# DETERMINATION OF SUGAR IN NON-ALCOHOLIC BEVERAGES USING NEAR INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS

Miss Sureerat Makmuang

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Chemistry
Department of Chemistry
Faculty of Science
Chulalongkorn University
Academic Year 2018

Copyright of Chulalongkorn University

# การตรวจวัดน้ำตาลในเครื่องดื่มที่ไม่มีแอลกอฮอล์โดยใช้สเปกโทรสโกปีอินฟราเรดย่านใกล้ร่วมกับเคโมเมทริกซ์

น.ส.สุรีรัตน์ มากเมือง

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเคมี ภาควิชาเคมี
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

| | |
|---|---|
| Thesis Title | DETERMINATION OF SUGAR IN NON-ALCOHOLIC BEVERAGES USING NEAR INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS |
| By | Miss Sureerat Makmuang |
| Field of Study | Chemistry |
| Thesis Advisor | Assistant Professor KANET WONGRAVEE, Ph.D. |
| Co Advisor | Assistant Professor Prompong Pienpinijtham, Ph.D. |

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

............................................................ Dean of the Faculty of Science
(Associate Professor Polkit Sangvanich, Ph.D.)

THESIS COMMITTEE

............................................................ Chairman
(Associate Professor Vudhichai Parasuk, Ph.D.)
............................................................ Advisor
(Assistant Professor KANET WONGRAVEE, Ph.D.)
............................................................ Co-Advisor
(Assistant Professor Prompong Pienpinijtham, Ph.D.)
............................................................ Examiner
(Assistant Professor Passapol Ngamukot, Ph.D.)
............................................................ Examiner
(Associate Professor VIWAT VCHIRAWONGKWIN, Dr. rer. nat)
............................................................ External Examiner
(Assistant Professor Pitiporn Ritthiruangdej, Ph.D.)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สุรีรัตน์ มากเมือง : การตรวจวัดน้ำตาลในเครื่องดื่มที่ไม่มีแอลกอฮอล์โดยใช้สเปกโทรสโกปีอินฟราเรดย่านใกล้ร่วมกับเคโมเมทริกซ์. ( DETERMINATION OF SUGAR IN NON-ALCOHOLIC BEVERAGES USING NEAR INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS) อ.ที่ปรึกษาหลัก : ผศ. ดร.คเณศ วงษ์ระวี, อ.ที่ปรึกษาร่วม : ผศ. ดร.พร้อมพงศ์ เพียรพินิจธรรม

การตรวจวัดด้วยเทคนิคสเปกโทรสโกปีอินฟราเรดย่านใกล้ร่วมกับเคโมเมทริกซ์ถูกนำมาใช้อย่างแพร่หลายสำหรับการควบคุมคุณภาพของผลิตภัณฑ์อาหาร ในงานวิจัยนี้นำเสนอวิธีการสร้างแบบจำลองที่มีชื่อว่า “แบบจำลองสากล” โดยแบบจำลองดังกล่าวสร้างจากข้อมูลที่ได้โดยการวัดด้วยเทคนิคเปกโทรสโกปีอินฟราเรดย่านใกล้ของสภาวะแรก เพื่อใช้ทำนายค่าตอบสนองจากสภาวะอื่น ๆ (สภาวะที่สอง) ได้อย่างถูกต้อง งานวิจัยนี้ใช้แบบจำลองสากลเพื่อทำนายปริมาณน้ำตาลกลูโคสในเครื่องดื่มที่ไม่มีแอลกอฮอล์ วิธีการสร้างแบบจำลองสากลมี 3 ขั้นตอนหลัก คือ ปรับสเปกตรัม เลือกช่วงที่สำคัญของข้อมูล และดึงข้อมูลที่เป็นส่วนประกอบหลัก ดังนั้น เพื่อเป็นการพิสูจน์แนวคิดของแบบจำลองสากลดังกล่าว ผู้วิจัยจึงได้จำลองเปกตรัมของอินฟราเรดย่านใกล้ที่ถูกรบกวนจากสิ่งอื่น ๆ ที่ไม่เกี่ยวข้องกับระบบขึ้น จากผลการทดลองพบว่าแบบจำลองสากลสามารถทำนายความเข้มข้นของน้ำตาลกลูโคสได้ถูกต้องมากขึ้น 30 เปอร์เซนต์เมื่อเปรียบเทียบกับแบบจำลองปกติ จากนั้นแบบจำลองสากลดังกล่าวถูกนำไปทำนายความเข้มข้นของน้ำตาลกลูโคสในน้ำชา โกโก้ และกาแฟ ผลการคำนวณพบว่าค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนจากการทำนายคือ 0.72 ($r^2$ = 0.9977) 0.99 ($r^2$ = 0.9965) และ 0.54 ($r^2$= 0.9940) ตามลำดับ ดังนั้นอาจกล่าวได้ว่าแบบจำลองสากลนี้สามารถทำนายปริมาณน้ำตาลกลูโคสในเครื่องดื่มที่ไม่มีแอลกอฮล์ได้โดยไม่ต้องสร้างแบบจำลองมาตรฐานใหม่

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สาขาวิชา เคมี

ปีการศึกษา 2561

ลายมือชื่อนิสิต ...............................................

ลายมือชื่อ อ.ที่ปรึกษาหลัก .............................

ลายมือชื่อ อ.ที่ปรึกษาร่วม ..............................

# # 5972085423 : MAJOR CHEMISTRY

KEYWORD:

Sureerat Makmuang : DETERMINATION OF SUGAR IN NON-ALCOHOLIC BEVERAGES USING NEAR INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS. Advisor: Asst. Prof. KANET WONGRAVEE, Ph.D.,Asst. Prof. Prompong Pienpinijtham, Ph.D.

The measurement of Near infrared (NIR) spectroscopy, combined with chemometric techniques, has been widely employed for quality control in food products. This study presents a methodology to optimize the calibration models, called “universal model” of NIR spectra of primary condition (glucose solutions) and maintain the accurate prediction of secondary conditions. For instance, the models were designed for determination of glucose concentration in non-alcoholic drinks. Three stages of methodology including pre-processing, feature selection and main component extraction were applied to spectral data in order to obtain the universal calibration model. The simulated NIR spectra with different noise levels were used to ensure that the model from our methods is able to estimate amount of sugar in any conditions with high accuracy. From the analysis, the universal model improves the prediction for the test set (unseen data) for at least 30 percent compared to the other predictions. Then, it was used to quantify amount of glucose in non-alcoholic beverages (tea, cocoa and coffee in the case). The promising value for root mean square error of prediction (RMSEP) were obtained to be 0.72 ($r^2$ = 0.9972), 0.99 ($r^2$ = 0.9965) and 0.54 ($r^2$ = 0.9940) corresponding to tea, cocoa and coffee system, respectively. Therefore, it might be implied that our universal model approach can be used to estimate glucose concentrations in other non-alcoholic drinks without any requirement of a new calibration model.

| | | |
|---|---|---|
| Field of Study: | Chemistry | Student's Signature ............................... |
| Academic Year: | 2018 | Advisor's Signature .............................. |
| | | Co-advisor's Signature ......................... |

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

# ACKNOWLEDGEMENTS

Firstly, I would like to express my graduate and regards to my handsome and cute Asst. Prof. Dr. Kanet Wongravee for his support and helpful guidance, which help me to fulfill my work through various stages. Thanks for his joke that make me lost my stress. Additionally, thanks for his encouragement that always give me in all my situations not only work, but also in daily life.

I would like to thank my co-advisor Asst. Prof. Dr. Prompong Pienpinijtham who also help me completed my work through continuous good recommendation for improving my work.

I would like to thank all of my relative and friends who willingly shared the moment and their cheerfulness with me

I highly appreciate the valuable contribution of SRU member. Thank you for their kindness, friendliness and comfortable feeling over the time of studying. Moreover, thank for their facilities support (Scientific instrument) help me achieve my research project.

A special thank of mine goes to Science Achievement Scholarship of Thailand (SAST) for financial support.

Finally, I would like to thank my family for all their encouragement and great support to me. I want to thank my parents who always beside me no matter where I am in any situation.

Sureerat Makmuang

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

# TABLE OF CONTENTS

Page

ABSTRACT (THAI) ........ iii

ABSTRACT (ENGLISH) ........ iv

ACKNOWLEDGEMENTS ........ v

TABLE OF CONTENTS ........ vi

LIST OF TABLES ........ viii

LIST OF FIGURES ........ ix

LIST OF STMBOLS AND ABBREVIATIONS ........ xii

CHAPTER I  INTRODUCTION ........ 1

- 1.1 Introduction ........ 1
- 1.2 Objective of this work ........ 8
- 1.3 Scope of this work ........ 8

CHAPTER II THEORETICAL BACKGROUND ........ 9

- 2.1 Near infrared spectroscopy ........ 9
- 2.2 Chemometrics ........ 11
  - 2.2.1 Principal component analysis (PCA) ........ 11
  - 2.2.2 Partial least square (PLS) ........ 13

CHAPTER III EXPERIMENTS ........ 16

- 3. Materials and Methods ........ 16
- 3.1 Chemicals and Materials ........ 16
- 3.2 Sample preparation ........ 16
- 3.3 Spectral acquisition ........ 19
- 3.4 Reference measurement ........ 19
- 3.5 Data simulation ........ 21
- 3.6 Wavelength selection ........ 22
- 3.7 Chemometrics ........ 24

CHAPTER IV  RESULTS AND DISSCUSSION ........ 29

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.1 NIR spectrum acquisition ....................29

4.1.1. Preprocessing techniques ....................30

4.1.2. Variations of the NIR spectra of glucose solution with different concentrations....................31

4.2 Primary calibration set of samples....................35

4.3 Simulated datasets....................38

4.4 Prediction of glucose concentration in non-alcoholic drinks....................47

CHAPTER V CONCLUSIONS ....................54

REFERENCES ....................56

VITA....................lx

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

# LIST OF TABLES

Table 1.1 Literature reviews of determination of sugar in non-alcoholic drink using Near-Infrared spectroscopy combined with chemometrics……………………………...................................... 7

Table 4.1 Details of 6 different RMSE indices……….……………………….. 45

Table 4.2 Details of the calculation models in low and high level of system using partial least square (PLS) on calibration samples to estimate glucose concentration of validation samples. Number of major components extracted by PCA. Significant wavelengths were selected using CSWMPLS. Number of PLS components was selected by leave one-out cross validation (LOOCV). Root mean square error of calibration (RMSE) index, root mean square error of cross validation (RMSECV) and root mean square error of prediction (RMSEP) were used as index of prediction error. $r^2$ represents the correlation of predicted and actual value…………….. 48

Table 4.3 Details of the calculation models in mixture sugar using multivariate data analysis approach (PLS) of calibration sample set, validation sample set. Number of major components extracted by PCA. Number of PLS components was selected by leave one-out cross validation. Root mean square error of calibration (RMSE) index, root mean square error of cross validation (RMSECV) and root mean square error of prediction (RMSEP) were used as index of prediction error………………………………………………………. 53

# LIST OF FIGURES

Figure 1.1 Summarize model generation by partial least square regression (PLS), with N obtains total number of variable (wavelength), $M_{\text{trian}}$ obtains total number of sample from training set, $M_{\text{test}}$ obtains total number of sample from test set……………………………………… 3
Figure 1.2 An overview of the step for universal calibration model to predict unknown sample in another system ………………………………… 4
Figure 2.1 The range of electromagnetic radiation in UV (10 nm to 400 nm), visible (400 to 700 nm), infrared (700 nm to 1 mm) and NIR (700-2500 nm)........................................................................................... 9
Figure 2.2 Vibration transition of diatomic molecule………………………....... 10
Figure 2.3 Near infrared overtone absorptions………………………………….. 11
Figure 3.1 Preparation procedure of stock solution of non-alcoholic drink including (A) tea, (B) cocoa and (C) coffee solution at low level (5g /DI 200 mL) which were further use as solvent for secondary condition. In order to obtain high level, 10g /DI 200 mL of tea, cocoa and coffee was used……..……………………....................... 17
Figure 3.2 Procedure for preparation of glucose solution of non-alcoholic drink (3% w/w in the case) using the stock solution of (A) tea, (B) cocoa and (C) coffee as solvent. The procedure will be repeated for glucose concentrations at 5, 7, 10, 12, 14, 16, and 18 %w/w…........... 19
Figure 3.3 A set up of NIR spectrometer for spectrum acquisition…………….. 20
Figure 3.4 Comparison of glucose contents in water and tea between present glucose concentration and glucose concentration determined from HPLC at food research and testing laboratory (FRTL) Chulalongkorn university…………………………………….......  20
Figure 3.5 Concept idea of NIR simulated calculation………………................ 21

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 3.6 (A) Scheme for explanation of CSMWPLS with n is number of wavelength and w is sub window size including 3, 5 and 7. (B) Optimized region of NIR spectra for prediction of glucose in function of RMSECV on PLS model, applying CSMWPLS……….................. 24

Figure 3.7 Conceptual view of the calculation methodology model using multivariate data analysis (PLS) with conventional way (full spectrum) and global model (using PCA)……………………………. 28

Figure 4.1 NIR spectra of water acquired with different path-length in range of 0.4 mm – 10 mm using integration time of 1 and 32 averaged scans with smoothing window of 1………………………………………. 29

Figure 4.2 Savitzky-Golay smoothing NIR spectra of glucose solutions after performing different preprocessing methods including (A) raw data (B) the second spectral derivative and (C) Standard Normal Variate (SNV). The red and blue line represent the NIR spectrum of the highest (18 %w/w) and the lowest (3%w/w) glucose concentration, respectively………………………………………………………… 32

Figure 4.3 NIR spectrum (A) combination spectrum of pure water, pure glucose and glucose solution (18%w/w) in water following blue, black and red line, respectively. (B) glucose solution in water system (18%w/w) after performed baseline correction using standard normal variate (SNV), (C) variance sample of average NIR spectrum of glucose in water (primary condition)……………………………….................... 34

Figure 4.4 (A) the error of each sample calculated from ypredict – yactual (blue spots) including the average error of all samples (red dot line) and three standard deviation of average error (black dot line) (B) Bar plot between root mean square error of cross validation (RMSECV) (gray bar) and root mean square error of calibration (RMSEC) (pink bar) on NIR spectra of glucose concentration in different repeated batches…………………………………………………………… 36

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 4.5 Loading profiles (PC1 to PC6) of the NIR spectra from batch 1-3 after performing PCA…………………………………………………. 37

Figure 4.6 (A) Simulated NIR spectra and (B) Variance plot of the simulated NIR spectra at different additional noise levels (0-40%)……………... 39

Figure 4.7 (A) RMSEP (black line on the left Y axis) and r2 (blue line on the right Y axis) plot against different additional noise levels, (B) concentration correlation between actual values (X-axis) and predicted values (Y-axis) of glucose concentrations at different levels of noise (%w/w)……………………………………………………. 41

Figure 4.8 RMSEP calculated using different number of inserted samples (1-40 samples). The inset figure shows the relative percent error of RMSEP compared with RMSEP obtained from using single inserted sample........................................................................................... 43

Figure 4.9 (A) Simulated NIR spectra with 9%w/w additional noise. (B) the selected wavelengths using CSWMPLS (window size of 3, 5, and 7). The yellow highlight bands represent the sub-regions which are strongly correlate to the prediction of glucose concentration……….. 44

Figure 4.10 Plots of six different strategies including Model I: RMSEC$_{(1st,\ 1st,\ full,\ 0)}$, Model II: RMSECV$_{(1st,\ 1st\ CV,\ full,\ 0)}$, Model III: RMSEP$_{(1st,\ 2nd,\ full,\ 0)}$, Model IV: RMSEP$_{(1st,\ 2nd,\ pc,\ 0)}$, Model V: RMSEP$_{(1st,\ 2nd,\ pc,\ \omega)}$ and Model VI: RMSEP$_{(2nd,\ 2nd,\ full,\ 0)}$. In this case, our proposed universal model was represented as model V…………………………………... 47

Figure 4.11 A-F show correlation plots between actual value (X-axis) and predicted value (Y-axis) of glucose concentration in high/low level of tea, high/low level of cocoa, high/low level of coffee, respectively, using calibration model A, B and C…………………………………... 51

Figure 4.12 NIR spectra of glucose, fructose, galactose and sucrose solution (8%w/w) in water system following black, pink, blue and green line, respectively………………………………………………………… 52

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

# LIST OF STMBOLS AND ABBREVIATIONS

| | |
|---|---|
| NIR | : near infrared spectroscopy |
| UV | : ultraviolet |
| $\boldsymbol{X}$ | : predictor variables |
| $\boldsymbol{X}_{\text{primary}}$ | : data set from primary condition |
| $\boldsymbol{X}_{\text{secondary}}$ | : data set from secondary condition |
| $\boldsymbol{X}_{\text{obs}}$ | : is observation matrix which is a combination of $\boldsymbol{X}_{\text{primary}}$ and $\boldsymbol{X}_{\text{secondary}}$ |
| $y$ | : response variable |
| $y_{\text{cv}}$ | : estimated glucose concentration by leave one out cross validation |
| $y_{\text{c}}$ | : estimated glucose concentration of the sample in calibrations set |
| $y_{\text{p}}$ | : estimated glucose concentration of the sample in validation set |
| $b$ | : correlation coefficient between data matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$ |
| $\boldsymbol{N}$ | : total number of variable (wavelength) |
| $n$ | : number of selected wavelength |
| $M$ | : total number of sample (concentration) |
| $m$ | : number of sample |
| PLSR | : partial least squares regression |
| PC | : principal component |
| PCA | : principal component analysis |
| DS | : direct standardization |
| PDS | : Piecewise direct standardization |
| MLR | : multiple linear regression |
| LOD | : limit of detection |
| $\mu$ mol/l | : micromole per liter |
| ppm | : parts per million |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

| | |
|---|---|
| HPLC | : high-performance liquid chromatography |
| SD | : standard deviation |
| RSD | : relative standard deviation |
| RMSEC | : root mean square error of calibration |
| RMSECV | : root mean square error of cross validation |
| RMSEP | : root mean square error of calibration |
| $\tilde{\nu}$ | : frequency of vibration |
| $\acute{s}$ | : speed of light |
| $\varepsilon$ | : force constant (5 x $10^5$ dynes/cm) |
| cm | : centimeter |
| $u$ | : mass |
| NIPALS | : Nonlinear Iterative Partial Least Squares |
| $\boldsymbol{P}$ | : loading matrix (of size $A$ x $N$) |
| $\boldsymbol{T}$ | : score matrix (of size $M$ x $A$) |
| $\boldsymbol{P}_{\text{primary}}$ | **:** data of loading of primary condition |
| $\boldsymbol{T}_{\text{primary}}$ | **:** data of score of primary condition |
| $\boldsymbol{P}_{\text{secondary}}$ | **:** data of loading of secondary condition |
| $\boldsymbol{T}_{\text{secondary}}$ | **:** data of score of secondary condition |
| $K$ | : total number of PLS component |
| $k$ | : selected number of PLS |
| $A$ | : PCA component |
| $a$ | : selected number of PCA component |
| $e$ | : residual information in predictor variables ($\boldsymbol{X}$) |
| $f$ | : residual information in response variable ($y$) |
| $\boldsymbol{H}$ | : matrix of weights (of size $N$ x $A$) |
| $q$ | : PLS loading vector |
| $t$ | : PLS score vector |
| DI | : distilled deionized water |
| mL | : milliliter |
| g | : gram |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

e.g. : for example

°C : degree Celsius

%w/w : percent weight by weight

SMA : Sub Miniature Version A

LS : Light Storm

rmp : revolutions per minute

$\mu$m : micrometer

mm : millimeter

nm : nanometer

cm : centimeter

SNV : standard normal variate

mL/min : milliliter per minute

RID : refractive index detection

ASTM : American Society of Testing Materials

FRTL : food research and testing laboratory

$\boldsymbol{c}_{\text{water}}$ : water content (%w/w)

$\boldsymbol{x}_{\text{water}}$ : pure spectrum of water

$\boldsymbol{c}_{\text{glucose}}$ : glucose content (%w/w)

$\boldsymbol{x}_{\text{glucose}}$ : pure spectrum of glucose

$\boldsymbol{c}_{\text{noise}}$ : noise content (%w/w) that was simulated base on distribution

$\boldsymbol{c}_{\text{total}}$ : total content (100 %w/w)

$\boldsymbol{X}_{\text{simulate}}$ : simulated dataset

CSMWPLS : changeable size moving window partial least square

$i$ : spectral elements

$w$ : window size

$\alpha$ : calibration

$\beta$ : validation

$\lambda$ : number of principal component

$\omega$ : selected wavelength

$^{T}$ : Transpose

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Model I : the calibration model was built from a dataset from primary condition and used to estimate the responses of a dataset from primary condition.

Model II : the calibration model was built from a dataset from primary condition and used to estimate the responses of a dataset from primary condition using Leave-One-Out cross validation.

Model III : the calibration model was built from a dataset from primary condition and used to estimate the responses of a dataset from secondary conditions (without extract any main components).

Model IV : the calibration model was built from the main components extracted from a dataset from primary condition and then it was used to estimate the responses of a dataset from secondary condition.

Model V : the calibration model was built from the main components extracted from a dataset from primary condition with only selected wavelengths and used to estimate the responses from secondary condition. This represents our universal model.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Model VI : the calibration model was built from a dataset from secondary condition and directly used to estimate the responses of the secondary condition. This basic idea was represented as a conventional model.

Model A : the calibration model was built from a dataset from drinks (tea, cocoa, and coffee system) and directly used to estimate the glucose concentration in the drinks.

| | |
|---|---|
| Model B | : the calibration model was built from a dataset of glucose solution (primary condition). Then, it was used to estimate the glucose concentration in the drinks (tea, cocoa, and coffee system) without extracting any main components. |
| Model C | : the calibration model was built from the main components extracted from a dataset of glucose solution (primary condition) with only selected wavelengths and then it was used to estimate the glucose concentration in the drinks (tea, cocoa, and coffee system). This represents |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

# CHAPTER I
# INTRODUCTION

## 1.1 Introduction

In recent years, near infrared spectroscopy (NIR) has broad prospects for the quality measurement system in various fields such as agriculture, pharmaceuticals and food industry. It displays to be reliable one of the promising nondestructive techniques. The attraction of NIR lie in its advantages over other analytical techniques such as no requirement for sample pretreatment, very fast and easy to implement[1]. NIR spectroscopy is among the vibrational techniques that measure wavelengths from 800 nm to 2500 nm[2]. The band regions of NIR are based on molecular overtones and combination vibrations of C-H, O-H and N-H which are the primary functional groups of organic molecules[3]. It enables qualitative and quantitative assessment via spectral information and multivariate calibration models especially for complex chemicals in food and drinks such as protein[4], carbohydrate[5] , sugar[6] and lipids[7]. There are evident that NIR has simplified and helped to quantify a variety of element in food such as moisture, protein, wet gluten and fat[8]. With the advance of technology, the prices of commercial NIR instruments in the current market is relatively cheap as there are the development of micro-electromechanical system (MEM) technology. Therefore, it offers the NIR detection using such a small sensor chip. It is foreseeable that the current NIR spectrometer can produce a large amount of data. However, due to overtone and combination bands of NIR spectrum contain very complex and many overlapping signals. Consequently, the distinguished and characteristic peaks of an analyze are difficult to identify by conventional band assignment methods and spectral analysis method[9]. More sophisticated approaches with mathematical and statistical tools are required to extract analytical information from the corresponding NIR spectra.

Chemometrics is an application of mathematical and statistical methods to data that is underlying chemical in nature to obtain relevant information[8]. Multivariate data analysis on visualization, calibration and classification are among the most important and widely used in chemometrics methods. In this study, only multivariate calibration is discussed. The multivariate calibration investigates the relationship between two set of variables which usually defines as "predictor" variables (dependent '***X***' block) and "response" variables (independent '***y***' block) which can be in form of vector or matrix. The predictor variable is an independent variable that is being manipulated in an experiment while the response variable is the effects from whose variation is being studied. Examples of the predictor variable including physical-chemical measurements are wavelengths in the case of NIR spectra. The responses are properties of interest such as concentrations in the case. Multivariate calibration always involves two major stages: (1) Modelling where a calibration model is constructed using samples with known properties as "training set" and (2) Prediction which involves the prediction of unknown samples as "test set" based on the built relationship information obtained from the first stage. Overall model was shown in Figure 1.1. From figure 1.1, the training set consists NIR spectra of $N$ variable and $M$ sample and a response variable vector ($y$). A coefficient vector ($b$) is calculated based on the maximum correlation coefficient between data matrix $\boldsymbol{X}$ and response vector $y$. Then, these coefficients were used to predict response of the external test set ($\boldsymbol{y}_{\text{predict}}$).

To build a good calibration curve from a single defined peak from NIR spectra might not sufficient. Chemometrics has most often been used to extract specific features to specific chemical components in the NIR spectra for effective interpretation. A main part of chemometics is multivariate data analysis, which is pivotal for quantitative and qualitative assay based on NIR spectra. Multivariate data analysis techniques such as principal component analysis (PCA)[10] and partial least squares (PLS)[11] are used to mathematically predict the pure component spectra and pure component concentration profiles from the set of NIR spectra. Mostly, PLS have been frequently used to build the calibration model from NIR spectra[12]. A calibration model is a mathematical relationship between the acquired spectra and factor of interest and generated calibration model can be used to predict the response of the unknown samples. By conventional way, each calibration model is required for each system. To obtain an appropriate quantification, a new calibration model must be constructed for any new system. It is inevitable, in case of many systems need to build a new calibration model

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

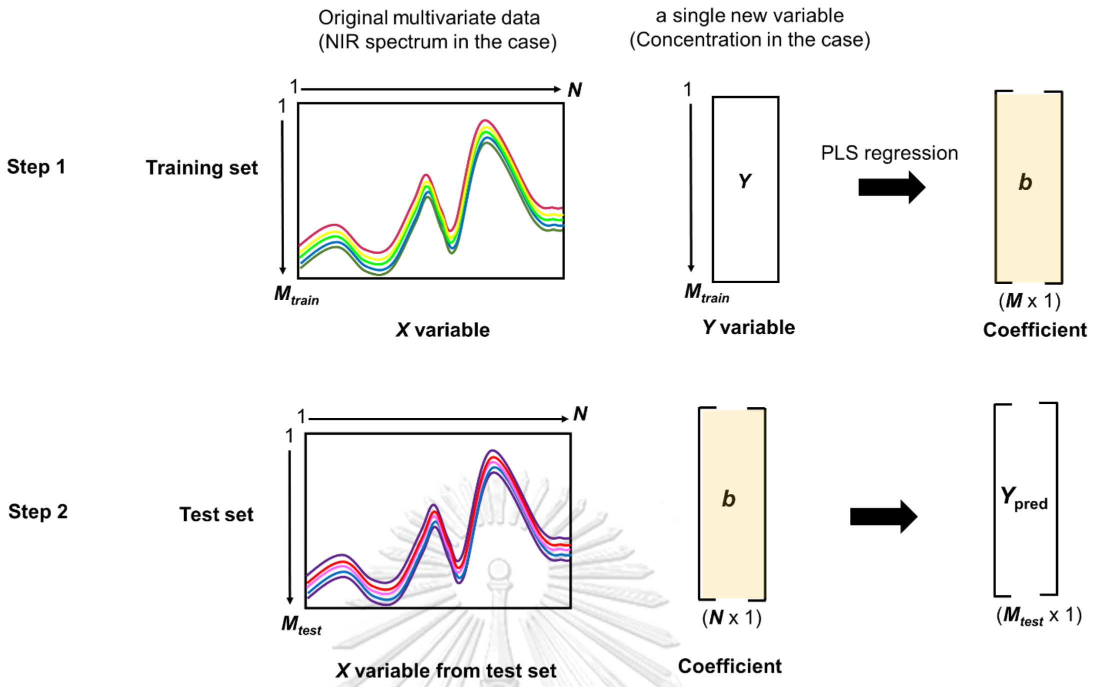


**Figure 1. 1 Summarize model generation by partial least square regression (PLS), with $N$ obtains total number of variable (wavelength), $M_{\mathrm{trian}}$ obtains total number of sample from training set, $M_{\mathrm{test}}$ obtains total number of sample from test set**

all times rendering a multivariate calibration model invalid. Hence, the limitation of prediction capabilities in multivariate spectral calibration model was occurred because time consuming, costly, involving selection and preparation of a large numbers of calibration sample sets. In reality, it is not possible to obtain all calibrations due to the limitation of laboratory and the measurement conditions. Therefore, the process of searching for the chemometric approaches to interpret and improve the predictive ability on future samples in different system is called “Calibration maintenance”[13]. Model maintenance can be roughly defined as the ongoing upkeep of calibration model of primary condition to maintain their predictive abilities of secondary conditions. The goal of model maintenance is to preserve or to improve models over time and changing conditions with the least amount of effort, cost and it should be done automatically. In practical application, it is preferable to produce the universal calibration model that can be used to predict another system without any requirement of set up new calibration curve for quantitative as shown in Figure 1.2.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

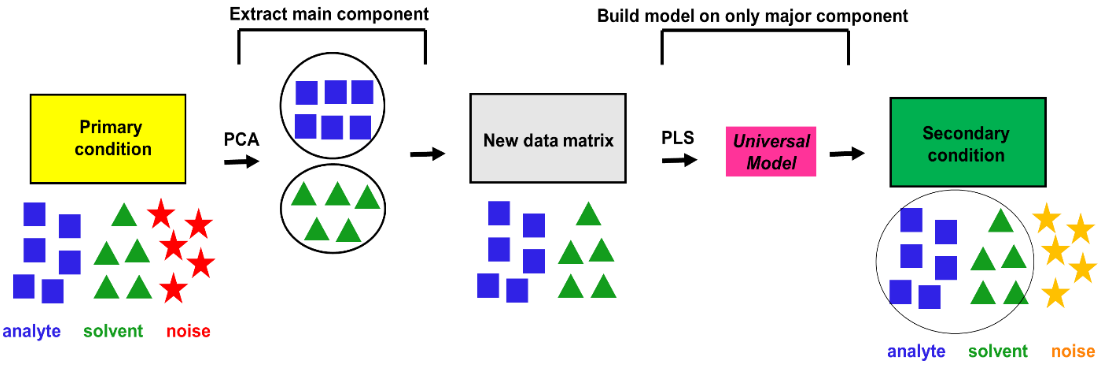


**Figure 1.2 an overview of the step for universal calibration model to predict unknown sample in another system**

From Figure 1.2, it is a brief summary of our model maintenance idea. Initially, primary condition consists of 3 components including analyte (glucose), solvent (water) and noise (other chemical content). Subsequently, principal component analysis was performed on the data of primary condition to extract main components (analyte + solvent) that will later represent in form of new data matrix. After that, the calculation model was constructed from the matrix as a universal model by using PLS. The generated model involves only the relation of the extracted main components, in the case, the noise might not affect the model. Therefore, this model can be used to quantify amount of analyze in secondary conditions with high accuracy and precision.

In the reality, there are many maintenance methods such as simple univariate slope and bias correction method[14] that is one of the most widely methods for correcting predictive value to standardize the calibration models. Therefore, the calibration developed on primary condition has an ability to predict the response in secondary condition. Direct standardization (DS) is a common calibration maintenance that uses the correlation coefficients between matrices from primary and secondary condition to standardize the calibration model[15-16]. In the extended standardization called “Piecewise direct standardization (PDS)” were developed[17]. From the method, the data is segmented into small sub-windows and the correlation coefficients are determined using PLS rather than the simple multiple linear

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

regression (MLR) in DS[18]. However, the disadvantages of these methods are that utilizes all the variations in the data must be utilizes therefore the variation of primary condition must be standardized together with all other conditions. The prediction cannot be accurate when the model is used to predict a sample from an unknown condition. The approach does not produce "universal" model as demonstrated previously.

## Glucose in non-alcoholic drink

Glucose is an aldolic monosaccharide that is essential in the processes of photosynthesis and respiration, serving as an energy storage and metabolic fuel in most organisms[19]. Moreover, glucose plays an important role in our daily life in form of food and beverage. Non-alcoholic beverages are the soft drinks heavily consumed mainly because of their nutritional values and companies promote and market them everywhere. Naturally, glucose is the major content in mostly soft drink. They provide energy for the body and also in the physiological processes within the human body. To receive of glucose may lead to excessive energy intake, increasing the risk of overweight and obesity[20]. For this reason, determination of glucose content in non-alcoholic beverages are important. In last decade, various techniques for the determination of glucose have been published. Three standard techniques were used to determine amount of pure sugar and in mixing sugar solution including density measurement, refractive index measurement and enzymatic assay[21]. In 1999, Harms *et al*. reported a new method for determination of glucose in soft drinks base on the glucose oxidase-catalyzed oxidation, resulting the limit of detection (LOD) is 10 μmol/l (1.8 ppm)[22]. Although this method is specific, rapid and reproducible, but they require single determination for each compound, which is time consuming and expensive. Meanwhile, several methods to quantify amount of glucose have been developed as well. High-performance liquid chromatography (HPLC) is a standard method used for analyzing glucose in non-alcoholic beverages. In 1992, Akiyama *et al*. developed column packing material and applied to the separation of many kinds of sugar to determine amount of sugar (glucose, sucrose) in soft drinks. This method can be used to predict amount of sugar of approximately 4.2 g per 100 mL in drink[23]. However, they require tedious sample preparation and a relatively long analysis time

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

for each analysis. In contrast, vibrational methods are non-destructive, easy to use, rapid, and do not require sample preparation. Near infrared spectrometry (NIR) are novel and useful alternative to the classical methods mentioned above. Rambra and Guardia reported the method for the direct determination of sugar in fruit juice samples[12]. This method base on the partial least square (PLS) calculation on the first derivative near infrared (NIR) spectra. The limit of detection values are in the range of 0.2 g/100 ml total sugar and 0.2 g/100 ml for glucose. In 2009, Xie *et al.* used near-infrared (NIR) spectroscopy to detect and quantify glucose, fructose and sucrose in bayberry juice[24]. For the result, root square error of cross validation in range of 0.1-0.5, it can be noticed that this method provided an accurate and precise way for determination of glucose, fructose and sucrose in real samples. However, from literature reviews, it can be seen that the determination of glucose using NIR spectroscopy reveals the accuracy of the evaluating models which were not very perfect compared to HPLC method. Nevertheless, NIR spectroscopy is preferable by reasons of their ability to dramatically reduce consuming time and cost of monitoring without any chemical treatment.

In this work, a new alternative method for calibration maintenance was proposed. The idea involves 2 major steps. Firstly, the major components were extracted from the calibration of primary condition (glucose solutions). The calibration model for prediction was built using only major components for prediction of glucose in secondary conditions. Secondly, the optimization of the calibration model including number of principal components, number of PLS components and effective wavelength regions were performed. This proposed method was totally automatic, therefore, it can be applied in other systems. In this study, the simulated datasets with different added noise levels were generated in order to prove our proposed concept. For practical system, quantification amount of glucose in non-alcoholic beverages (tea, cocoa and coffee) were chosen to demonstrate this particular application of the idea because it is a simple system and it benefits in many aspects such as nutritional labeling, detection of adulteration, food quality and economics. This protocol can be used in any secondary condition contain water and glucose as major components without any requirement of set up new calibration curve.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

**Table 1.1 Literature reviews of determination of sugar in non-alcoholic drink using Near-Infrared spectroscopy combined with chemometrics**

| Year | Journal | System | Chemometrics | Detection limit | Accuracy | Ref |
|---|---|---|---|---|---|---|
| 1981 | Journal of food of science | ▪ glucose,fructose and sucrose<br>▪ dried apple tissue | MLR | 20.03 %w/w | Predicted error = 4.6<br>SD = 0.90 | 25 |
| 1984 | Journal of food of science | ▪ glucose,fructose and sucrose<br>▪ fruit juice sample | PLSR | 2.22-14.90 g/100ml | | 26 |
| 1997 | Analytica Chimica Acta | ▪ total sugar, sugars, glucose, sucrose,fructose<br>▪ fruit juice sample | PLSR | 0.2 g/mol | RSD = 0.4-2.3% | 12 |
| 2006 | Journal agriculture and food chemistry | ▪ glucose, fructose and sucrose<br>▪ apple juice | PLSR | 0.059 g/100g | RMSEP = 0.201<br>RMSEC = 0.275 | 27 |
| 2009 | Food of chemistry | ▪ glucose,fructose and sucrose<br>▪ bayberry juice | PLSR | 2.10 g/100g | RMSEP = 0.093<br>RMSEC = 0.0826 | 24 |
| 2015 | Journal of near infrared spectro scopy | ▪ glucose,fructose and sucrose<br>▪ roasted green tea | PLSR | 8.00 g/100g | RMSEP = 0.408<br>RMSEC = 0.313 | 28 |

Note:

| | |
|---|---|
| MLR | : Multiple Linear Regression |
| PLSR | : Partial Least Squares Regression |
| RMSEP | : Root Mean Square Error of Prediction |
| RMSEC | : Root Mean Square Error of Calibration |
| SD | : Standard Deviation |
| RSD | : Relative Standard Deviation |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

However, glucose in alcoholic beverages was neglected in this study. Since molecular structures of alcohol ($C_2H_5OH$) contains only ethyl group ($C_2H_5$-) and hydroxyl group (-OH) which are similar to the main functional groups of sugar, therefore, the overtone patterns of alcohol will be strongly affected to the overtones of sugar. Moreover, the ethyl alcohol is easily volatile that the contents cannot be controlled during the detection. Furthermore, most of alcoholic drinks contain less or without sugar which is not suitable for use in sugar detection.

## 1.2 Objective of this work

To develop a calculation procedure based on chemometrics for determination of glucose concentration in non-alcoholic beverages using Near Infrared Spectroscopy

## 1.3 Scope of this work

This study involves the development of procedure based on chemometrics to perform the universal calibration model. The model was built from primary conditions (glucose in water solution). The developed procedure was performed in order to use the model (from primary condition) to determine glucose concentration in secondary conditions (tea, cocoa and coffee in the case). The limitation and performance of developed procedure was evaluated by using the simulated dataset (NIR spectra) with different added noise level.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

# CHAPTER II

# THEORETICAL BACKGROUND

## 2.1 Near infrared spectroscopy

Over the past 30 years, on/in-line near infrared NIR spectroscopy has been developed to be one of the most efficient and advanced technique for controlling and estimation of quality assessment not only in the food processing but also gain wide acceptance in pharmaceutical industry, biotechnology, plastics and textiles[29]. NIR spectroscopy is a vibrational spectroscopic technique among the infrared light spectrum with close to visible region that can be expressed in range of 750 nm and 2500 nm as shown in Figure 2.1

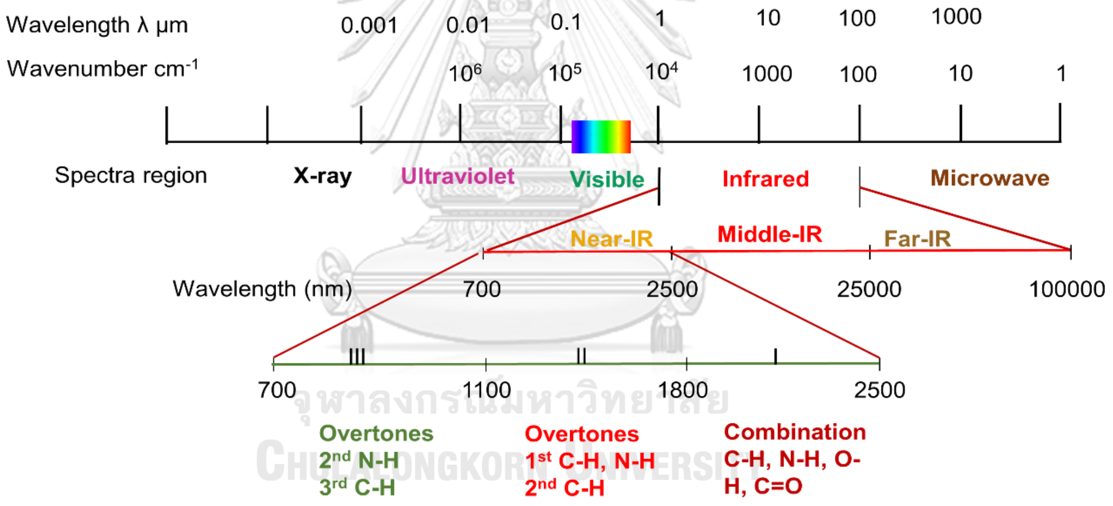


**Figure 2. 1 the range of electromagnetic radiation in UV (10 nm to 400 nm), visible (400 to 700 nm), infrared (700 nm to 1 mm) and NIR (700 -2500 nm)**

NIR spectroscopic method is based on molecular overtone and combination vibrations of C-H, O-H and N-H. Combination bands originate by concurrently interaction between two or more vibrations[30-31]. Generally, even a normal mode of vibrational following to internal atomic motions in which all atoms move in phase with same frequently but different amplitude. Moreover, these normal vibration transition was called overtone. According to the selection rules of quantum mechanics

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

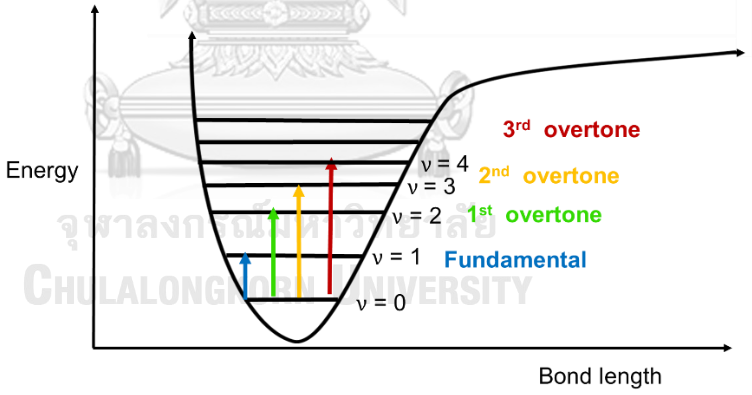
mention normal transition are prohibited cause molar absorptivity in the NIR region is very small[30]. One of the rules that govern the basics idea of vibrational spectroscopy is Hooke’s law. Hooke’s law states that, for two body harmonic oscillators, the frequency of vibration is

$$\tilde{\nu} \text{ (in cm}^{-1}\text{)} \quad = \quad \frac{1}{2\pi\acute{s}}\sqrt{\frac{\varepsilon(u_1+u_2)}{u_1 u_2}}$$

Where, $\acute{s}$ = speed of light, $\boldsymbol{\varepsilon}$ = force constant (5 x $10^5$ dynes/cm)

$u_1$ and $u_2$ is mass of molecule 1 and molecule 2, respectively.

Normally, fundamental vibration for diatomic molecules can be calculated by Hooke’s law. To make it easy to understand, the simple example was shown in Figure 2.2.



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 2. 2 vibration transition of diatomic molecule**

Transition from ground ($v$ = 0) to the first excited state, namely fundamental bands which absorbs strongly light in IR render to high intense band. Transition from the ground state to the second exited state with absorption of NIR that perform weak bands was called 1st overtone in NIR. Transition from ground state to the third exited state with the absorbance of NIR cause to weak band, namely 2nd overtone. In a similar way, transition from the ground state to fourth and fifth exited state with absorbance of NIR will be provided 3rd and 4th overtone, respectively. Additionally, the near infrared absorption region that correspond to vibrational transition (mentioned above) was shown in Figure 2.3

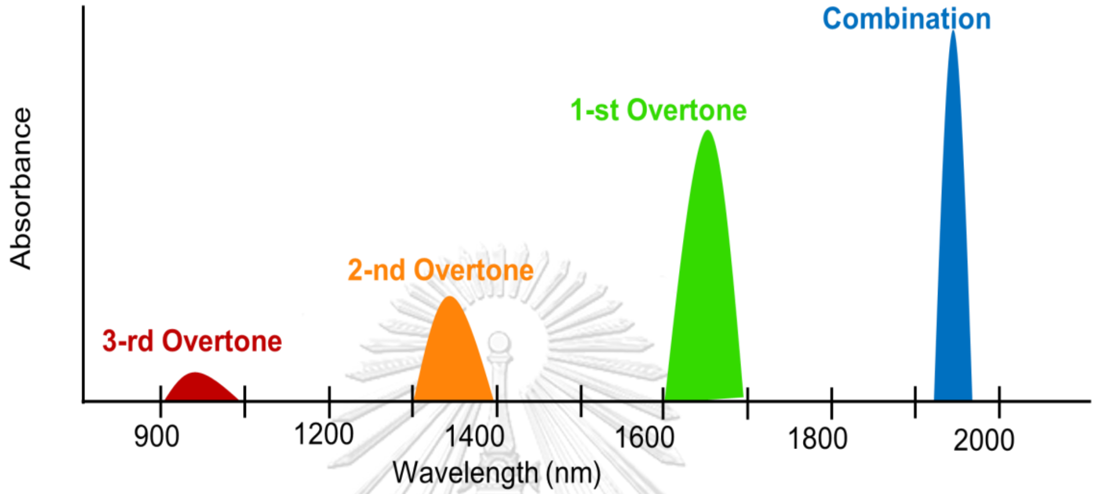


**Figure 2. 3 Near infrared overtone absorptions**

Obviously, weakly absorbed bands occur in the NIR regions due to the overtone and combination bands. As a result, it difficult for use of NIR spectral information for analytical purpose. Therefore, mathematical and statistical method is usually combined with NIR spectra for extracting as much relevant information as possible from analytical data[1]. Chemometric is one of the methods in order to extract the necessary information for further analysis.

CHULALONGKORN UNIVERSITY

## 2.2 Chemometrics

Chemometrics is an application of mathematical and statistical methods to extract only the essential component from NIR spectra comprising complicated overlapping absorption bands. Multivariate data analysis on visualization, calibration and classification are among the most important and widely used in chemometrics methods.

### 2.2.1 Principal component analysis (PCA)

Principal component analysis (PCA) is one of tool from multivariate statistics that help to drastically reduce dimensionality in a large dataset, while that most of the essential information is preserved[32]. Basically, PCA was used to extract the main component from data matrix base on two principal idea, including the number of significant PCs which ideally equal to the number of significant component (such there are three components in the mixture, then only three PCs was expected), the other one is characterization of each PC by loadings and scores.

NIPALS (Nonlinear Iterative Partial Least Squares) is a common, iterative algorithm often used for PCA[33]. Briefly, it extracts components one at a time, and can be stopped after the desired number of PCs has been obtained. The steps are as follows:

Initialization

1. Originate a data matrix $\boldsymbol{X}$ which is used for PCA.

New Principal Component

2. Take a column of this matrix (often the column with greatest sum of squares) as the first guess of the scores first principal component; called ${}^{\text{initial}}\hat{t}$.

Iteration for each principal component

3. Calculate $$\hat{\boldsymbol{p}}_{\text{unnorm}} = \frac{\hat{\boldsymbol{t}}^{\mathrm{T}}{}_{\boldsymbol{initial}} \cdot \boldsymbol{X}}{\sum \hat{\boldsymbol{t}}^{\boldsymbol{2}}}$$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4. Normalize the guess of the loading, so

$$\hat{\boldsymbol{p}} = \frac{\hat{\boldsymbol{p}}_{unnorm}}{\sqrt{\sum \hat{\boldsymbol{p}}_{unnorm}}}$$

5. Now calculate a new guess of the score:

$$\hat{\boldsymbol{t}}_{new} = \boldsymbol{X} \,.\, \hat{\boldsymbol{p}}^{\mathrm{T}}$$

Check for Convergence

6. Check if this new guess differs from the first guess; a simple approach is to look at the size of the sum of square difference in the old and new scores, i.e. $\sum(\hat{\boldsymbol{t}}_{initial} - \hat{\boldsymbol{t}}_{new})^2$. If this is small, the PC has been extracted, set the PC scores ($\boldsymbol{t}$) and loading ($\boldsymbol{p}$) for the current PC to $\hat{\boldsymbol{t}}$ and $\hat{\boldsymbol{p}}$. Otherwise, return to step 3, substituting the initial scores by the new scores.

Compute the Component and Calculate Residuals

7. Subtract the effect of the new PC from the data matrix to obtain a residual data matrix:

$$\boldsymbol{X}_{resid} = \boldsymbol{X} - \boldsymbol{t}\,.\,\boldsymbol{p}$$

Further PCs

8. If it desires to compute further PCs, substitute the residual data matrix for $\boldsymbol{X}$ and go to step 2.

2.2.2 Partial least square (PLS)

In order to construct a calibration model, partial least square (PLS) is one of the most popularly used multivariate calibration methods. Its purpose is to predict a dependent variable, $\boldsymbol{y}$ (of size $M$ x 1 where $M$ is the number of samples), from a matrix of independent variables or predictors, $\boldsymbol{X}$ (of size $M$ x $N$ where $N$ is number of wavelengths), by projecting $\boldsymbol{X}$ and $y$ to the latent subspaces that maximise the covariance between them[34]. This criterion combines high variance of $\boldsymbol{X}$ and high correlation with the interesting property of $y$. According simple structure of this latent variable (LV) model, $\boldsymbol{T}$ is a score matrix obtaining $K$ LVs, $K \leq N$ (of size $M$ x $K$); $\boldsymbol{P}$ is a loading matrix (of size $K$ x $N$) and $q$ (of size $K$ x 1) are matrices of coefficients that relate $\boldsymbol{T}$ to predictor (wavelength) and predicted variable (sample), respectively; $e$ and $f$ represent the residual information in $\boldsymbol{X}$ and $\boldsymbol{y}$ after $K$ LVs, respectively[34].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$\boldsymbol{X} = \boldsymbol{T}\,.\,\boldsymbol{P} + e \tag{1}$$

$$\boldsymbol{y} = \boldsymbol{T}\,.\,\boldsymbol{q} + f \tag{2}$$

In order to estimate $\boldsymbol{T}$ value, the general form was shown in equation (3), where $\boldsymbol{H}$ is a matrix of weights (of size $N$ x $K$), usually estimated using the NIPALS algorithm.

$$\boldsymbol{T} = \boldsymbol{X}\,.\,\boldsymbol{H} \tag{3}$$

Subsequently, $\boldsymbol{T}$ in equation (2) was substituted by equation (3) leads to the simple equation for prediction of y (eq. 4), (where y corresponds to the matrix of predicted variables (each column of which follows to a different number of LVs) and $b$ (of size $N$ x 1) obtain the matrix of estimated regression coefficients.

$$y = \boldsymbol{X} \, . \, b \tag{4}$$

All parameters can be calculated following this step:

Initialization

1. To obtain matrix $\boldsymbol{X}$ which is used for PLS.
2. Take the concentration vector $\boldsymbol{y}$ and preprocess it to give the vector $\boldsymbol{c}$ which is used for PLS. Note that if data matrix $\boldsymbol{X}$ is centred down the columns, the concentration vector must also be centred. Generally, centring is the only from of preprocessing useful for PLS. Start with an estimate of $\hat{\boldsymbol{c}}$ that is vector of 0s (equal to the mean concentration if the vector is already centred).

New PLS component

3. Calculate the vector

$$\boldsymbol{H} = \boldsymbol{X}^{\mathrm{T}} . \, c$$

4. Calculate the score, which are simply given by

$$\boldsymbol{T} = \frac{\boldsymbol{X} \, . \, \boldsymbol{H}}{\sqrt{\sum \boldsymbol{H}^2}}$$

5. Calculate the $\boldsymbol{X}$ loadings by

$$\boldsymbol{P} = \frac{\boldsymbol{T}^{\mathrm{T}} . \, \boldsymbol{X}}{\sum \boldsymbol{T}^2}$$

6. Calculate the c loading (a scalar) by

$$\boldsymbol{q} = \frac{\boldsymbol{c}^{\mathrm{T}} \, . \, \boldsymbol{T}}{\sum \boldsymbol{T}^2}$$

Compute the component and calculate residuals

7. Subtract the effect of the new PLS component from the data matrix to get a residual data matrix:

$$\boldsymbol{X}_{\mathrm{resid}} = \boldsymbol{X} - \boldsymbol{T} \, . \, \boldsymbol{P}$$

8. Determine the new concentration estimate by

$$\hat{\boldsymbol{c}}_{\mathrm{new}} = \hat{\boldsymbol{c}}_{\mathrm{initial}} + \boldsymbol{T} \, . \, \boldsymbol{P}$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

and sum the contribution of all component calculated to give an estimated $\hat{c}$. Note that the initial concentration estimate is 0 (or the mean) before the first component has been computed. Calculate

$$\boldsymbol{c}_{\text{resid}} = \boldsymbol{c}_{\text{true}} - \hat{c}_{\text{new}}$$

where $\boldsymbol{c}_{\text{true}}$ is, like all values of $\boldsymbol{c}$, after the data have been preprocessed (such as centring).

Further PLS Components

9. If further components are required, replace both X and $\boldsymbol{c}$ by the residuals and return to step 3.

   Note that in the implementation used in this text the PLS loading are neither normalize nor orthogonal.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

# CHAPTER III
# EXPERIMENTS

## 3. Materials and Methods

### 3.1 Chemicals and Materials

Analytical grade of D (+) – glucose was purchased from Ajex Finechem. Ingredients for the preparation of non-alcoholic drinks including tea, cocoa and coffee were bought from local supermarkets (Tesco lotus at Chamchuri Square, Bangkok, Thailnd). The dried tea leaves were purchased from Three Horses Tea Co.,Ltd., while Dutch cocoa powder were bought from Pongjit Company Limited. Instant Coffee Mixed with Finely Ground Roasted and Coffee were purchased from Quality Coffee Products Ltd. All of them was used without any further pretreatments. In this work, glucose solutions and non-alcoholic drink were prepared by using distilled deionized water (DI). All glassware was cleaned up with detergent followed by DI water for several times.

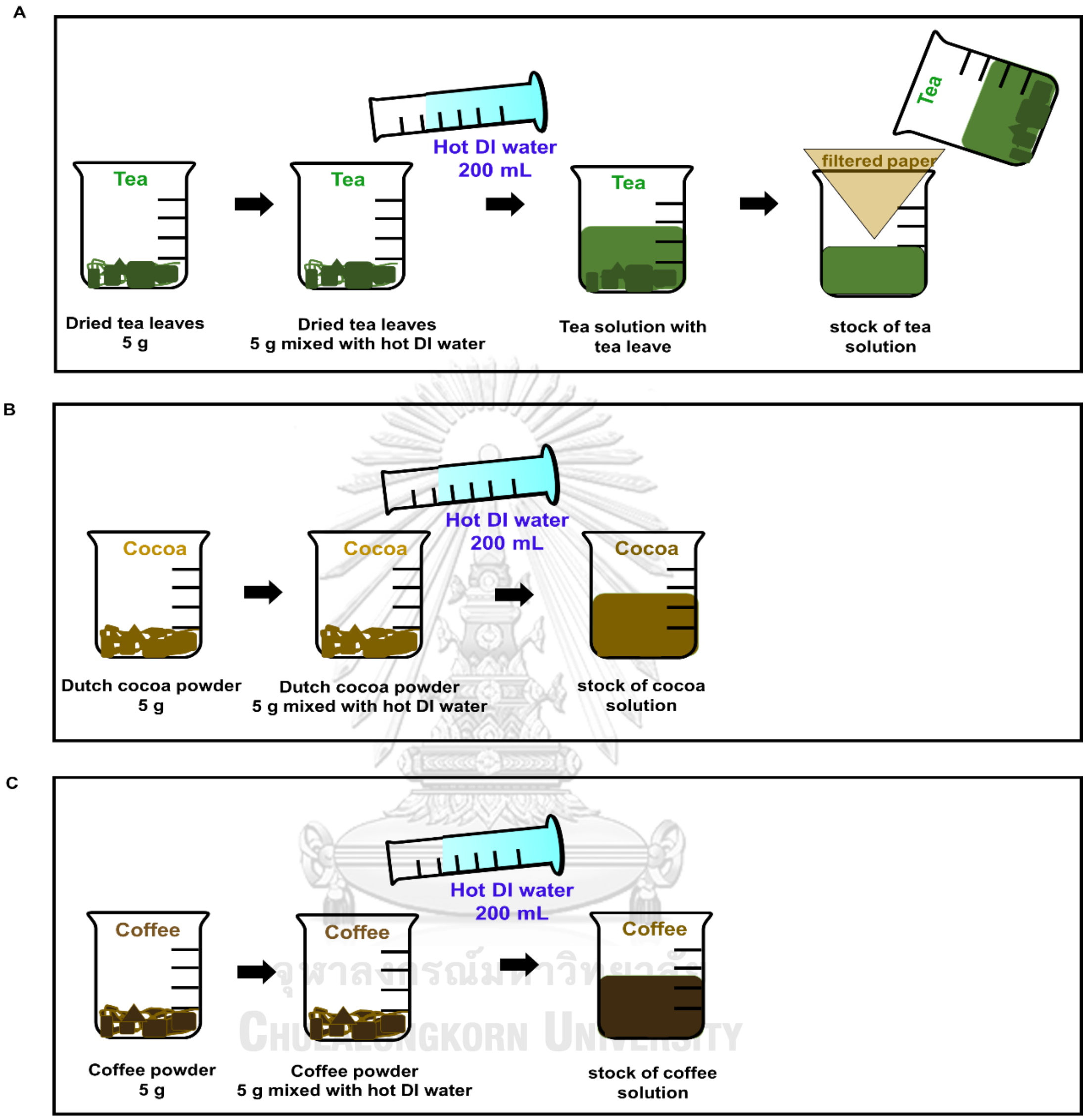
### 3.2 Sample preparation

For preparation solution of non-alcoholic drinks, the dried tea leaves were measured for 5 g (low) and 10 g (high) which was incubated in the 200 mL of hot DI water for 5 minutes. Then the solution was filtered to separate tea leaves in order to obtain the stock of tea solution. In case of soluble ingredients (e.g. cocoa and coffee), the stock solutions were prepared using 5 g (low) and 10 g (high) of the cocoa powder were dissolved in a 200 mL of hot DI water and were stirred until all powders were dissolved. This preparation protocol was repeatedly performed using the instant coffee. From this step, the stock solutions with high and low level of tea, cocoa and coffee were successfully prepared and were undisturbedly left until the temperature of the solution were cooled to the room temperature (25˚C)

Subsequently, the glucose solutions were prepared using DI water as primary condition and the stock of non-alcoholic drink solution including tea, cocoa and coffee as secondary conditions. A calibration set of samples was prepared with

CHULALONGKORN UNIVERSITY

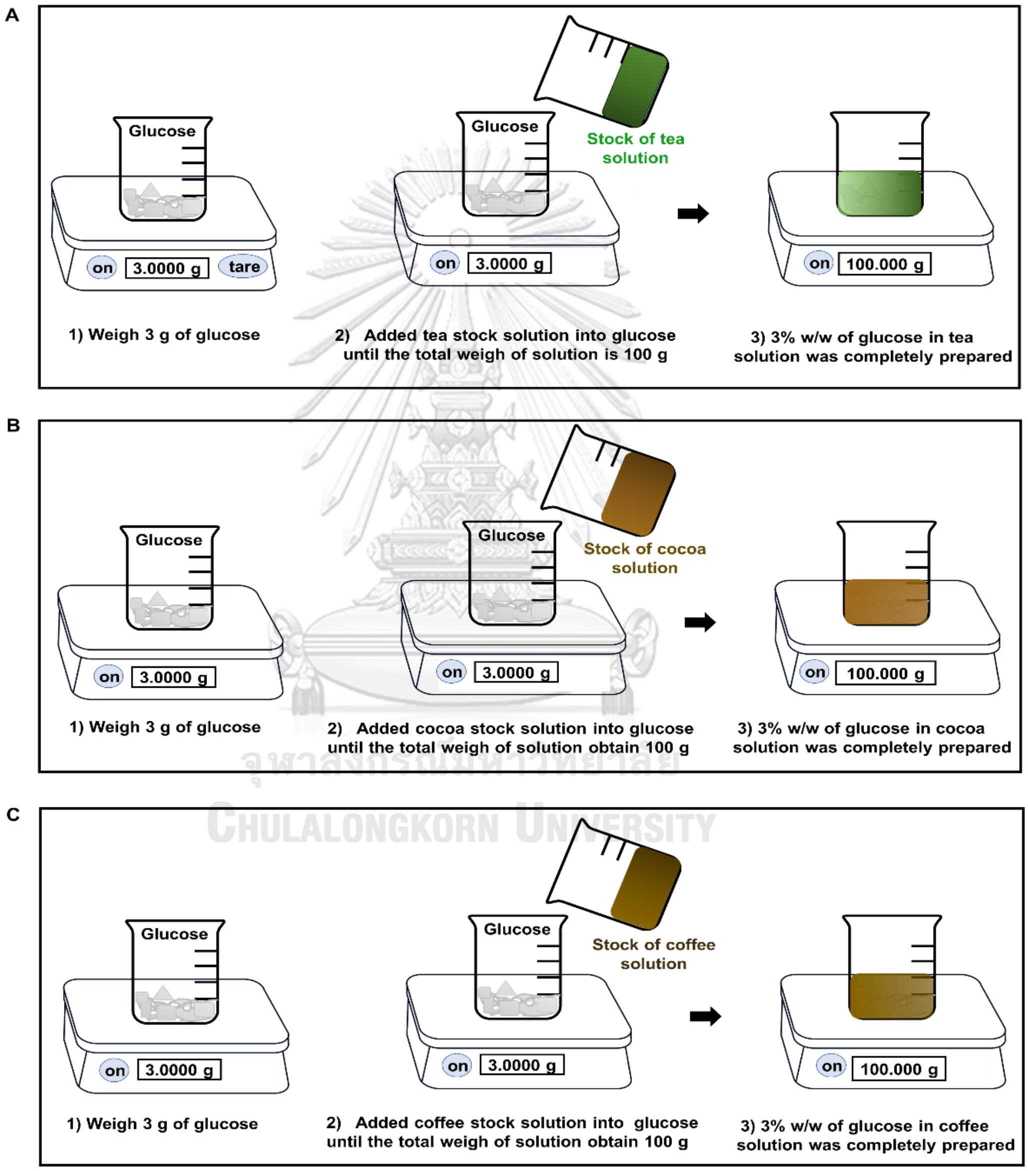
glucose concentrations at 3, 5, 7, 10, 12, 14, 16, and 18 %w/w using DI water as solvent. To prepare percent weight of solution, all steps were performed on the balances with 4 digits. These ranges of the glucose concentrations (3-18 %w/w) were chosen from the average of total sugar in commercial non-alcoholic beverages (100 different types of drink from 20 bands). To perform the other calibration sets, the stock solution of tea, cocoa and coffee (secondary conditions) were used as solvents



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 3. 1 Preparation procedure of stock solution of non-alcoholic drink including (A) tea, (B) cocoa and (C) coffee solution at low level (5g /DI 200 mL) which were further use as solvent for secondary condition. In order to obtain high level, 10g /DI 200 mL of tea, cocoa and coffee was used**

instead of DI water. The validation set was prepared with the glucose concentration at 4, 8, 13 and 17 %w/w which are different from the concentration in the calibration set. This set was used to validate and evaluate the calibration model build from the calibration set of samples. The scheme of the preparation process was shown in Figure 3.1 and 3.2.


A
Glucose
on
3.0000 g
tare
1) Weigh 3 g of glucose
Glucose
Stock of tea solution
on
3.0000 g
2) Added tea stock solution into glucose until the total weigh of solution is 100 g
on
100.000 g
3) 3% w/w of glucose in tea solution was completely prepared
B
Glucose
on
3.0000 g
1) Weigh 3 g of glucose
Glucose
Stock of cocoa solution
on
3.0000 g
2) Added cocoa stock solution into glucose until the total weigh of solution obtain 100 g
on
100.000 g
3) 3% w/w of glucose in cocoa solution was completely prepared
C
Glucose
on
3.0000 g
1) Weigh 3 g of glucose
Glucose
Stock of coffee solution
on
3.0000 g
2) Added coffee stock solution into glucose until the total weigh of solution obtain 100 g
on
100.000 g
3) 3% w/w of glucose in coffee solution was completely prepared

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 3. 2 Procedure for preparation of glucose solution of non-alcoholic drink (3% w/w in the case) using the stock solution of (A) tea, (B) cocoa and (C) coffee as solvent. The procedure will be repeated for glucose concentrations at 5, 7, 10, 12, 14, 16, and 18 %w/w**

### 3.3 Spectral acquisition

NIR spectrometer with NIR256-2.5 detector, LS-1 tungsten halogen light source and fiber optic connector (SMA 905 to 0.22 numerical aperture single-strand optical fiber) purchased from Ocean Optics was used to acquire NIR spectra of the samples. To obtain homogeneous glucose solution, the samples were vigorously stirred for one hour before the NIR acquisition. In order to control the temperature of solution, all of samples were incubated in the water batch controlled at 25$^{o}$C and the humidity was kept at a steady level in the laboratory prior the detection. In case of secondary conditions, the glucose solutions of tea, cocoa and coffee were centrifuged at 5000 rpm for 5 minutes using a temperature-controlled centrifuge (Andreas Hettich GmbH & Co. KG, Germany), then filtered through a 0.45 $\mu$m nylon filter in order to remove all small particles that might scatter the incident light during the NIR detection. According to the high absorptivity, the sample holder was developed and the path length was controlled by spacer of 0.4 mm put between the two individual quartz slide. The setup scheme of NIR instrument used in this work is shown in Figure 3.3. The NIR spectra of the samples were collected using transmittance mode in the range of 1350 nm - 2350 nm using path length 0.4 mm, integration time of 1 millisecond and 32 averaged scans with smoothing windows of 1. Each sample was measured three replicated times. NIR spectra were preprocessed using standard normal variate (SNV) to remove multiplicative interferences of scatter and particle size. The preprocessed NIR spectra was used for the further multivariate data analysis.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

### 3.4 Reference measurement

The accuracy of glucose contents in the solutions were verified by high performance liquid chromatography (HPLC) from food research and testing laboratory (FRTL) Chulalongkorn university to avoid mislabeled samples. The separation column in HPLC was Zorbax $NH_2$ (4.6 x 250 mm, 5 μm) column with mobile phase of Acetonitrile : $H_2O$ (70:30), flow rate of 1.5 mL/min, run time of 15 min and refractive index detection (RID). The determined amount of glucose from the standard HPLC was used as the benchmark of our prepared glucose solutions.

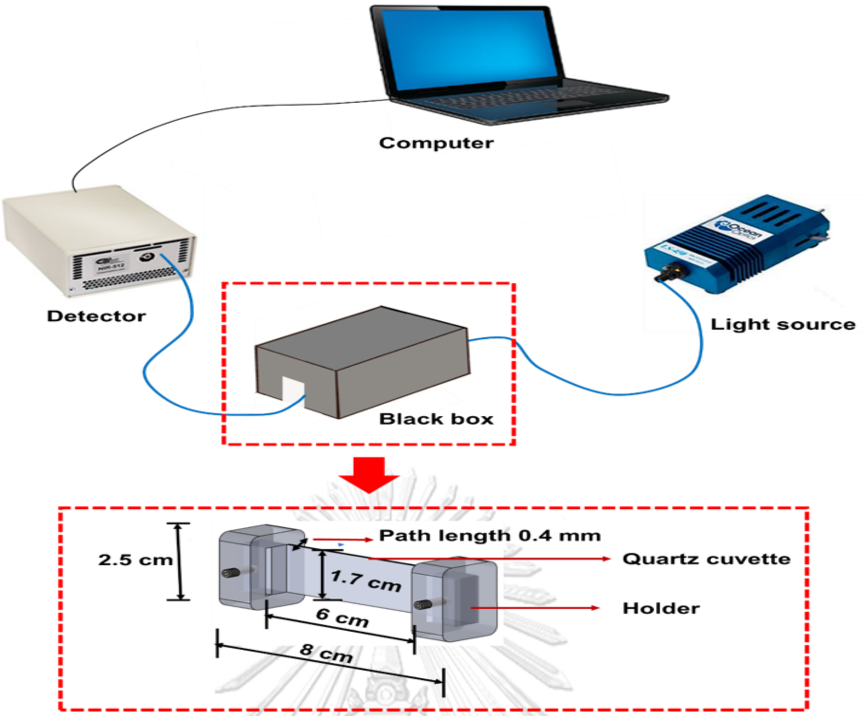


**Figure 3.3 A set up of NIR spectrometer for spectrum acquisition**

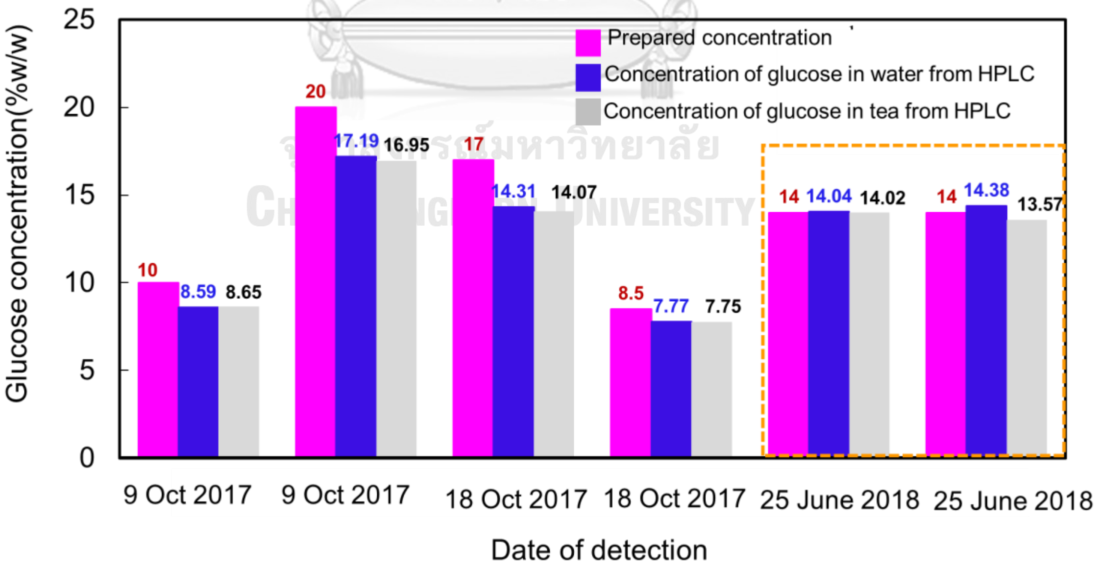


**Figure 3.4 Comparison of glucose contents in water and tea between present glucose concentration and glucose concentration determined from HPLC at food research and testing laboratory (FRTL) Chulalongkorn university**

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 3.4 shows the comparison between presetting glucose concentration (%w/w) shown in pink bar chart and concentration of glucose in water and tea determined by HPLC representing in blue and gray chart, respectively. For sample prepared in 2017, it can be seen that presetting glucose concentration were slightly different from the concentration determined by HPLC. This might due to the preparation protocol and error from instruments. For next testing (2018), the balance was calibrated with the standards of the American Society of Testing Materials (ASTM E617). After calibrating balance, the accurate results with < 0.5% difference were obtained.

## 3.5 Data simulation

According to Beer-Lamberts law, the absorbance of a mixture is a linear combination of the pure spectrum of chemical species and their concentrations. The synthetic NIR spectra were generated by summation of spectra generated from water, glucose and noise. In this case, the noise level can be controlled in order to investigate the performance and limitation of our developed calibration model. Pure spectra of water and glucose were obtained by the acquired spectrum of pure water and pure melt glucose. Noise spectra were generated using the latter PC loading from the spectra of glucose solutions. The simulated NIR spectra was built by summation of spectra from water, glucose and noise as shown in Figure 3.5.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

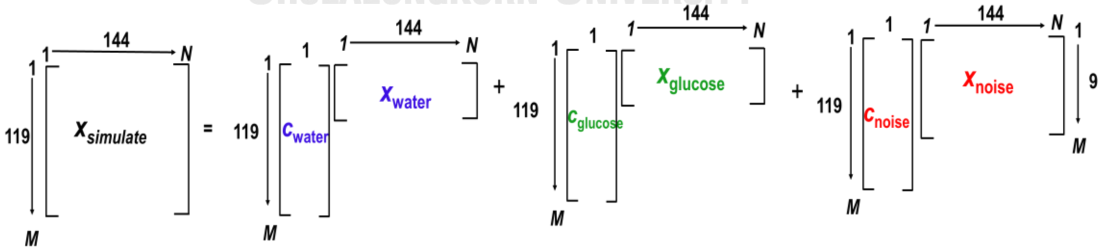


**Figure 3.5 Concept idea of NIR simulated calculation**

**Constraint** : $c_{\text{total}} = c_{\text{water}} + c_{\text{glucose}} + \sum c_{\text{noise}}$ % 100 =w/w all of parameter

$$c_{\text{water}} = 100 - (c_{\text{glucose}} + \sum c_{\text{noise}})$$

where $M$ is total number of sample
$N$ is total number of wavelength
$\boldsymbol{c}_{\mathrm{water}}$ is water content (%w/w)
$\boldsymbol{x}_{\mathrm{water}}$ is pure spectrum of water
$\boldsymbol{c}_{\mathrm{glucose}}$ is glucose content (%w/w)
$\boldsymbol{x}_{\mathrm{glucose}}$ is pure spectrum of glucose
$\boldsymbol{c}_{\mathrm{noise}}$ is noise content (%w/w) that was simulated base on distribution
$\boldsymbol{x}_{\mathrm{noise}}$ is noise spectrum that come from latter loading PC
$\boldsymbol{c}_{\mathrm{total}}$ is total content (100 %w/w)
$\boldsymbol{X}_{\mathrm{simulate}}$ is simulated dataset

The absorbance spectrum of $M$ mixtures containing different concentrations of a diluent (water) and a species of interest (glucose) generated with $N$ wavelengths can be grouped in a data matrix ($\boldsymbol{X}$) where each row represents the spectrum of mixture and each column is wavelength. In this case, the pure spectrum of each species was constrained in all mixtures, while the concentration fraction was controlled by concentration vector of each species. The mass balance was used to limit the total mass of all species summed up to 100 %w/w. In the data simulation, glucose content was varied from 3% w/w – 18%w/w as this can be controlled in real experiment. The fractions of noise were controlled from 1% to 40 %w/w which randomly generated from normal distribution with standard deviation of 10%. Then, the water contents were inversely proportional to the summation of glucose and noise contents which can be provided as $\boldsymbol{c}_{\mathrm{water}} = 100 - \boldsymbol{c}_{\mathrm{glucose}} - \Sigma \boldsymbol{c}_{\mathrm{noise}}$. The simulated spectra with 0% w/w of noise corresponds to the pure spectra of glucose solutions as a primary condition, while the simulated data with 1% - 40 %w/w noise represent the glucose solution in secondary condition.

### 3.6 Wavelength selection

To quantify amount of glucose more precisely, a selection of signal regions from the target analyze might be necessary. Therefore, the wavelengths correlated to the variations of glucose were determined and selected. There are several methods of

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

wavelength selection. One of method is changeable size moving window partial least square (CSMWPLS)[35]. Briefly, basic idea of CSMWPLS is spectral region prospection the window size is determined, then, it moves all over the spectral region. PLS model was performed on each sub window size in order to search the most important regions for improvement of state glucose prediction (low RMSE)[36]. Extended detailed following in this step:

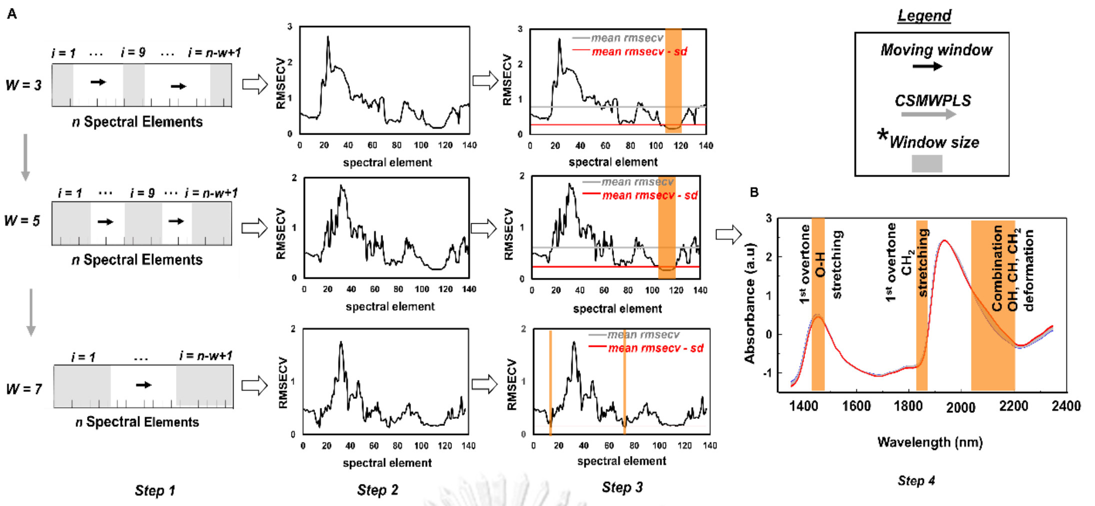
**Step 1:** The NIR spectrum were divided into small sub windows. This window is made by certain number of spectral elements ($i$) and called window size ($w$). In this study, windows size ($w$) was divided into 3 sizes, including 3, 5 and 7. For each window size, there are $\boldsymbol{N}$-$w$+1 windows over the whole spectra, where $\boldsymbol{N}$ is the number of wavelength.

**Step 2:** According to each sub window, PLS models are performed to generate a predictive model. The prediction performance was evaluated by Leave-one -out-cross validation to obtain RMSECV of the sub window. Step 2 will be repeat until RMSECV all sub windows is determined.

**Step 3:** The predictive ability was evaluated by RMSECV presented by each sub window. Region with RMSECV that lower than average of RMSECV – standard deviation of RMSECV was selected as shown in several highlight bands in Figure 3.6A (step 3). It might be implied that significant dependence of the interest variable (glucose in the case) was occurred in that region.

**Step 4:** The selected regions of each window sizes (3, 5 and 7) were unionly selected together as show in highlight bands in Figure 3.6B (step4)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



* Window position at *ith* spectra element (with size w) corresponding to the group of spectral elements contained between element *(i)* and *(i+w-1)*, used to create *i* PLS model

**Figure 3. 6 (A) Scheme for explanation of CSMWPLS with n is number of wavelength and w is sub window size including 3, 5 and 7. (B) Optimized region of NIR spectra for prediction of glucose in function of RMSECV on PLS model, applying CSMWPLS**

## 3.7 Chemometrics

A proposed chemometrics approach is to extract only the variation from the interested component which might involve only water, glucose and their interactions for establishing the appropriate calibration model to quantify the glucose concentrations in secondary conditions. In this study, the calculation methodology was separated into two parts. Firstly, spectral decomposition was involved to separate the variation in the NIR spectra into smaller parts. Then, only the effective variations were chosen for the future analysis. This involves spectral decomposition method based on use of simple PCA decomposition. Secondly, the calibration model using PLSR was built from the selected variations in order to form a universal model. The calculation methodology was performed in the following steps:

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Step of this global model followings as:**

**Step 1** The raw NIR spectra of both primary ($\boldsymbol{X}_{\text{primary}}$) and secondary ($\boldsymbol{X}_{\text{secondary}}$) condition were smoothed using Savitsky-Golay with 5 window sizes and follows by SNV in order to eliminate the influence of background shift. The *N* row and *M* columns of the data matrix represents the sample different concentration of glucose and wavelengths, respectively.

**Step 2** To obtain the major variations, sample set from secondary condition were added to the calibration set from primary condition. In this step, number of inserted samples from $\boldsymbol{X}_{\text{secondary}}$ secondary condition was varied at 1, 5, 10, 20 and 40 to investigate the influences of the variations from secondary conditions.

$$\boldsymbol{X}_{obs} = \begin{bmatrix} \boldsymbol{X}_{primry} \\ \boldsymbol{X}_{secondary} \end{bmatrix}$$

where $\boldsymbol{X}_{\text{obs is}}$ observation matrix which is a combination of $\boldsymbol{X}_{\text{primary}}$ and $\boldsymbol{X}_{\text{secondary}}$
$\boldsymbol{X}_{\text{primary}}$ is NIR spectra from primary condition
$\boldsymbol{X}_{\text{secondary}}$ is NIR spectra from secondary condition

**Step 3** Perform PCA as a mathematical transformation to extract loading and scores matrix which correlate to the major variation of $\boldsymbol{X}_{\text{obs}}$. In this step, the maximum number of principal component were set to 10 PCs (the total variance up to >99.99%)

General form: $\boldsymbol{X}_{\text{obs}} = \boldsymbol{T}_{\text{obs}} \, \boldsymbol{.} \, \boldsymbol{P}_{\text{obs}} + \boldsymbol{E}$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Matrix form: $\begin{bmatrix} \boldsymbol{X}_{\text{primary}} \\ \boldsymbol{X}_{\text{secondary}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_{\text{primary}} & . & \boldsymbol{P}_{\text{primary}} \\ \boldsymbol{T}_{\text{secondary}} & . & \boldsymbol{P}_{\text{secondary}} \end{bmatrix}$

where $\boldsymbol{T}$ ($M$ x $A$) is the scores with $M$ rows correspond to number for sample and $A$ column correspond to number of PC
$\boldsymbol{P}$ is the loadings ($A$ x $N$) with $A$ row (number of PC) and $N$ column (wavelength)

**Step 4** Excluding the extracted data including score and loading from secondary condition ($\boldsymbol{T}_{\text{secondary}}\boldsymbol{P}_{\text{secondary}}$). The data of score and loading of primary condition were remained for further calculation, however, the variations from secondary condition were already included in $\boldsymbol{T}_{\text{primary}}\boldsymbol{P}_{\text{primary}}$. For further analysis, the new observation spectra ($\boldsymbol{X}_{\text{new\_primary}}$) was generated with only significant of PCs (PC1 to PC10).

**Step 5** The calibration model of $\boldsymbol{X}_{\text{new\_primary}}$ is generated using Partial Least Squares Regression (PLSR). In its simplest form, a linear model specifies the linear relationship between response $\boldsymbol{y}$ (glucose concentrations), and a set of variables of the $\boldsymbol{X}_{\text{primary}}$ (wavelength) which can be expressed by

$$\boldsymbol{y} = \boldsymbol{X}_{\text{new\_primary}} \,.\, \boldsymbol{b} + \boldsymbol{E} = \boldsymbol{T} \,.\, \boldsymbol{q} + \boldsymbol{E}$$

where $\boldsymbol{T}$ and $\boldsymbol{q}$ are PLS score matrix and PLS loading vector, respectively. $\boldsymbol{b}$ is the regression coefficient vector estimated as follows:

$$\boldsymbol{b} = \boldsymbol{H} \,.\, \boldsymbol{q}$$

where $\boldsymbol{H}$ is the PLS weight matrix (described in Section 2.2).
In our study, maximum number of PLS component was limited to 25.

**Step 6** This step involves the optimization of number of PCs and number of PLS component which give the smallest Root-Mean-Square Error using Leave one out cross validation (RMSECV). The number of PCs and PLS component were optimized using grid search approach with row and column of grid represents number of PCs and PLS component, respectively.

$$\begin{array}{cc} & 1 \xrightarrow{\hspace{6em}} K \quad \text{PLS component} \\ \begin{array}{c} 1 \\ \big\downarrow \\ A \\ \textit{PC component} \end{array} & \begin{bmatrix} \mathit{RMSECV(1,1)} & \cdots & ,\ \mathit{RMSECV(1,K)} \\ \vdots & \ddots & \vdots \\ \mathit{RMSECV(A,1)} & \cdots & \mathit{RMSECV(A,K)} \end{bmatrix} \end{array}$$

Where $A$ (in row) represent as PC component (The maximum number of PCs is 10)

$K$ (in column) represent as PLS component (The maximum number of PLS is 25)

The optimal PC and PLS was selected by the coordination (PC comp, PLS comp) which gives the lowest *RMSECV*.

**Step 7** Using an optimal number of PC from step 6 to create the new $\boldsymbol{X}_{\text{new}}$ and then separated into $\boldsymbol{X}_{\text{new\_primary}}$ and $\boldsymbol{X}_{\text{new\_secondary}}$ again.

General form: $$\boldsymbol{X}_{\text{new}} = \boldsymbol{T}_{\text{obs\_PC}} \cdot \boldsymbol{P}_{\text{obs\_PC}}$$

Matrix form: $$\begin{bmatrix} \boldsymbol{X}_{\text{new\_primary}} \\ \boldsymbol{X}_{\text{new\_secondary}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_{\text{new\_primary}} & \cdot & \boldsymbol{P}_{\text{new\_primary}} \\ \boldsymbol{T}_{\text{new\_secondary}} & \cdot & \boldsymbol{P}_{\text{new\_secondary}} \end{bmatrix}$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

**Step 8** The PLS calibration model including PLS regression coefficients ($\boldsymbol{b}$) was calculated from $\boldsymbol{X}_{\text{new\_primary}}$. Then, the generated model was used to predict the response of $\boldsymbol{X}_{\text{new\_secondary}}$ as a validation sample set

$$\boldsymbol{y}_{\text{secondary}} = \boldsymbol{X}_{\text{new\_secondary}} \cdot \boldsymbol{b}_{\text{new}}$$

**Step 9** Estimate the model performance using RMSE as a validation index.

The performance of calibration model was evaluated in terms of root mean square error of calibration (*RMSEC*) index, root mean square error of cross validation (*RMSECV*) and root mean square error of prediction (*RMSEP*).

They can be denoted as,

$$RMSEC = \sqrt{(\sum_{m=1}^{M}(y - y_{c)}{}_{m}^{2}/M}$$

$$RMSECV = \sqrt{(\sum_{m=1}^{M}(y - y_{cv)}{}_{m}^{2}/M}$$

$$RMSEP = \sqrt{(\sum_{m=1}^{M}(y - y_{p)}{}_{m}^{2}/M}$$

Where $y$ contains the actual value (glucose concentration in the case)
$y_{cv}$ contains the estimated glucose concentration by leave one out cross validation
$y_c$ contains the estimated glucose concentration of the sample in calibration set
$y_p$ contains the estimated glucose concentration of the sample in validation set
$m$ is the selected number of sample in the data set.
$M$ is total number of sample in the data set.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A conceptual view was showed in Figure 3.7

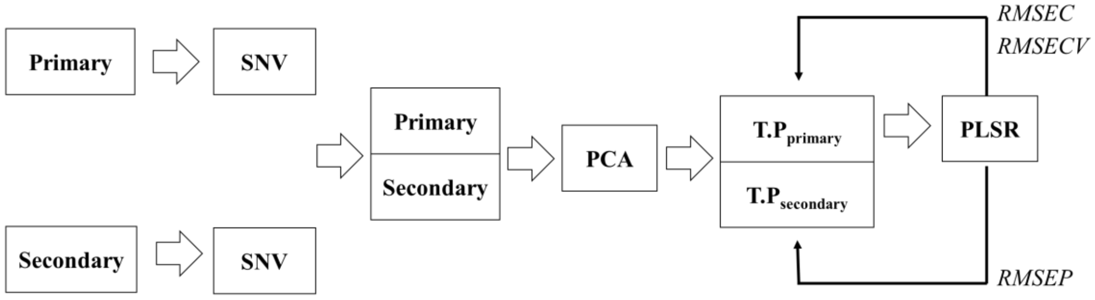
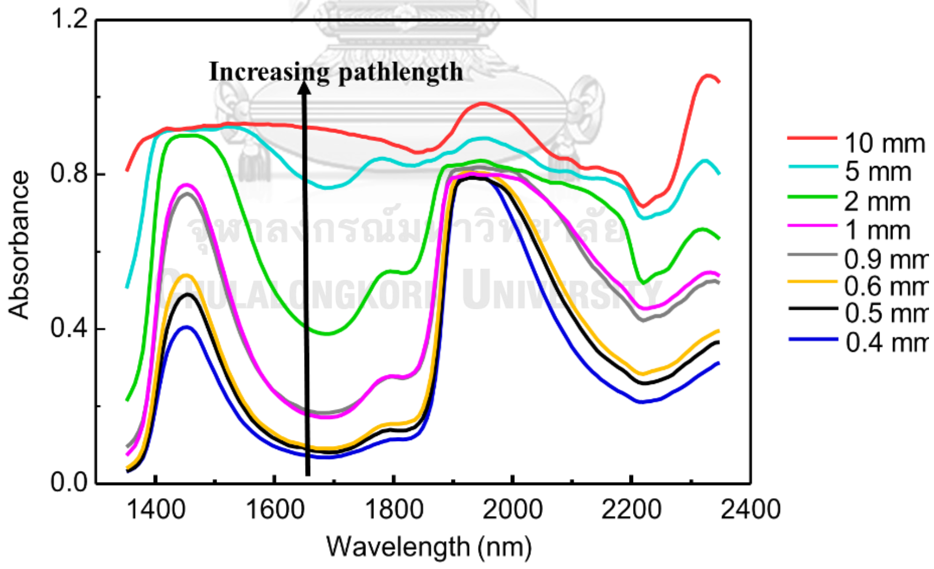


**Figure 3.7 Conceptual view of the calculation methodology model using multivariate data analysis (PLS) with conventional way (full spectrum) and global model (using PCA)**

# CHAPTER IV
# RESULTS AND DISSCUSSION

## 4.1 NIR spectrum acquisition

The NIR spectra in transmitting mode would be affected by different position of spectral measurement and different detection parameters of spectral scan such as path-length, scanning rate, integration time, smoothing windows. Therefore, the spectral measurement must be carried through under the uniform experimental conditions. Firstly, the measurement parameters were optimized in order to obtain the appropriate NIR spectra which demonstrate all overtones and not over absorbed. In this case, integration time of 1 millisecond and 32 averaged scans with smoothing window of 1 was set according the presetting of NIR instrument. The path-length of the detection was varied at 0.4 – 10 mm in order to obtain the maximum informative spectra from the detection as shown in Figure 4.1.



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 4.1 NIR spectra of water acquired with different path-length in range of 0.4 mm – 10 mm using integration time of 1 and 32 averaged scans with smoothing window of 1.**

According to Beer-Lambert law, the higher path-length, the higher absorbance obtains. However, the problem of over adsorb band will occur when an excessive path length is used. Figure 1 show set of NIR spectra of DI water acquired using different path-lengths. The part of NIR spectra especially at the range of 1,900 – 2,000 nm exhibit over absorbance which represent by cut off at the top of peak, when large path-length > 0.6 mm was used. The NIR spectra acquired using path-length of 0.4 mm show the minimum background shift. Therefore, the path-length of 0.4 mm was chosen for further NIR acquisition.

### 4.1.1. Preprocessing techniques

The NIR spectra data preprocessing is an essential part of chemometric modeling. The NIR spectra of samples are mostly influenced by the physical properties of the samples and other effects from environments e.g. human errors, outside incident lights, holder positions etc.  The purpose of preprocessing is to increase the important information and to minimize the contribution of irrelevant information. The proper options of the preprocessing technique depend on the nature of data and difficult to assess before the model validation. Therefore, the preprocessing technique of the acquired NIR spectra is carried out through trial and error approach. This study applies three basic techniques involving the Savitzky-Golay smoothing, the second spectral derivative and Standard Normal Variate (SNV). Figure 4.2 shows the smoothed NIR spectra of raw spectra, derivative spectra and normalized spectra. It can be seen that preprocessing method affects the behavior of signal patterns. Original data without any signal pretreatment is highly susceptible to noise, inconsistency and baseline shift causing to the low quality of the data. In order to improve the quality of the NIR spectra, smoothed with second derivative which used to eliminate baseline errors and resolve overlapped peak was applied (Figure 4.2). A second derivative spectrum was calculated for each measurement by using the Savizky-Golay algorithm (5 smoothing points). However, the improvement of the peak resolution is still unclear.  Even domination bands of water between 1400 – 1600 nm still show fluctuation and noises. One of a major reason is wavelength gap spectral resolution due to the slit aperture of the instrument. For our NIR spectrometer, the slit aperture of 7 nm was constantly operated. Therefore, the change

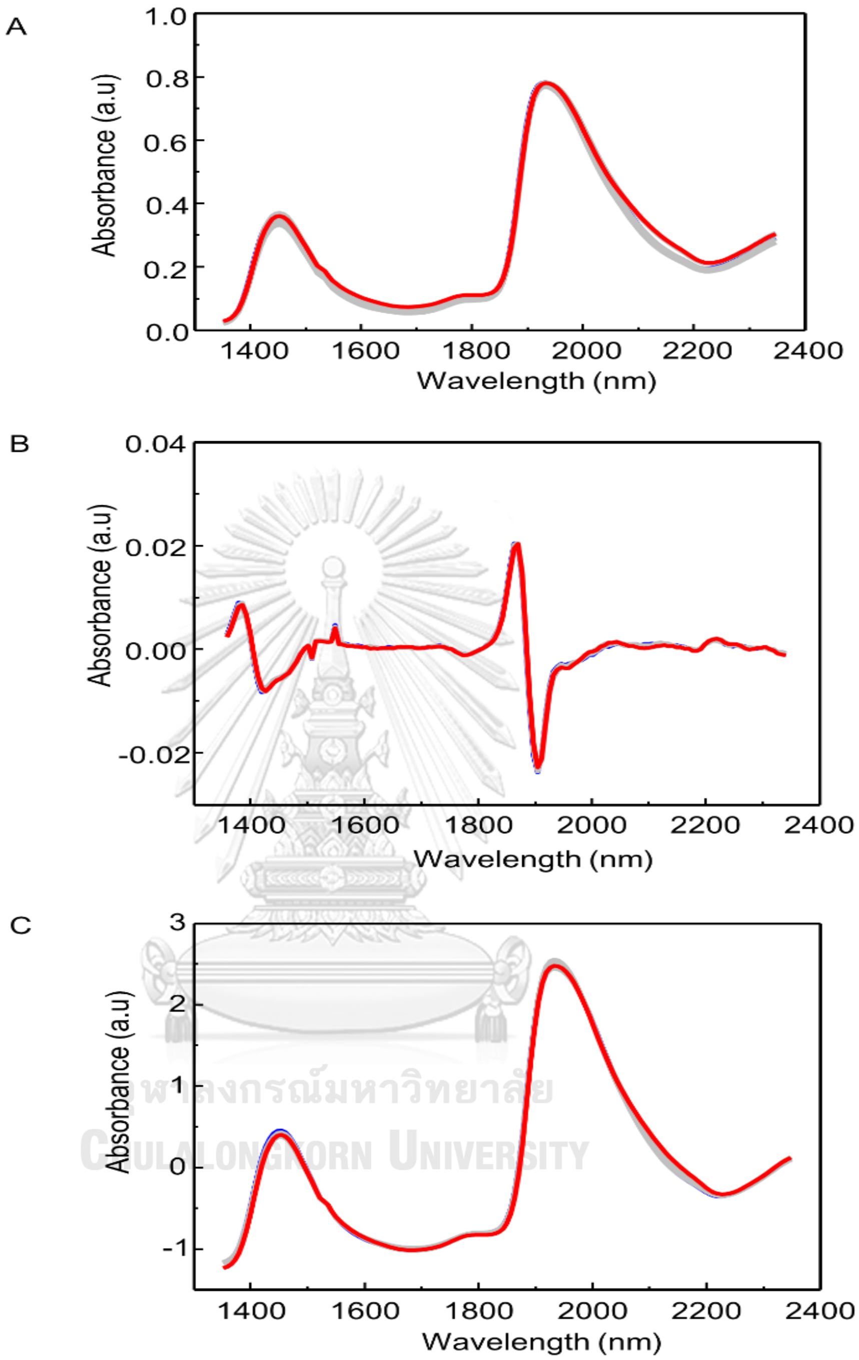
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

of slope in some part of spectrum might be strongly fluctuated due to a high wavelength shift of the detection. Second derivatives can also be employed to decrease baseline shifts and curvilinearity, but noise and complexity of the spectra increases. Another preprocessing method including standard normal variate (SNV) is used to removes the multiplicative interference of scatter and particle size[37] which causes the baseline shift. SNV is designed to operate on individual sample spectra, therefore, it is unaffected with the spectrum of other samples. From figure 2c, it is obvious seen that the baseline shifts were corrected and intensity were more correlated to the responses. Therefore, it might be indicated that smooth with standard normal variate was selected as an appropriate preprocessing method which was used to pretreat the raw NIR spectra for the further multivariate data analysis.

### 4.1.2. Variations of the NIR spectra of glucose solution with different concentrations

In the previous section, we already mention on the importance of signal preprocessing method on the acquired NIR spectra. In this study, smooth with standard normal variate (SNV) was selected because it can reduce noise and remove multiplicative interferences of scatter and particle size. To visualize the characteristic overtones of samples, NIR spectrum of pure water (blue line), pure glucose (black line) and 18%w/w of glucose solutions (red line) were shown in figure 4.3A. It can be seen that these three samples have the same dominated peak at 1450 nm and 1900 nm corresponding to 1st vibration overtone O-H stretching and combination of O-H deformation[24], respectively. However, the overtone patterns of pure glucose (black line) shows a tiny overtone peak at 1800 nm due to 1st overtone C-H stretching of glucose which do not appear in glucose solution (red line)[24]. From figure 4.3B, it is difficult to observe the characteristic overtone band of glucose from glucose solution directly. Therefore, mathematic and statistic approach in form of variance was calculated to reveal the major variations in the NIR spectra of glucose solution prepared with different concentrations. The variance of the NIR spectra was calculated and plotted as shown in Figure 4.3C

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 4.2 Savitzky-Golay smoothing NIR spectra of glucose solutions after performing different preprocessing methods including (A) raw data (B) the second spectral derivative and (C) Standard Normal Variate (SNV). The red and blue line represent the NIR spectrum of the highest (18 %w/w) and the lowest (3%w/w) glucose concentration, respectively**

According to the variance plot in Figure 4.3C, the three major bands with high variance were observed. They include 1st overtone O-H stretching of water at 1450 nm, O-H and $CH_2$ combination band at 1900 nm and 2100 nm, respectively. It could be implied that these bands are strongly correlated to the concentration of the glucose solutions. The intensities of these bands are influenced by the amount of glucose in the solution. The variance plot indicates the most variation regions in the spectra, however, they cannot provide the information about the direction of variation. Therefore, in order to visualize the variability direction of NIR spectra along with overtone patterns of glucose solutions (in Figure 4.3B) is required.

The band assignment of the major components (in glucose solution) are briefly summarized in Figure 4.3B. Figure 4.3B shows the average NIR spectra of glucose solution at 3, 5, 7, 10, 12, 14, 16 and 18 %w/w. It can be seen that the NIR spectra are dominated by water absorption bands at 1450 corresponding to 1st overtone O-H stretching of water[24] and 1950 nm relating to the combination bands of stretching and deformation of the O-H group in water[38] (as mentioned above). The characteristic band of glucose associated to 1st overtone C-H stretching (-$CH_3$ and –$CH_2$-) in the range of 1600 to 1700 nm is unfortunately low intensity, while the band at 2100 nm corresponding to the 1st set of C-H combination band is very strong[24]. From variance plots, this suggests that the intensity of NIR spectra was changed depending on glucose concentrations.

To demonstrate the direction of variability, the intensity of the assigned band was magnified and shown as the insets of Figure 4.3. The intensity of 1st overtone of CH (1600-1700 nm and combination bands at 2100-2200 nm) increases when the glucose concentration increases. This represents the direct variation due to characteristic band of glucose. On the other hand, the 1st overtone of OH stretching (1450 nm) and combination bands (1900 – 2000 nm) show an inverse variation because these overtone regions are correlated with water content. These trends are in good agreement with the variance plot.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

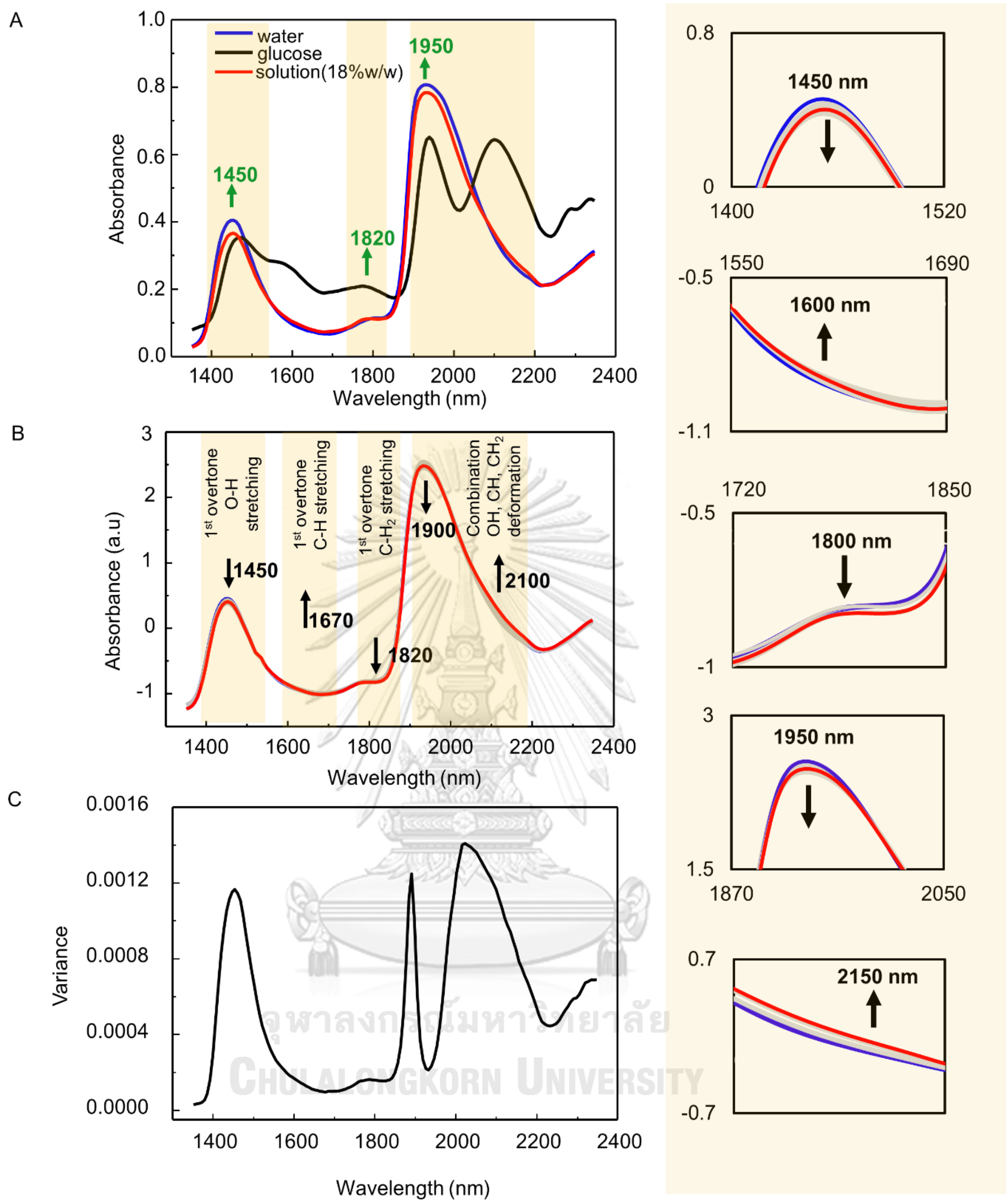


**Figure 4.3** NIR spectrum (A) combination spectrum of pure water, pure glucose and glucose solution (18%w/w) in water following blue, black and red line, respectively. (B) glucose solution in water system (18%w/w) after performed baseline correction using standard normal variate (SNV), (C) variance sample of average NIR spectrum of glucose in water (primary condition)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

### 4.2 Primary calibration set of samples

Selection of representative calibration samples is important for the success of the further statistical modeling. Moreover, it would be important to note that the samples selected for the calibration should cover the variation of future samples. The main concept of this work is to establish a universal calibration model from a primary condition (glucose solution in the case) which can be used to predict glucose concentrations in a secondary condition (non-alcoholic drinks). To search for an appropriate normal operating samples (NOS) is our primary target. The calibration set was constructed from glucose solution with 8 different actual concentrations (3, 5, 7, 10, 12, 14, 16 and 18 %w/w) with 3 repetitions involving total 120 samples ($y_{\text{actual}}$). To remove anomalous observations, the outlier detection has been performed. PLS regression with leave-one-out cross validation approach was performed on the all samples in the calibration sets in order to obtain the prediction of concentration ($y_{\text{predict}}$). The error of $y_{\text{predict}} - y_{\text{actual}}$ was calculated for all samples. Any sample with the error more than three standard deviations from the average errors was determined as an outlier[39]. Figure 4.4A shows the error of each sample (blue dot), the average error (red dashed line) and three stand derivations (black dashed line). It can be seen that only one sample (sample number 1, batch 1) was determined as an outlier. Therefore, this sample was removed from the calibration set. From the prediction error, it suggests that our experiment set up is consistent in each repetition as only one outlier exists. Next, the combination of the calibration set was optimized by RMSECV and RMSEC values which are an index to evaluate the appropriate NOS. Lower value of RMSE, higher accuracy of the prediction was occurred. Besides, the gap between RMSECV and RMSEC was also under consideration to estimate model. The small gap represents non-overfitting model. Figure 4.4B show acceptable value of RMSEC and RMSECV of approximately 0.05 and 0.11, respectively, occurred in individual batch and its combination. Meanwhile, gap between RMSEC and RMSECV in each batch was obtained of 0.06 which exhibit insignificantly different. Thus, the combination of batch 1, 2 and 3 was selected as a benchmark of NOS to construct a universal model.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

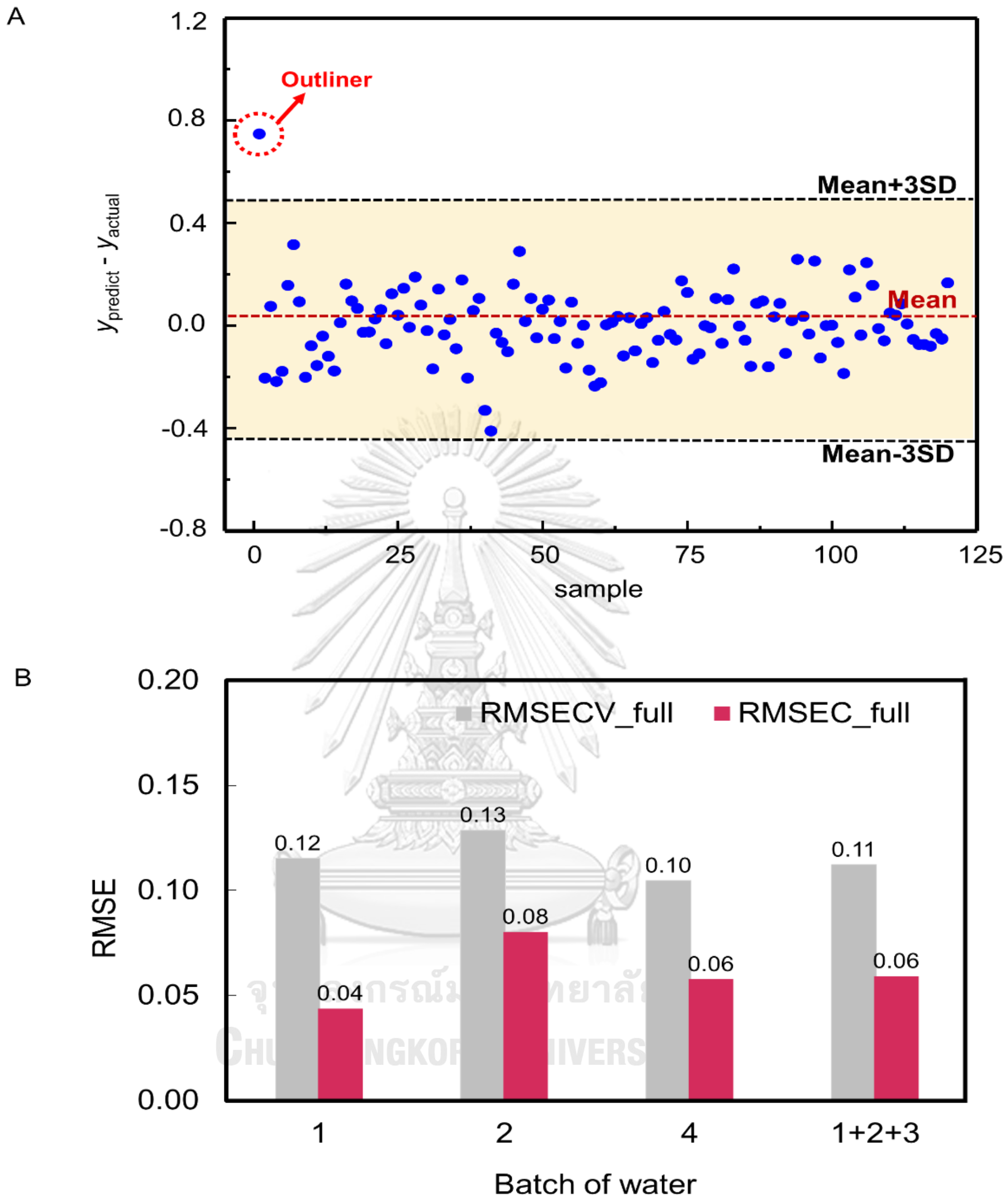
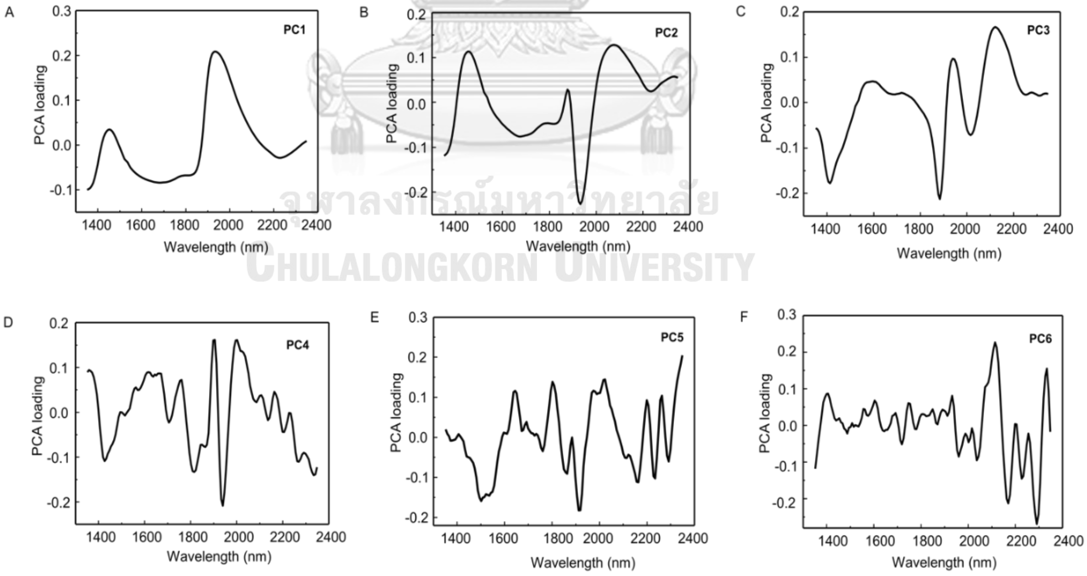


**Figure 4.4** (A) the error of each sample calculated from ypredict – yactual (blue spots) including the average error of all samples (red dot line) and three standard deviation of average error (black dot line) (B) Bar plot between root mean square error of cross validation (RMSECV) (gray bar) and root mean square error of calibration (RMSEC) (pink bar) on NIR spectra of glucose concentration in different repeated batches

Prior to the further data analysis, principal component analysis (PCA) was used to extract the major components of the data matrix which is the NIR spectra of all batches. The characteristic pattern of major components is revealed by loading of each PC as shown in Figure 4.5. It can be seen that the loading of PC1 shows the characteristic patterns for water because it reveals a strong absorbance at 1450 nm (1st overtone of O-H bond stretching) and 1950 nm (O-H combination band). In case of loading of PC2, it shows a similar pattern of PC1 loading but the direction of band at 1950 nm is inversed. This suggests that there is some interaction between glucose and water though the combination band of O-H stretching. Whereas, the noticeably band of the PC3 loading was appeared at 1890 nm and 2050 nm which are corresponding to the $CH_2$ stretching of glucose and the 1st set of C-H combination band of glucose, respectively. These PC loadings are in good agreement with the pure spectra (Figure 4.3A). However, the latter PC loading do not show any distinctly signals compared to the loading from PC1-PC3. This suggest that the main components are occurred only in the first few PCs, therefore, the higher PC might not be necessary in the model.



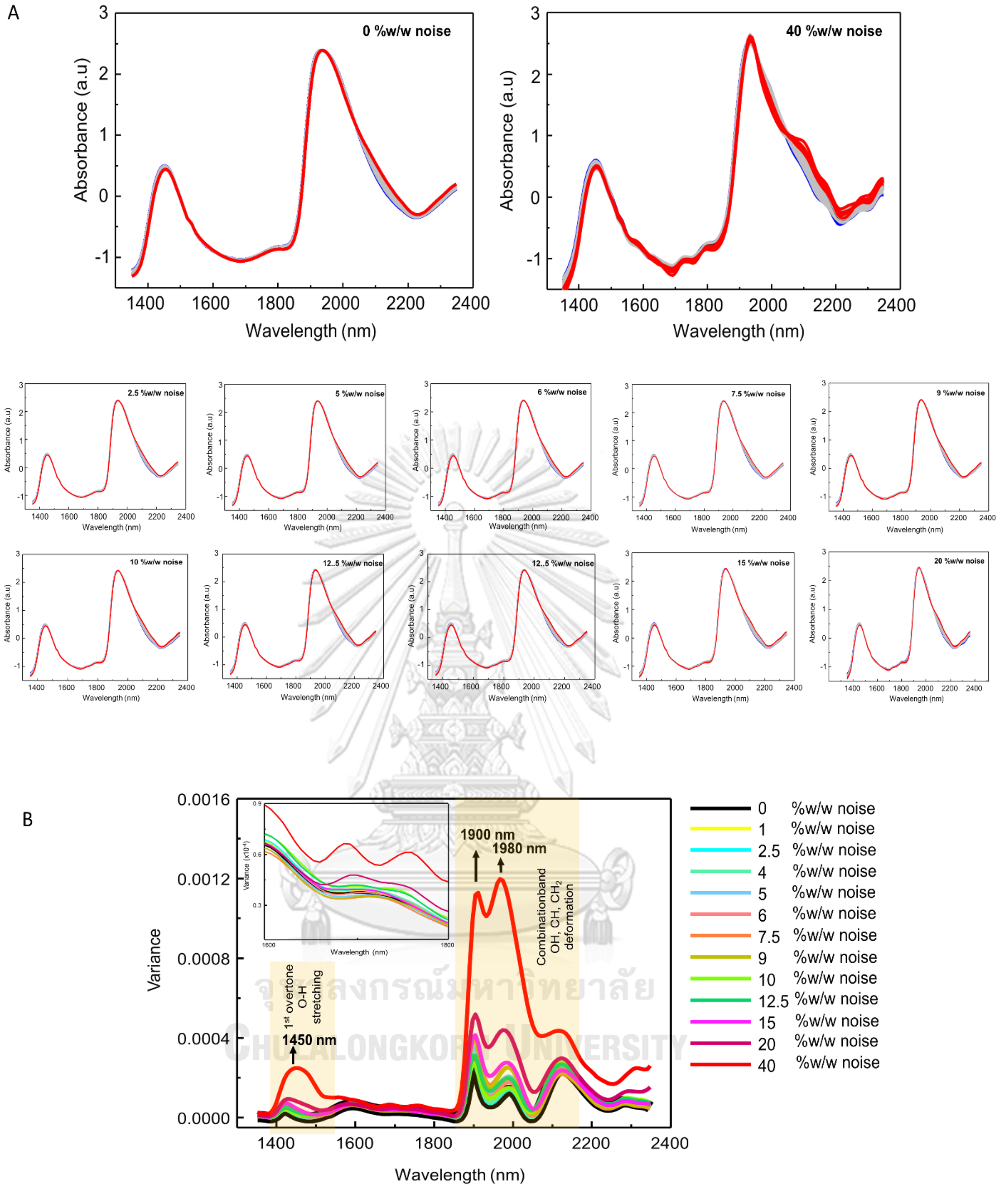
จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 4.5 Loading profiles (PC1 to PC6) of the NIR spectra from batch 1-3 after performing PCA.**

## 4.3 Simulated datasets

A motivation of the simulations was to produce the data with controllable underlying distribution of correlations that is similar to those found in the real datasets. The simulated NIR spectra were calculated using summation between concentration and pure spectra of the main components (water and glucose in the case) with the additional noises at several different levels. However, the total concentration of noise spectra was controlled by mass balance of 1% - 40 % w/w. The concentration of glucose was constrained, therefore, the added noise would only affect to the proportion of water spectrum. The simulated spectra with the different additional noise levels and the corresponding sample variances were shown in Figure 4.6. It can be seen that low noise level does not affect the spectrum pattern, since they preserve as much as possible the characteristic of the original pure simulated NIR spectra until 12.5%w/w noise. The baseline shift was initially occurred in the spectra simulated with noise level over 15%w/w, while the pattern of the simulated NIR spectra was totally changed at noise level of 40%w/w. The variance plots of each simulated spectra were shown in Figure 4.6B. It can be seen that the variance plot demonstrates the strong variations correlated with presetting glucose concentrations which are including 1st overtone O-H stretching (1450 nm) and combination O-H deformation (1900 nm and 1980 nm) of water. The simulated NIR spectra without noise come from summation between concentration and pure spectra of the main components (water and glucose in the case), so added noise level would be effect to only the overtone regions of water. Obviously, the higher added noise level, the more variations on the pattern of spectra occurs. From the simulations, it can be noticed that an added noise is limited to 1-12.5 %w/w as it can be able to maintain the identity of NIR spectrum which is in good agreement with the real spectra from the experiments.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Figure 4.6 (A) Simulated NIR spectra and (B) Variance plot of the simulated NIR spectra at different additional noise levels (0-40%)**

To evaluate the prediction performance, the PLS calibration model was built from the simulated NIR spectra without any additional noises. Then, the generated PLS model was used to predict the glucose concentration of the other independent simulated datasets with the noise at different levels. The root mean square error of prediction (RMSEP) and coefficient of determination ($r^2$) were calculated in order to express the model performance. The lower RMSEP and the higher $r^2$ represent a prediction performance of the model. A plot of RMSEP (black line) and $r^2$ (blue line) against noise level is shown in figure 4.7A. It can be seen that the RMSEP values are slightly increased from 0.12 – 1.66 for the data with noise 1- 12.5%w/w, respectively. After using noise level over 15%w/w, the RMSEP value dramatically raises from 2.05 to 7.19. This suggests that the capability of a predictive model is directly related to noise in the data. The higher noise level, the lower predictive ability of the model occurs. These observations are in good agreement with $r^2$ plot. The $r^2$ value is over 0.99 for the prediction of the data with noise only in the range of 1-12.5%w/w. The $r^2$ value is lower than 0.99 when the noise level is up to 15%w/w. However, the $r^2$ value is still good (>0.98) even for the prediction of the data with noise 40%w/w. To demonstrate the ability to measure each sample independently, the concentration correlation plots between actual preset glucose concentration and the predictive value are presented in Figure 4.7B. In all the plots, the prediction points all fall on the ideal diagonal line with no apparent systematic variation for the data with 0% noise level. However, the appearance of variation on prediction especially for high glucose concentration will occur when the noise level was increased. This suggests that our PLS calibration model can predict more accurately at the low glucose concentration rather than the high one.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

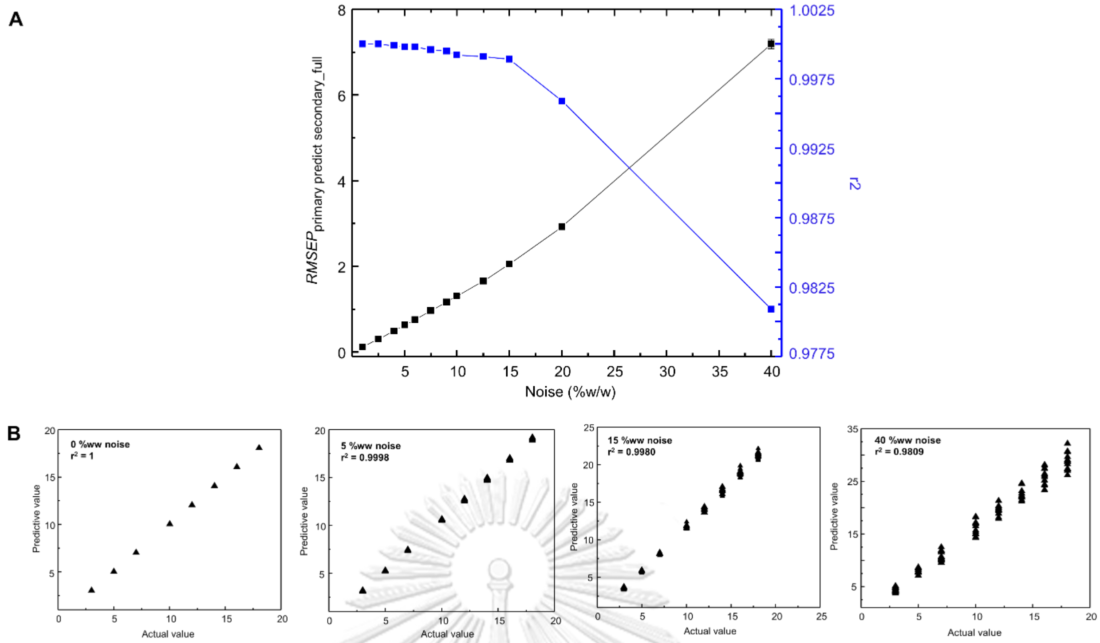


**Figure 4.7 (A) RMSEP (black line on the left Y axis) and $r^2$ (blue line on the right Y axis) plot against different additional noise levels, (B) concentration correlation between actual values (X-axis) and predicted values (Y-axis) of glucose concentrations at different levels of noise (%w/w)**

In the previous prediction, the PLS calibration models were generated using the dataset containing all samples and wavelengths without any additional noise. This calibration model was used to predict the glucose concentration of the other datasets with different noise levels. In this study, assessment of model selectivity is critical for achieving a prediction of glucose concentration. In order to extract the pure components of the dataset, principal component analysis (PCA) were performed on the data to extract the major components (glucose and water). To generate the universal model, the PCs of all major components contributed to the data matrix including calibration set (primary condition) and also system (secondary condition) must be determined and selected. Then, the new data matrix built from only the selected PCs was calculated for further data analysis. It should be noted that the inserted number of sample from secondary condition might affect the total variance of the data matrix (from primary condition) and the prediction. In this section, the influences from different number of inserted samples on the PC selection and model

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

prediction are investigated. The error of prediction ($\mathrm{RMSEP}_{\mathrm{pc}}$) of the model built from major components with different number of inserted sample from 1– 40 are plotted against the added noise level as shown in Figure 4.8. The $\mathrm{RMSEP}_{\mathrm{pc}}$ generally increases when noise level of the data increases. It raises from 0.11 to 2.93 for noise level 1% and 40%, respectively. Moreover, the $\mathrm{RMSEP}_{\mathrm{pc}}$ seem increase when the number of inserted samples increases. This shows the number of inserted samples which come from secondary condition have strongly influenced on the variance of the data from primary condition and the prediction. Although, the lowest $\mathrm{RMSEP}_{\mathrm{pc}}$ would be obtained by using only single inserted sample, the calculation will not be practical when the external sets of samples are large. In this case, we would like to determine the optimal number of inserted samples that would insignificantly alter the prediction. Relative percent error of $\mathrm{RMSEP}_{\mathrm{pc}}$ using different number of inserted sample compared with single inserted sample was calculated:

$$\text{Relative percent error} = \left(\frac{RMSEP_{\mathrm{pc\_single\ added\ sample}} - RMSEP_{\mathrm{pc\_}m\ \mathrm{added\ sample}}}{RMSEP_{\mathrm{pc\_single\ added\ sample}}}\right) x\ 100$$

Where $\mathrm{RMSEP}_{\mathrm{pc\_single\ added\ sample}}$ is $\mathrm{RMSEP}_{\mathrm{pc}}$ that calculate from single added sample

$\mathrm{RMSEP}_{\mathrm{pc\_}m\ \mathrm{added\ sample}}$ is $\mathrm{RMSEP}_{\mathrm{pc}}$ that calculate from various numbers of added sample with $m$ obtain added number of sample

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

The calculation result was show in an inset of Figure 4.8. The $\mathrm{RMSEP}_{\mathrm{pc}}$ will be increased more than 5% when the number of inserted sample were up to 20. On the other hand, the $\mathrm{RMSEP}_{\mathrm{pc}}$ was changed for less than 2% when the number of inserted sample equal to 10. For this observation, it might be suggested that limitation of inserted number of sample is around 20 samples and the optimal number sample of 10 was selected to further multivariate data analysis. However, it should be note that the limitation of inserted sample strongly depends on the total sample number of the dataset from primary condition. The larger size of the dataset, the limitation might be raised.

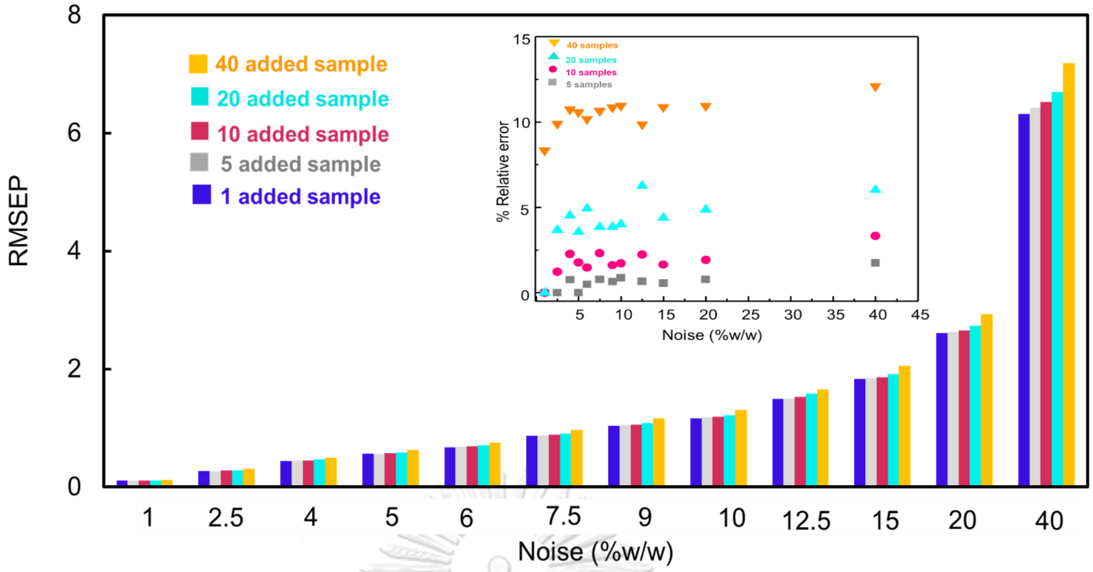


**Figure 4.8 RMSEP calculated using different number of inserted samples (1-40 samples). The inset figure shows the relative percent error of RMSEP compared with RMSEP obtained from using single inserted sample**

In order to be able to quantify amount of glucose more precisely, a selection of signal from the target analyze is indispensable. Therefore, the wavelengths correlated to the variation of glucoses was determined and selected. There are several methods of wavelength selection available in the literatures. One of an efficient method is to use moving window partial least square[35]. The protocol detail was already discussed in section3.6, chapter 3. In this case, NIR spectrum of 9%w/w additional noise was selected to represent in this wavelength selection. Sub regions of 1st overtone O-H bond stretching of water (1350-1420 nm), 1st overtone $CH_2$ stretching of glucose (1820-1946), OH, CH, $CH_2$ combination bands are selected as shown in Figure 4.9B. It should be noted that the selected sub-regions might be changed due to the noise level of the dataset.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

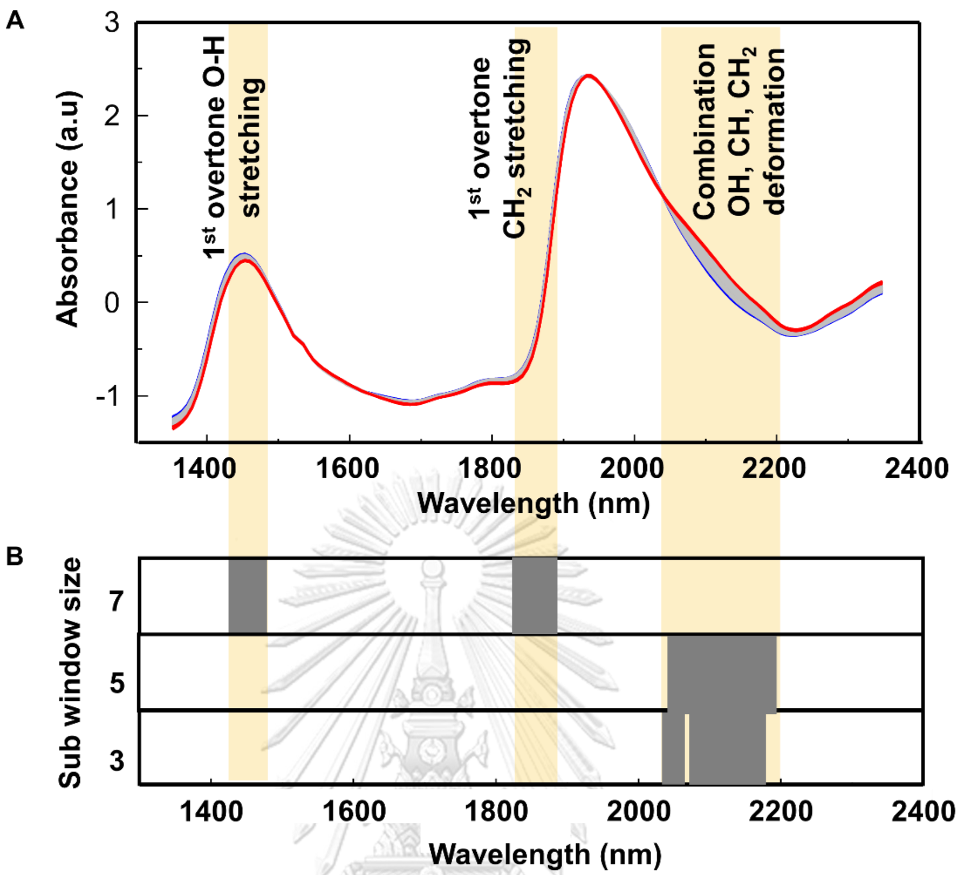


**Figure 4.9 (A) Simulated NIR spectra with 9%w/w additional noise. (B) the selected wavelengths using CSWMPLS (window size of 3, 5, and 7). The yellow highlighted bands represent the sub-regions which are strongly correlate to the prediction of glucose concentrations**

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

RMSE is an index to evaluate the prediction performance of a model. In this section, there are 6 different RMSE indices which are summarized in Table 4.1.

**Table 4.1 Details of 6 different RMSE indices**

| Model | RMSE$_{(\alpha, \beta, \lambda, \omega)}$ | Calibration | Validation | Number of PCs | Selected wavelength |
|---|---|---|---|---|---|
| | | $\alpha$ | $\beta$ | $\lambda$ | $\omega$ |
| I | RMSEC$_{(1st, 1st, full, 0)}$ | 1st | 1st | full | 0 |
| II | RMSECV$_{(1st, 1st CV, full, 0)}$ | 1st | 1st CV | full | 0 |
| III | RMSEP$_{(1st, 2nd, full, 0)}$ | 1st | 2nd | full | 0 |
| IV | RMSEP$_{(1st, 2nd, pc, 0)}$ | 1st | 2nd | pc | 0 |
| V | RMSEP$_{(1st, 2nd, pc, \omega)}$ | 1st | 2nd | pc | $\omega$ |
| VI | RMSEP$_{(2nd, 2nd, full, 0)}$ | 2nd | 2nd | full | 0 |

Note : 1st = Dataset from primary condition
1st CV = Primary set that are estimate by cross validation
2nd = Dataset from secondary condition
full = full spectrum with all wavelengths (without extract any main component)
pc = extract main components using PCA

Description of the model (I-VI)

Model I: RMSEC$_{(1st, 1st, full, 0)}$ is RMSE of prediction when the calibration model was built from a dataset from primary condition. Then, it was used to estimate the responses of a dataset from primary condition.

Model II: RMSECV$_{(1st, 1st CV, full, 0)}$ is RMSE of prediction when the calibration model was built from a dataset from primary condition. Then, it was used to estimate the responses of a dataset from primary condition using Leave-One-Out cross validation.

Model III: RMSEP$_{(1st, 2nd, full, 0)}$ is RMSE of prediction when the calibration model was built from a dataset from primary condition. Then, it was used to estimate the responses of a dataset from secondary conditions (without extract any main components).

Model IV: RMSEP$_{(1st, 2nd, pc, 0)}$ is RMSE of prediction when the calibration model was built from the main components extracted from a dataset from primary condition and then it was used to estimate the responses of a dataset from secondary condition.

Model V: RMSEP$_{(1st, 2nd, pc, \omega)}$ is RMSE of prediction when the calibration model was built from the main components extracted from a dataset from primary condition with only selected wavelengths and then it was used to estimate the responses from secondary condition. This represents our universal model.

Model VI: RMSEP$_{(2nd, 2nd, full, 0)}$ is RMSE of prediction when the calibration model was built from a dataset from secondary condition and directly used to estimate the responses of the secondary condition. This basic idea was represented as a conventional model.

In this study, model VI represents a conventional prediction that a calibration model was built and predicted the samples from the identical system. This will obviously provide the lowest RMSEP compared to model III and model V because these two models were built from primary condition and they were used to predict samples from the other secondary conditions. However, an improvement of the prediction of model V (our universal model) was expected. Figure 4.10 shows a summary of the comparison of RMSE based on 6 different strategies. Generally, RMSEP of all models linearly increases when noise level increases.

From the results, the RMSEP of model VI provides very low value in between 0.006 to 0.49 for noise level 1% w/w to 40 %w/w, respectively. On the other hand, the RMSEP of model III shows the highest value in the range of 0.11 to 6.53 for noise level 1%w/w to 40 %w/w, respectively. This shows that it was not possible to obtain an accurate prediction (low RMSE) when a model was generated using one condition and was used to estimate the responses form the other conditions without any pretreatment. In this case, our proposed strategy was applied as model V to maintain the model from primary condition to accurately estimate the responds from secondary conditions. The RMSEP$_{(1st,\ 2nd,\ pc,\ \omega)}$ of model V was reduced to 0.07 - 4.76 for noise level 1% w/w to 40 %w/w, respectively. It can be seen that our proposed universal model (Model V) can be used to predict an unknown sample from other condition more accurate compared with the model III. In order to visualize more clearly, percent improvement of model V compared to model III was shown as an inset of Figure 4.10.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$\text{Percent improvement} = \left(\frac{RMSEP_{(1st,2nd,full,0)} - RMSEP_{(1st,2nd,pc,\omega)}}{RMSEP_{(1st,2nd,full,0)}}\right) x\ 100$$

It might be seen that the accuracy of prediction from our universal model (model V) has improved up to 30 percent based on noise level of 1-40%w/w. Therefore, it could be indicated that our universal model shows high ability prediction compared to model III. One of this reason, it might be noted that this universal model composed of only variations of glucose without any influences from noise while model III was constructed from full spectrum including noise as interferences, so ability prediction of model V was more corrected. By the way, RMSEP$_{(1st,\ 2nd,\ pc,\ \omega)}$ (orange triangle) of the universal model (model V) still higher than RMSEP$_{(2nd,\ 2nd,\ full,}$

$_{0)}$ of model VI (pink triangle). This observation might originate from influences of interference in the systems because noise level is inversely proportional to amount of glucose. That means the higher noise level, the lower glucose concentration was occurred. Therefore, it causes to prediction accuracy of glucose in our universal model becomes more error.

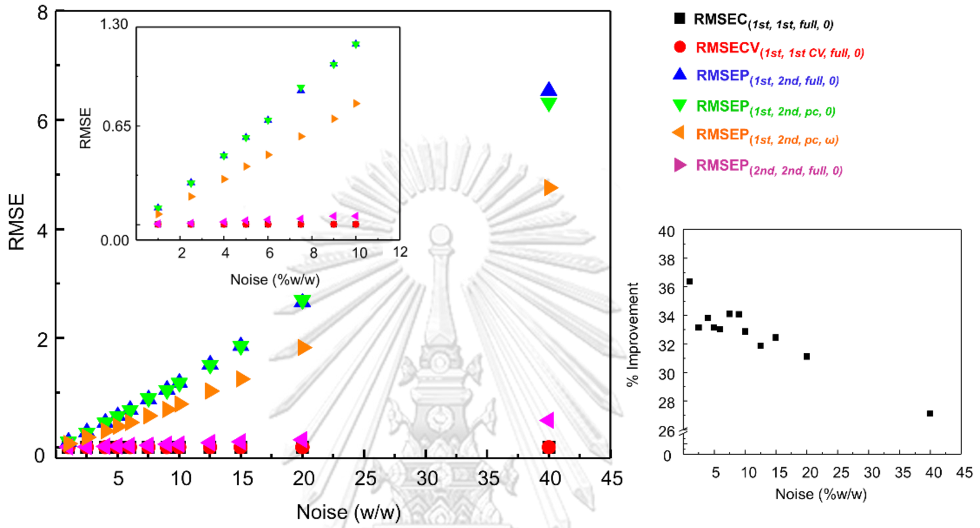


**Figure 4.10 plots of six different strategies including Model I: RMSEC$_{(1st,\ 1st,\ full,\ 0)}$ , Model II: RMSECV$_{(1st,\ 1st\ CV,\ full,\ 0}$, Model III: RMSEP$_{(1st,\ 2nd,\ full,\ 0)}$, Model IV: RMSEP$_{(1st,\ 2nd,\ pc,\ 0)}$, Model V: RMSEP$_{(1st,\ 2nd,\ pc,\ \omega)}$ and Model VI: RMSEP$_{(2nd,\ 2nd,\ full,\ 0)}$. In this case, our proposed universal model was represented as model V.**

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

## 4.4 Prediction of glucose concentration in non-alcoholic drinks

The performance and limitation of our universal model were discussed in the previous section. In this section, our proposed strategies to build a universal model were applied to the real non-alcoholic drinks in order to estimate amount of glucose in the drinks which are tea, cocoa and coffee. The glucose solution using DI water as solvent was defined as primary condition, while the glucose in the drinks was defined as secondary condition. In this study, influent levels from chemical contents of the drinks were set to 2 different levels (high and low). The details of experimental were set up already discussed in section 3.2, chapter 3. Table 4.2 shows the results obtained from multivariate data analysis on experimental details that was calculated using 3 models (A, B and C).

**Table 4.2 Details of the calculation models in low and high level of system using partial least square (PLS) on calibration samples to estimate glucose concentration of validation samples. Number of major components extracted by PCA. Significant wavelengths were selected using CSWMPLS. Number of PLS components was selected by leave one-out cross validation (LOOCV). Root mean square error of calibration (RMSE) index, root mean square error of cross validation (RMSECV) and root mean square error of prediction (RMSEP) were used as index of prediction error. $r^2$**

| | Tea low system | | | Tea high system | | | Cocoa low system | | | Cocoa high system | | | Coffee low system | | | Coffee high system | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | A | B | C | A | B | C | A | B | C | A | B | C | A | B | C | A | B | C |
| Calibration set of samples | tea | water | water | tea | water | water | cocoa | water | water | cocoa | water | water | coffee | water | water | coffee | water | water |
| Validation set of samples | tea | tea | tea | tea | tea | tea | cocoa | cocoa | cocoa | cocoa | cocoa | cocoa | coffee | coffee | coffee | coffee | coffee | Coffee |
| Number of PCs | - | - | 6 | - | - | 6 | - | - | 9 | - | - | 5 | - | - | 6 | - | - | 5 |
| Number of PLS components | 10 | 22 | 5 | 10 | 22 | 4 | 8 | 22 | 4 | 6 | 22 | 4 | 9 | 22 | 5 | 7 | 22 | 5 |
| RMSECV | 0.11 | 0.11 | 0.27 | 0.13 | 0.11 | 0.28 | 0.14 | 0.11 | 0.28 | 0.19 | 0.11 | 0.28 | 0.10 | 0.11 | 0.27 | 0.17 | 0.11 | 0.27 |
| RMSEC | 0.05 | 0.06 | 0.27 | 0.05 | 0.06 | 0.27 | 0.08 | 0.06 | 0.29 | 0.09 | 0.06 | 0.27 | 0.06 | 0.06 | 0.26 | 0.10 | 0.06 | 0.26 |
| RMSEP | 0.92 | 1.12 | 0.72 | 0.59 | 1.39 | 1.39 | 0.69 | 0.96 | 0.99 | 0.73 | 1.87 | 0.84 | 0.90 | 3.28 | 0.54 | 0.33 | 5.64 | 2.80 |
| $r^2$ | 0.9938 | 0.9969 | 0.9977 | 0.9902 | 0.9729 | 0.9941 | 0.9942 | 0.9981 | 0.9965 | 0.9944 | 0.9951 | 0.9941 | 0.9839 | 0.9952 | 0.9940 | 0.9989 | 0.9983 | 0.9976 |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Description of the model (A-C)

Model A: the calibration model was built from a dataset from drinks (tea, cocoa, and coffee system) and directly used to estimate the glucose concentration in the drinks.

Model B: the calibration model was built from a dataset of glucose solution (primary condition). Then, it was used to estimate the glucose concentration in the drinks (tea, cocoa, and coffee system) without extracting any main components.

Model C: the calibration model was built from the main components extracted from a dataset of glucose solution (primary condition) with only selected wavelengths and then it was used to estimate the glucose concentration in the drinks (tea, cocoa, and coffee system). This represents universal model

The model A represents the conventional way that the calibration model of each system (tea, cocoa and coffee in the case) was built to predict glucose concentrations of the system. For example, to predict glucose concentration in tea and cocoa, two calibration models were built from set of samples using tea and cocoa solutions, respectively. In case of model B, it was built from full spectrum without extracted any important component for building the calibration model. In order to improve performance of calibration model, the strategy to generate model C was proposed. For model C, it is represented as a universal model. This involves the determination and extraction of only significant components and wavelengths which are strongly related to the variations of glucose in the system (as mentioned in section 3.6 chapter 3) prior to build the calibration model. For the universal model (model C), it was found that there are 5-9 significant components contributed in NIR spectra of all systems. These components might be related to the patterns of pure NIR spectrum of (i) water, (ii) glucose, interaction between (iii) glucose-water and (iv) glucose-glucose, the other components might be effect by noise which might originate from chemical contents of each system.

RMSE values represent the average prediction error of the model. It can be seen that RMSEC, which represent root mean square error of calibration samples, are quite low in all models (A, B, and C) with 0.05 - 0.29. This suggests that our calculation procedure including optimization of significant components and prediction using PLS is correct and appropriate to quantify glucose concentrations. In case of RMSECV, the average prediction error is slightly higher (0.11-0.28) but it still in the acceptable range. This represents the capability of the model in order to be used to predict the glucose concentration of an unknown solution. Next step, the generated model was used to predict the glucose concentration of the validation set. In case of low system, RMSEP of our universal model (model C) is in range of 0.54 – 0.99 which are lower than model A (conventional model) and model B (using full spectrum). These show approximately 30 and 60 % improvement of the prediction. To make it easier to understand, an example in case of low level of tea system was demonstrated. The RMSEP of 0.92, 1.12 and 0.72 corresponds to model A (conventional way), model B (full spectrum) and model C (our universal model), respectively, they can be seen that RMSEP of model C lower than RMSEP of model

A and B. These observations were also found in low level of cocoa and coffee system. Therefore, it might be concluded that this universal model has a capability to predict concentration of glucose in non-alcoholic drinks without the requirement of building a new calibration model.

To evaluate the limitation of prediction using universal model, the model was used to estimate amount of glucose in more complex system. In this case, more complex system was represented as extremely amount of tea, cocoa and coffee that was dissolved in water in order to prepare as a solvent (20 g in water 100 mL), also called high level system. It seems to be that RMSEP of model C (0.84 – 2.80) lower than RMSEP of model B (1.39 – 5.64) while slightly higher than RMSEP of model A (0.33 – 0.73). One of the main reason could be come from interference which originate from high amount of chemical components in solutions.  From this result, it might be suggested that our universal model has limitation in predicting the amount of glucose due to the interferences from level of chemical content in the solution. These results are in good agreement with the calculations on the simulated datasets in the previous section.

Further investigation, correlation plot between actual values and the predicted value of glucose concentrations in different models were shown in Figure 4.11. It can be seen that the higher glucose concentration, the higher variation on prediction is occurred. This suggests that our universal model (model C) can be able to accurately predict glucose concentration in non-alcoholic drink (tea, cocoa, and coffee) in the same level as using model A (conventional model) especially for low concentration of glucose.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

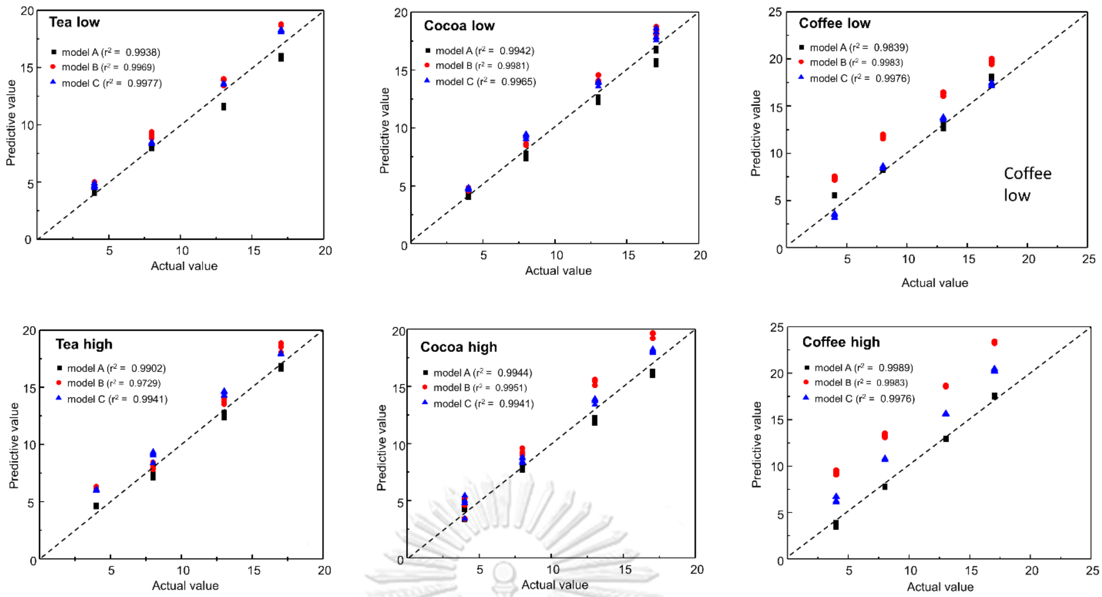


**Figure 4. 11A-F show correlation plots between actual value (X-axis) and predicted value (Y-axis) of glucose concentration in high/low level of tea, high/low level of cocoa, high/low level of coffee, respectively, using calibration model A, B and C**

To examine the selectivity of the prediction, the universal model was used to estimate glucose concentration in the solution mixed with several types of sugar including monosaccharide (fructose, galactose) and disaccharide (sucrose). The calibration model that was constructed from primary condition (only glucose in the solution), and it was used to predict glucose concentration in the mixed sugar solutions. In the solution, the total concentration of sugar is 16%w/w which consist of 8%w/w glucose + 8%w/w of another sugar. For the result (Table 4.3), RMSEP from the prediction of glucose concentrations in all solutions show high error, while RMSEP from the prediction of total sugar shows high accuracy. This result suggests that our universal model is capable to predict total sugar, not for glucose. One possible reason because of all sugar (glucose, fructose, galactose and sucrose in this case) provide similar characteristic of NIR spectra as shown in Figure 4.12

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

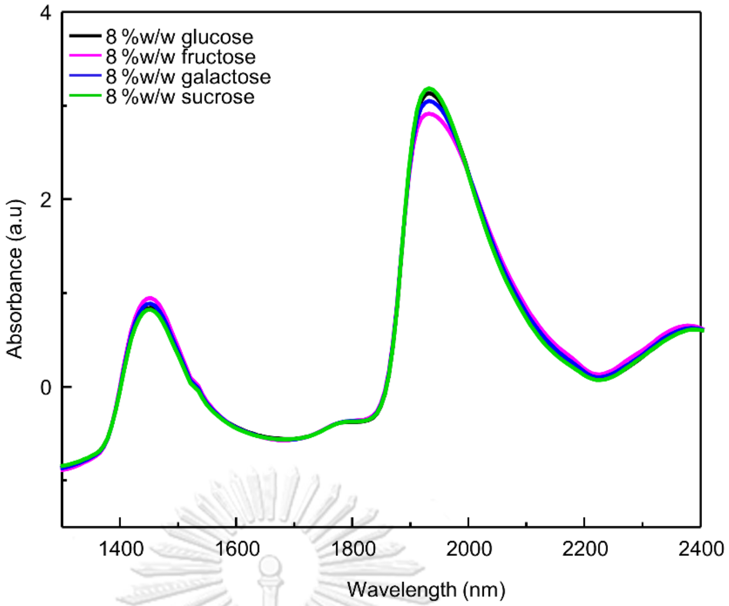


**Figure 4.12 NIR spectra of glucose, fructose, galactose and sucrose solution (8%w/w) in water system following black, pink, blue and green line, respectively.**

From Figure 4.12, it can be seen that overtone patterns of sugars (glucose, fructose, galactose and sucrose) show high similarity. Moreover, variation of intensity of all sugars are not significant different. From these spectra, our analyte (glucose) cannot differentiated from interferences (fructose, galactose and sucrose)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

**Table 4.3 Details of the calculation models in mixture sugar using multivariate data analysis approach (PLS) of calibration sample set, validation sample set. Number of major components extracted by PCA. Number of PLS components was selected by leave one-out cross validation. Root mean square error of calibration (RMSE) index, root mean square error of cross validation (RMSECV) and root mean square error of prediction (RMSEP) were used as index of prediction error.**

| Model | B | C | B | C | B | C |
|---|---|---|---|---|---|---|
| Calibration set of samples | water | water | water | water | water | water |
| Validation set of samples | 8%glucose +8%fructose | 8%glucose +8%fructose | 8%glucose +8%galactose | 8%glucose +8%galactose | 8%glucose +8%sucrose | 8%glucose +8%sucrose |
| Number of PCs | FS | 5 | FS | 8 | FS | 8 |
| Number of PLS components | 22 | 5 | 22 | 6 | 22 | 7 |
| RMSECV | 0.11 | 0.26 | 0.11 | 0.25 | 0.11 | 0.25 |
| RMSEC | 0.06 | 0.26 | 0.06 | 0.24 | 0.06 | 0.24 |
| RMSEP compare with 8%w/w | 8.70 | 9.10 | 8.32 | 8.39 | 7.27 | 7.45 |
| RMSEP compare with 16%w/w | 0.72 | 1.12 | 0.35 | 0.42 | 0.72 | 0.53 |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Description of the model (B-C)

Model B: the calibration model was built from dataset of glucose solution (primary condition). Then, it was used to estimate the glucose concentration in the mixture sugar solution (8%w/w glucose + 8%w/w of others sugar) without extracting any main components.

Model C: the calibration model was built from main components extracted from a dataset of glucose solution (primary condition). Then, it was used to estimate the glucose concentration in the mixture sugar solution (8%w/w glucose + 8%w/w of other sugars)

# CHAPTER V
# CONCLUSIONS

This study has demonstrated the potential of NIR spectroscopy combined with chemometric to quantify amount of glucose in non-alcoholic beverages. A methodology of “universal model” was proposed in order to construct the universal calibration model from primary condition and use the model to predict glucose concentrations in secondary condition (tea, cocoa, and coffee in the case). Three stages of methodology including pre-processing, feature selection and main component extraction were applied to spectral data in order to obtain the universal calibration model. The NIR spectra of the samples were collected using transmittance (*T*) mode in the range of 1350 nm - 2350 nm using path length 0.4 mm, integration time of 1 and 32 averaged scans with smoothing windows of 1. SNV is useful pretreatment for raw data spectra in order to remove background signals. According to characteristic overtones of glucose (3-18%w/w) in water system, it shows dominated peaks of water at 1450 nm and 1950 nm corresponding to 1st vibration overtone O-H stretching and combination of O-H deformation of O-H group, respectively. The characteristic bands of glucose associated to 1st overtone C-H stretching (-$CH_3$ and –$CH_2$-) in the range of 1600 to 1700 nm is unfortunately low intensity, while the band at 2100 nm corresponding to the 1st set of C-H combination band is very strong. These results are consistent with the variance plot. In this study, variation of our universal model was separated into two parts including simulation and experimentation.

In case of simulated part, the noise spectra were controlled by mass balance of 1% - 40 % w/w. Different additional noise levels, difference simulated NIR patterns were occurred. It was found that the spectra with noise level in range of 1 – 12.5 %w/w can be able to maintain the identity pattern of NIR spectra which are in good agreement with the real spectra from the experiments, while the simulated spectra with >15%w/w noise level render to baseline shift. Additionally, influences from different number of inserted samples on the PC selection and model prediction are investigated base on relative percent error of $RMSEP_{\text{pc}}$ using different number of inserted sample compared with single inserted sample. For the result, the $RMSEP_{pc}$ will be increased more than 5% when the number of inserted sample were up to 20 while the $RMSEP_{pc}$ was changed for less than 2% when the number of inserted sample equal to 10. So, it could be suggested that the limitation of inserted number of sample is around 20 samples and the optimal number sample of 10 was selected to further multivariate data analysis. From the analysis, the universal model improves the prediction accuracy for the test set (unseen data) for at least 30 percent.

In case of experimental part, it was used to quantify amount of glucose in non-alcoholic beverages (tea, cocoa and coffee in the case). The promising values for root mean square error of prediction (RMSEP) were obtained to be 0.72, 0.99 and 0.54 corresponding to tea, cocoa and coffee system, respectively. These observations are in good agreement with $r^2$ plot. The $r^2$ value is over 0.99 for the prediction of the data with noise only in the range of 1-12.5%w/w. Therefore, it might be implied that our universal model approach can be used to estimate glucose concentrations in other non-alcoholic drinks without any requirement of a new calibration model. However, the universal model cannot be used to determine glucose selectivity in the solution mixed with other sugar (fructose, galactose and sucrose)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

# REFERENCES

1. Blanco, M.; Villarroya, I., NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* **2002,** *21* (4), 240-250.
2. Nicolaï, B. M.; Beullens, K.; Bobelyn, E.; Peirs, A.; Saeys, W.; Theron, K. I.; Lammertyn, J., Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology* **2007,** *46* (2), 99-118.
3. Roggo, Y.; Chalus, P.; Maurer, L.; Lema-Martinez, C.; Edmond, A.; Jent, N., A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of pharmaceutical and biomedical analysis* **2007,** *44* (3), 683-700.
4. Lanza, E., Determination of moisture, protein, fat, and calories in raw pork and beef by near infrared spectroscopy. *Journal of Food Science* **1983,** *48* (2), 471-474.
5. Segtnan, V.; Isaksson, T., Evaluating near infrared techniques for quantitative analysis of carbohydrates in fruit juice model systems. *Journal of Near Infrared Spectroscopy* **2000,** *8* (2), 109-116.
6. Tewari, J. C.; Dixit, V.; Cho, B.-K.; Malik, K. A., Determination of origin and sugars of citrus fruits using genetic algorithm, correspondence analysis and partial least square combined with fiber optic NIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2008,** *71* (3), 1119-1127.
7. Shiroma, C.; Rodriguez-Saona, L., Application of NIR and MIR spectroscopy in quality control of potato chips. *Journal of Food Composition and Analysis* **2009,** *22* (6), 596-605.
8. Wang, W.; Paliwal, J., Near-infrared spectroscopy and imaging in food quality and safety. *Sensing and Instrumentation for Food Quality and Safety* **2007,** *1* (4), 193-207.
9. Ozaki, Y., Near-infrared spectroscopy—its versatility in analytical chemistry. *Analytical sciences* **2012,** *28* (6), 545-563.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

10. Abdi, H.; Williams, L. J., Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2010,** *2* (4), 433-459.

11. Gowen, A. A.; Downey, G.; Esquerre, C.; O'Donnell, C. P., Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics* **2011,** *25* (7), 375-381.

12. Rambla, F.; Garrigues, S.; De La Guardia, M., PLS-NIR determination of total sugar, glucose, fructose and sucrose in aqueous solutions of fruit juices. *Analytica Chimica Acta* **1997,** *344* (1-2), 41-53.

13. Shahbazikhah, P.; Kalivas, J. H., A consensus modeling approach to update a spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems* **2013,** *120*, 142-153.

14. Du, W.; Chen, Z.-P.; Zhong, L.-J.; Wang, S.-X.; Yu, R.-Q.; Nordon, A.; Littlejohn, D.; Holden, M., Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. *Analytica Chimica Acta* **2011,** *690* (1), 64-70.

15. Fonollosa, J.; Fernández, L.; Gutiérrez-Gálvez, A.; Huerta, R.; Marco, S., Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization. *Sensors and Actuators B: Chemical* **2016,** *236*, 1044-1053.

16. Wang, Y.; Veltkamp, D. J.; Kowalski, B. R., Multivariate instrument standardization. *Analytical chemistry* **1991,** *63* (23), 2750-2756.

17. Wülfert, F.; Kok, W. T.; Noord, O. E. d.; Smilde, A. K., Correction of Temperature-Induced Spectral Variation by Continuous Piecewise Direct Standardization. *Analytical Chemistry* **2000,** *72* (7), 1639-1644.

18. Mou, Y.; Zhou, L.; Yu, S.; Chen, W.; Zhao, X.; You, X., Robust calibration model transfer. *Chemometrics and Intelligent Laboratory Systems* **2016,** *156*, 62-71.

19. Galant, A. L.; Kaufman, R. C.; Wilson, J. D., Glucose: Detection and analysis. *Food Chemistry* **2015,** *188*, 149-160.

20. Agbazue, V.; Ibezim, A.; Ekere, N., Assessment of Sugar levels in Different Soft Drinks. *Journal of Chemical Science* **2014,** *12* (2), 327-334.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

21. Contreras, N. I.; Fairley, P.; McClements, D. J.; Povey, M. J., Analysis of the sugar content of fruit juices and drinks using ultrasonic velocity measurements. *International journal of food science & technology* **1992,** *27* (5), 515-529.
22. Harms, D.; Meyer, J.; Westerheide, L.; Krebs, B.; Karst, U., Determination of glucose in soft drinks using its enzymatic oxidation and the detection of formed hydrogen peroxide with a dinuclear iron(III) complex. *Analytica Chimica Acta* **1999,** *401* (1), 83-90.
23. Akiyama, S.; Nakashima, K.; Yamada, K., High-performance liquid chromatographic determination of sugars in an infusion and soft drinks using a silica-based 3-morpholinopropyl-bonded stationary phase. *Journal of Chromatography A* **1992,** *626* (2), 266-270.
24. Xie, L.; Ye, X.; Liu, D.; Ying, Y., Quantification of glucose, fructose and sucrose in bayberry juice by NIR and PLS. *Food Chemistry* **2009,** *114* (3), 1135-1140.
25. Giangiacomo, R.; Magee, J.; Birth, G.; Dull, G., Predicting concentrations of individual sugars in dry mixtures by near-infrared reflectance spectroscopy. *Journal of Food Science* **1981,** *46* (2), 531-534.
26. Lanza, E.; Li, B., Application for near infrared spectroscopy for predicting the sugar content of fruit juices. *Journal of Food Science* **1984,** *49* (4), 995-998.
27. Liu, Y.; Ying, Y.; Yu, H.; Fu, X., Comparison of the HPLC Method and FT-NIR Analysis for Quantification of Glucose, Fructose, and Sucrose in Intact Apple Fruits. *Journal of Agricultural and Food Chemistry* **2006,** *54* (8), 2810-2815.
28. Luqing, L.; Lingdong, W.; Jingming, N.; Zhengzhu, Z., Detection and quantification of sugar and glucose syrup in roasted green tea using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy* **2015,** *23* (5), 317-325.
29. Huang, H.; Yu, H.; Xu, H.; Ying, Y., Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review. *Journal of Food Engineering* **2008,** *87* (3), 303-313.
30. Aenugu, H. P. R.; Kumar, D. S.; Srisudharson, N. P.; Ghosh, S. S.; Banji, D., Near infra red spectroscopy–an overview. *International Journal of ChemTech Research* **2011,** *3* (2), 825-836.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

31. Pasquini, C., Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian chemical society* **2003,** *14* (2), 198-219.

32. Jolliffe, I. T.; Cadima, J., Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **2016,** *374* (2065), 20150202.

33. Brereton, R. G., *Chemometrics: data analysis for the laboratory and chemical plant*. John Wiley & Sons: 2003.

34. Gowen, A.; Downey, G.; Esquerre, C.; O'Donnell, C., Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics* **2011,** *25* (7), 375-381.

35. Du, Y. P.; Liang, Y. Z.; Jiang, J. H.; Berry, R. J.; Ozaki, Y., Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica Chimica Acta* **2004,** *501* (2), 183-191.

36. Ranzan, C.; Trierweiler, L. F.; Hitzmann, B.; Trierweiler, J. O., NIR pre-selection data using modified changeable size moving window partial least squares and pure spectral chemometrical modeling with ant colony optimization for wheat flour characterization. *Chemometrics and Intelligent Laboratory Systems* **2015,** *142*, 78-86.

37. Amigo, J. M.; Cruz, J.; Bautista, M.; Maspoch, S.; Coello, J.; Blanco, M., Study of pharmaceutical samples by NIR chemical-image and multivariate analysis. *TrAC Trends in Analytical Chemistry* **2008,** *27* (8), 696-713.

38. Niu, X.; Zhao, Z.; Jia, K.; Li, X., A feasibility study on quantitative analysis of glucose and fructose in lotus root powder by FT-NIR spectroscopy and chemometrics. *Food Chemistry* **2012,** *133* (2), 592-597.

39. Seo, S. A review and comparison of methods for detecting outliers in univariate data sets. University of Pittsburgh, 2006.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

# VITA

| | |
|---|---|
| NAME | Sureerat |
| DATE OF BIRTH | 16 September 1993 |
| PLACE OF BIRTH | Phitsanulok |
| HOME ADDRESS | 223/2 Moo.10 Phrompiram, Phrompiram district ,<br>Phitsanulok<br>65150 |
| PUBLICATION | Kikuchi, M., Makmuang, S., Izawa, S., Wongravee, K., & Hiramoto, M. (2019). Doped organic single-crystal photovoltaic cells. Organic Electronics, 64, 92-96.<br>Makmuang, S., Sricharoen, N., Pienpinijtham, P., Ekgasit, S., & Wongravee, K. (2018). Global calibration model for determination of glucose in non-alcoholic beverages using near infrared spectroscopy combined with chemometrics. Proceeding of Pure and Applied Chemistry International Conference 2018, the 60th Anniversary of His Majesty the King’s Accession to the Throne International Convention Center, Hat Yai, Songkhla, Thailand, pp AN22-AN27.<br>Chainok, K., Makmuang, S., & Kielar, F. (2016). Crystal structures of (E)-N′-(2-hydroxy-5-methylbenzylidene) isonicotinohydrazide and (E)-N′-(5-fluoro-2-hydroxybenzylidene) isonicotinohydrazide. Acta Crystallographica Section E: Crystallographic Communications, 72(7), 980-983. |

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University