowed to drink plain water). You will attend our clinic the next morning for the study. We will draw 12 ml fasting blood from you for the laboratory measurement. The tests will include Complete Blood Picture, Renal Function Test, Liver Function Test, Lipid Profile, Fasting Blood Glucose and Haemoglobin A1c. We will scan your 10 fingers, forearm, wrist, buccal cavity and ear lobe using our non-invasive glucose meter. You will also be asked to provide demographic data on gender, age and medical history. The whole procedure will take you about 30 minutes. You will receive your laboratory results together with explanation as a token.

Apart from the minimal needle puncture pain during the blood sampling procedure, it should not be any undue discomfort during the whole procedure. You have every right to withdrawn anytime from the study without penalty of any kind. All information related to you will remain confidential.

If you have any complaints about the conduct of this research study, please do not hesitate to contact Mr. Eric Chan, Secretary of the Human Subjects Ethics Sub-Committee of The Hong Kong Polytechnic University in person or in writing (c/o Human Resources Office in Room M1303 of the University).

If you would like more information about this study, please contact Joanne Wai-yee Chung on tel. no. 34003806.

Thank you for your interest in participating in our study.

Prof. Joanne Wai-yee CHUNG
Principal Investigator