Chemometric techniques for multivariate calibration and their application in spectroscopic sensors

Ke, Wang

2012

Ke, W. (2012). Chemometric techniques for multivariate calibration and their application in spectroscopic sensors. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/48640>

<https://doi.org/10.32657/10356/48640>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**CHEMOMETRIC TECHNIQUES FOR MULTIVARIATE
CALIBRATION AND THEIR APPLICATION IN
SPECTROSCOPIC SENSORS**

KE WANG

**SCHOOL OF CHEMICAL AND BIOMEDICAL
ENGINEERING**

2012

CHEMOMETRIC TECHNIQUES FOR MULTIVARIATE CALIBRATION AND THEIR
APPLICATION IN SPECTROSCOPIC SENSORS

KE WANG

2012

**CHEMOMETRIC TECHNIQUES FOR MULTIVARIATE
CALIBRATION AND THEIR APPLICATION IN
SPECTROSCOPIC SENSORS**

KE WANG

School of Chemical and Biomedical Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2012

Acknowledgements

Most importantly, I would like to express my sincere and deep gratitude to my supervisors, Prof. Chen Tao and Prof. Yang Yanhui, for their valuable and continuous guidance and support. They are always kind to listen and give advices. Their guidance and encouragement give me great help during my PhD study. Furthermore, they are not just supervisors but friends, and consider for every students' needs. I am very lucky to have so kind and gracious advisors.

My warm acknowledgements also go to School of Chemical and Biomedical Engineering, Nanyang Technological University for offering scholarship support for my PhD study.

I wish to express my sincere thanks to all the members in my research group: Dr Liu Dapeng, Dr Guo Zhen, Dr Zhou Chunmei, Dr Chen Hong, Dr Chen Yuanting, Mr Yan Wenjin, Mr Chi Guoyi, Ms Bai Linlu, Ms Yan Yibo, for their technical support and constructive suggestions. My sincere thanks also go to Prof. Chen Yuan and Ms Si Rongmei, for their cooperation in the project of chemometric determination of the length distribution of single walled carbon nanotubes.

Last but not the least, I would like to give my eternally loving thanks to my parents Mr Wang Zhanbiao and Ms Sun Xiang, for their unconditional support, understanding and encouragement all through my life. My special gratitude also goes to my brothers Mr Wang Xuning and Mr Wang Xucong, for their loving support.

Abstract

Chemometric modeling for multivariate calibration of spectroscopy is a crucial technique to ensure product quality and process performance at low cost in many industries. This technique provides fast, noninvasive and nondestructive analysis of sample/process by predicting analyte properties from measured spectra. Traditional multivariate calibration methods, such as principal component regression (PCR) and partial least squares (PLS), are only reliable when the relationship between analyte properties and spectra is linear. In practice, external disturbances, such as light scattering and baseline noise, will introduce non-linearity into spectral data, deteriorating the prediction accuracy of PCR and PLS. In this thesis, several chemometric strategies will be investigated to address this challenge, including pre-processing and non-linear calibration techniques. Pre-processing methods of the first (D1) and second derivatives (D2), standard normal variate (SNV), extended multiplicative signal correction (EMSC), and extended inverted signal correction (EISC), are proposed to remove the impact of disturbances first, so that the linear calibration method of PCR or PLS can be applied. In addition, a unique linear calibration strategy of optical path length estimation and correction (OPLEC), which involving the building of two linear calibration models, is also investigated. Non-linear calibration techniques aim to model the non-linearity directly, including the methods of artificial neural network (ANN), least squares support vector machine

(LS-SVM), and Gaussian process regression (GPR).

Through comparison of different linear and non-linear calibration techniques, it is found non-linear calibration techniques give more accurate prediction performance than linear methods in most cases. However, non-linear calibration models are not robust enough and small changes in training data or model parameters may result in significant changes in prediction. Therefore, the strategies of bagging/subbagging are investigated to improve the prediction robustness of non-linear calibration models. Furthermore, when using spectroscopic data to predict the analyte property, not all of the variables have contribution to the calibration model. Therefore, selecting the useful variables is effective to improve the prediction performance of calibration models. Two penalized regression algorithms with variable selection using LASSO (least absolute shrinkage and selection operator), penalized linear regression (PLR) and penalized Gaussian process regression (PGPR), are investigated to solve the variable selection problem. Finally, chemometric calibration techniques are applied for solving practical problems which involve predicting the length distribution of single walled carbon nanotubes (SWCNTs) through ultraviolet-visible near-infrared (UV-vis-NIR) spectroscopy and designing a soft sensor to monitor an industrial anaerobic wastewater treatment process.

List of Symbols

Roman letters

a	Dimension of analyte properties.
a_i	Coefficient of additive effect.
A	Absorbance.
$\mathbf{A}$	Matrix of regression coefficients between $\mathbf{Z}_{base}$ and $\mathbf{Z}_{rest}$ in OPLEC.
b_i	Coefficient of multiplicative effect.
$\mathbf{b}$	Vector of regression coefficients of multiplicative effect.
c_k	Concentration of the k^{th} constituent.
$\mathbf{C}$	Covariance matrix.
C_{ik}	Covariance matrix of $\mathbf{x}_i, \mathbf{x}_k$.
d_i	Smooth wavelength-dependent spectral variation in the linear term.
$\mathbf{D}$	Row vector of $[\mathbf{1}; \boldsymbol{\lambda}; \boldsymbol{\lambda}^2]$.
$\mathbf{D}^+$	Pseudo-inverse of $\mathbf{D}$.
$\mathbf{d}_k$	Vector of $\mathbf{s}_k(\mathbf{I} - \mathbf{D}^+\mathbf{D})$ in OPLEC.
$\Delta\mathbf{d}_k$	Vector of $\mathbf{d}_k - \mathbf{d}_2$.
e_i	Smooth wavelength-dependent spectral variation in the binomial term.
$\mathbf{e}_i$	Regression error in LS-SVM.
$\mathbf{E}$	Residuals in NIPALS.

$f(\cdot)$	Linear transfer function in ANN.
$\mathbf{f}_t$	Process future in CVA-SS.
F	State matrix in CVA-SS.
$g(\cdot)$	Sigmoid transfer function in ANN.
G	Input matrix in CVA-SS.
H	Number of principal components in PCR or latent variables in PLS.
H	Output matrix in CVA-SS.
I	Number of samples in spectra.
$\mathbf{I}$	Identity matrix.
J	Number of variables in spectra.
K_x	Number of the past input required to model the present output in ARX.
K_y	Number of the past output required to model the present output in ARX.
l	Path length travelled by photons.
L	Labeled dataset.
m	Serial number of randomly partition of a dataset.
$\mathbf{m}$	Reference spectra.
M	Number of randomly partition of a dataset.
N	Number of the hidden-layer neurons in ANN.
N	Number of the possible selected model parameters in wastewater treatment.
$\mathbf{p}_t$	Process past in CVA-SS.
$\mathbf{P}$	Loading matrix of $\mathbf{X}$.

q	Time lag in autocorrelation.
Q	Cost function in LS-SVM.
$\mathbf{Q}$	Loading matrix of $\mathbf{y}$.
r	Serial number of calibration models in bagging.
r_h	Autocorrelation coefficient.
r_m	RMSEP of a calibration method tested on the m^{th} partition.
$\bar{r}$	Mean of r_m .
$\tilde{r}_m$	$r_m - \bar{r}$.
R	Number of calibration models in bagging.
R_p	Pearson's correlation coefficient.
R_s	Spearman's correlation coefficient.
$\mathbf{s}_k$	Pure absorption spectra of the k^{th} constituent.
s_m	RMSEPs of a calibration method tested on the m^{th} partition.
$\bar{s}$	Mean of s_m .
$\tilde{s}_m$	$s_m - \bar{s}$.
$\mathbf{t}_h$	Vector of score matrix.
T	Transmission matrix in CVA-SS.
T_c	Transmission.
$\mathbf{T}$	Score matrix of $\mathbf{X}$.
U	Unlabeled dataset.
$\mathbf{U}$	Score matrix of $\mathbf{y}$.

v_j	Weight connecting the input- and hidden-layer neurons in ANN.
w_j	Regression parameter in GPR.
$\mathbf{w}$	Vector of model parameters in LS-SVM.
$\mathbf{w}$	Vector consisting of w_j .
x_{ij}	Predictor variable of the i^{th} sample at the j^{th} wavelength.
$\mathbf{x}_i$	Measured spectra of the i^{th} sample.
$\mathbf{x}_{i,chem}$	Theoretical spectra of the i^{th} sample.
$\mathbf{x}_{i,corr}$	Corrected spectra of the i^{th} sample in pre-processing techniques.
$\mathbf{X}$	Matrix of spectroscopic data.
X	Influent properties of wastewater.
$\mathbf{y}$	Vector of analyte property.
y_i	Analyte property of the i^{th} sample.
y_{ik}	Analyte property of the k^{th} constituent in the i^{th} sample.
Y	Effluent TOC of wastewater.
z	System states in CVA-SS.
$\mathbf{z}_i$	Vector of standardized spectra in OPLEC.
$\mathbf{Z}$	Matrix of standardized spectra in OPLEC.

Greek letters

α	Ridge constant in Ridge regression.
----------	-------------------------------------

β	Vector of regression coefficients in linear regression algorithms.
β_1	Regression vector used to predict the concentration in OPLEC.
β_2	Regression vector used to predict the concentration in OPLEC.
γ	Relative weight regression error in LS-SVM.
δ_i	Unknown and irrelevant variation in MSC.
δ_{ik}	Random error in GPR.
ε_n	Bias in the output layers in ANN.
ϵ_i	Residual spectra of the i^{th} sample.
o	Process errors with covariance matrix S in CVA-SS.
ρ	Process errors with covariance matrix R in CVA-SS.
σ	Standard deviation of SWCNTs length.
σ_X	Standard deviation of influent properties X .
σ_Y	Standard deviation of effluent TOC Y .
$\sigma_{a.k}$	Absorption cross section of the k^{th} chemical species.
$\sigma_{ext.k}$	Extinction cross section of the k^{th} chemical species.
$\sigma_{s.k}$	Scattering cross section of the k^{th} chemical species.
η	Vector of regression coefficients in PLR.
θ	Vector of hyper-parameters in GPR.
κ	Lagrange parameter.
λ	Wavelength of light.
λ	Vector of wavelengths.

μ	Mean of SWCNTs length.
μ_X	Mean of influent properties X .
μ_Y	Mean of effluent TOC Y .
ϕ_j	Bias in the hidden layers in ANN.
φ_t	$\varphi_t = (y_{t-1}, \dots, y_{t-K_y}, x_{t-1}, \dots, x_{t-K_x})$ in ARX.
$\boldsymbol{\varphi}$	Matrix with the elements of φ_t .
ω_n	Weight connecting the hidden- and output-layer neurons in ANN.

List of Figures

Figure 1.1 (a) Homogeneous sample (e.g. liquid mixture), absorption only. (b) Turbid sample with a high concentration of scatters: absorption and multiple scattering. (This figure is adapted from [9].)	2
Figure 3.1 RMSEPs of different calibration techniques for (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.	42
Figure 3.2 The standard deviations of RMSEPs for different calibration techniques: (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.....	46
Figure 3.3 RMSEPs of combining pre-processing and non-linear calibration models for (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.	47
Figure 4.1 Prediction errors of single (a) GPR (b) ANN and (c) PLS model, and bagging (d) GPR (e) ANN and (f) PLS model on “wheat kernels” set.....	53
Figure 4.2 RMSEPs of subbagging GPR, ANN and PLS at different sub-sampling ratios on “wheat kernels” dataset.....	56
Figure 5.1 The regression parameters in EMSC-PLS.....	66
Figure 5.2 The regression parameters $\boldsymbol{\eta}$ in (a) EMSC-PLR; (b) EMSC-PLR with variable selection.	67
Figure 5.3 The hyper-parameters $\boldsymbol{w}$ in (a) EMSC-PGPR; (b) EMSC-PGPR with variable selection.	68
Figure 6.1 UV-vis-NIR spectra of length sorted SWCNT fractions after scaling at 775	

nm. Note that “1f”, corresponding samples from the top of the centrifugation tube, contains few nanotubes, and thus its spectrum is not shown. The absorption spectrum of the starting material before separation (dashed line and denoted “AP”) is shown as a reference.	79
Figure 6.2 The length distribution of sorted SWCNT samples obtained from AFM analysis. The red solid line represents the fitted log-normal distribution. The bars represent the histogram from AFM measure.	80
Figure 6.3 The spectra before and after pre-processing by EISC.	81
Figure 6.4 LOO-CV calibration results where the log-mean is predicted by GPR and log-standard deviation by EISC. The red solid line represents the reference fitted distribution, while the blue dotted line indicates the predicted distribution. The bars represent the histogram from AFM measure.	84
Figure 6.5 LOBO-CV calibration results where the log-mean is predicted by GPR and log-standard deviation by EISC.	86
Figure A.1 The process of anaerobic wastewater treatment.	99
Figure A.2 The measured and predicted data of PLS.	111
Figure A.3 The autocorrelation of effluent TOC.	112
Figure A.4 The measured and predicted data of (a) ARX-PLS, (b) ARX-ANN, (c) ARX-GPR, and (d) CVA-SS.	114

List of Tables

Table 3.1 The number of selected components when using PLS in pre-processing or non-linear calibration techniques for (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture. The values are averaged over 50 repeated random partitions of data.....	39
Table 3.2 The p -values of paired t-test for different calibration techniques on (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.....	40
Table 4.1 Results of individual calibration methods (GPR, ANN, PLS) and those of subagging with different ratios of sub-sampled training data on “wheat kernels” dataset.	59
Table 4.2 A summary of results for individual calibration methods, bagging, and subagging by cross-validation on “wheat kernels” dataset.....	60
Table 4.3 Summary of results for individual calibration methods (GPR, ANN, PLS) and bagging GPR, ANN, PLS on “meat” dataset: (a) Moisture; (b) Fat; (c) Protein ..	61
Table 5.1 RMSEPs of EMSC, EMSC-PLR, EMSC-PLR-VS, EMSC-PGPR, and EMSC-PGPR-VS.....	70
Table 6.1 The LOO-CV calibration results.....	82
Table 6.2 The LOBO-CV calibration results..	87
Table A.1 Nomenclature	98
Table A.2 The selected 22 tags and their corresponding process parameters which	

contribute to the mathematical prediction model.....	106
Table A.3 The number of samples of the low frequency parameters from 2004 to 2009.	107
Table A.4 Correlation coefficients of influent properties of T-01 and change of TOC.	109
Table A.5 The relative root mean square error of prediction (RRMSEP) of different techniques.	115
Table A.6 Effluence of the past n days on RRMSEP for ARX-PLS, ARX-ANN, and ARX-GPR.....	115

List of Abbreviations

AFM	Atomic Force Microscope
ANN	Artificial Neural Network
API	Active Pharmaceutical Ingredient
ARD	Automatic Relevance Determination
ARX	AutoRegressive with eXogenous
BOD	Biological Oxygen Demand
COD	Chemical Oxygen Demand
CPDS	Continuous Piecewise Direct Standardization
CVA-SS	Canonical Variate Analysis based State Space
D1	The First Derivative
D2	The Second Derivative
DGU	Density Gradient Ultracentrifuge
EGSB	Expanded Granular Sludge Bed
EISC	Extended Inverted Signal Correction
EMSC	Extended Multiplicative Signal Correction
GPR	Gaussian Process Regression
HPLC	High Performance Liquid Chromatography
ISC	Inverted Signal Correction
LSS	Loading Space Standardization

LS-SVM	Least Squares Support Vector Machine
LVs	Latent Variables
MLLS	Minimum Length Least Squares
MLR	Multiple Linear Regression
MSC	Multiplicative Signal Correction
NIPALS	Non-linear Iterative Partial Least Squares
NIR	Near-Infrared
OPLEC	Optical Path Length Estimation and Correction
PCA	Principal Component Analysis
PCR	Principal Component Regression
PGPR	Penalized Gaussian Process Regression
PLR	Penalized Linear Regression
PLS	Partial Least Squares
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RMSEP	Root Mean Square Error of Prediction
RRMSE	Relative Root Mean Square Error
RRMSEP	Relative Root Mean Square Error of Prediction
SNV	Standard Normal Variate
SVD	Singular Value Decomposition
SVM	Support Vector Machine

SWCNTs	Single Walled Carbon NanoTubes
TEM	Transmission Electron Microscope
TKN	Total Kjeldahl Nitrogen
TOC	Total Organic Carbon
TSS	Total Suspended Solids
UVE	Uninformative Variable Elimination
VFA	Volatile Fatty Acids
VS	Variable Selection
WW	WasteWater

Contents

Acknowledgements.....	I
Abstract.....	II
List of Symbols.....	IV
List of Abbreviations.....	XIV
Contents	XVII
Chapter 1 Introduction	1
1.1 Chemometric spectroscopic multivariate calibration.....	1
1.2 Multivariate calibration techniques in the presence of external disturbances ...	3
1.3 Significance of spectroscopic multivariate calibration.....	7
1.4 Contributions of this thesis	9
Chapter 2 Literature Review and Research Methodology	10
2.1 Pre-processing techniques to deal with external variations	15
2.1.1 The first and second derivatives.....	15
2.1.2 Standard normal variate	16
2.1.3 Multiplicative signal correction and its extension.....	17
2.1.4 Inverted signal correction and its extension.....	19
2.1.5 Optical path length estimation and correction.....	19
2.2 Non-linear calibration techniques	22
2.2.1 Artificial neural network	22

2.2.2 Gaussian process regression.....	23
2.2.3 Least-square support vector machine.....	25
2.3 Ensemble modeling for robust calibration.....	26
2.4 Penalized regression with variable selection	29
2.4.1 Penalized linear regression with LASSO.....	30
2.4.2 Penalized Gaussian process regression with LASSO	31
Chapter 3 Comparative Study on Multivariate Calibration Methods to Correct Light Scattering Effect.....	34
3.1 Datasets	35
3.2 Statistical evaluation of calibration methods	36
3.3 Software	38
3.4 Results and discussion	39
3.5 Summary	48
Chapter 4 Bagging for Robust Non-linear Multivariate Calibration	49
4.1 Datasets	50
4.2 Statistical evaluation of bagging/subbagging	51
4.3 Results and discussion	53
4.5 Summary	57
Chapter 5 Penalized Regression Methodology with Variable Selection.....	63
5.1 Computation issues	65
5.2 Results and discussion	66

5.3 Summary	70
Chapter 6 Chemometric Determination of the Length Distribution of Single Walled Carbon Nanotubes.....	71
6.1 Introduction.....	71
6.2 Dataset.....	74
6.3 Chemometric prediction of the length distribution.....	75
6.4 Results and discussion	78
6.5 Summary	90
Chapter 7 Conclusions and Future Work	92
7.1 Conclusions of current work.....	92
7.2 Perspectives for future work	93
Appendix A Soft Sensor Design for an Industrial Anaerobic Wastewater Treatment Process	96
A.1 Introduction.....	96
A.2 The anaerobic wastewater treatment system.....	99
A.3 Methodology: dynamic regression algorithms.....	100
A.3.1 Canonical variate analysis based state space.....	101
A.3.2 AutoRegressive with eXogenous model	102
A.4 Data review and pre-treatment.....	103
A.5 Prediction results of dynamic techniques.....	111
A.6 Summary	116

References.....	117
-----------------	-----

Chapter 1 Introduction

1.1 Chemometric spectroscopic multivariate calibration

Continuous improvement of process performance and product quality at minimal cost is of paramount value to the process industry [1]. To achieve this objective, it is crucial to infer the values of key process variables (analyte properties¹) which are related to process performance and product quality [2]. A useful way to predict these properties of interest is through chemometric analysis of directly measured process variables, such as pressure, flow, temperature and infrared, Raman, or mass spectroscopy [3]. The relationship between analyte properties and measurements can be described by a mathematical conversion formula (the prediction equation). The choice of the mathematical formula and determination of the parameters in the formula form the task of “calibration”. Recently, spectroscopy in combination with calibration models has attracted increasing attention for process monitoring in pharmaceutical, petrochemical and food sectors [4-6].

Calibration has two branches: univariate and multivariate calibration. A univariate calibration model utilizes a single measurement at a chosen wavelength to determine the analyte property. Typically the wavelength corresponding to a peak maximum of

¹ Throughout this thesis, the term, analyte property, is adopted to refer to the variables (e.g. concentration) predicted from direct measurements (e.g. spectral absorbance).

spectra is selected [7]. However, univariate calibration is applicable when only the analyte of interest contributes to measured spectra [3]. If other substances also influence measured signals, the prediction will be biased since the significant peaks of these substances could overlap with each other [8]. Yet in practice it is difficult to fulfill the condition that only the analyte of interest contributes to measured spectra, unless samples are purified and stabilized prior to analysis. In contrast, multivariate calibration utilizes all measurements, and thus it is applicable even if other analytes affect the measurements of the target analyte. Therefore, multivariate calibration is superior to univariate models as a result of noise reduction and interference removal [7].

Figure 1.1 (a) Homogeneous sample (e.g. liquid mixture), absorption only. (b) Turbid sample with a high concentration of scatters: absorption and multiple scattering; adapted from [9].

In spectroscopic multivariate calibration, spectra measured at hundreds of wavelengths are capable of capturing the chemical/physical properties of samples being analyzed [10]. According to Beer-Lambert's law, when light passes through a homogeneous liquid mixture, as shown in Figure 1.1(a), the relationship between spectra and concentration (a typical analyte property to be predicted from spectroscopy) can be described as [9]

$$A(\lambda) = -\ln(T_c) = l \sum_{k=1}^K \sigma_{a,k}(\lambda) c_k \quad (1.1)$$

where A is the absorbance, T_c is the transmission, $\sigma_{a,k}$ is the absorption cross section of the k^{th} constituent, c_k is the concentration, λ is the wavelength of light, and l is the path length travelled by photons. According to Eq. (1.1), concentration depends linearly on spectra since the absorption cross section $\sigma_{a,k}$ and path length l are constant in homogeneous liquids. Therefore, linear calibration models can be developed to predict concentration from measured spectra. Principal component regression (PCR) [11] and partial least squares (PLS) [12, 13] are traditional linear multivariate calibration techniques used for building prediction models.

1.2 Multivariate calibration techniques in the presence of external disturbances

In practice, non-linear variations can be induced to spectra especially in the presence of external disturbances, making linear multivariate calibration techniques

giving inferior prediction performance. External disturbances include light scattering effect [14], temperature and pressure variation [15, 16], instrumental variation of background noise and baseline drift [17, 18]. Light scattering effect is one of the major sources of non-linearity, and will be mainly considered in this thesis.

Light scattering results from sample-to-sample variation in the optical path length, and this variation is typically caused by the differences in particle size, shape, distribution and concentration. Light scattering plays a significant role when spectroscopic instruments are applied to analyze heterogeneous samples, such as suspension and powder [14, 19]. When light passes through a heterogeneous sample as shown in Figure 1.1(b), Beer-Lambert's law needs to be modified as [9]

$$A(\lambda) = -\ln(T_c) = l \sum_{k=1}^K \sigma_{ext,k}(\lambda) c_k \quad (1.2)$$

where $\sigma_{ext,k} = \sigma_{a,k} + \sigma_{s,k}$ is the extinction cross section and $\sigma_{s,k}$ is the scattering cross section, which is a highly non-linear function of particle size and shape. In this case, the average path length travelled by photons and extinction cross section vary, and non-linear variations are induced between the concentration and spectra. The main source of variation in absorption is the path length of photons.

Several strategies have been proposed to account for light scattering effect to enhance calibration modeling. One possibility is to improve spectrometers by using advanced *hardware* configuration based on the principle of light propagation [20]. Alternatively, light scattering effect can be corrected or compensated using chemometric (*software*) techniques [21]. In principle, combining the hardware and

software strategies is expected to give better results than the use of either strategies [22]. Hardware solutions appear to be desired since they are based on fundamental physical principles. However, they may not be compatible with spectroscopic instruments currently used in practice, and thus significantly extra cost is incurred. Therefore, in this thesis, the focus is on chemometric methods to deal with light scattering effect for calibration modeling, which is more cost-effective for existing instruments.

Several chemometric approaches have been proposed in literatures to handle light scattering effect in spectra. One approach is the use of pre-processing methods to remove non-linear variations within spectral data prior to applying PLS, such as the first (D1) and second derivatives (D2) [17, 23], standard normal variate (SNV) [24], extended multiplicative signal correction (EMSC) [21], extended inverted signal correction (EISC) [25]. In addition, a special linear calibration strategy of optical path length estimation and correction (OPLEC) [26], which involving the building of two linear calibration models, is also devised to deal with light scattering. The second approach is to directly model non-linearity in spectral data, using non-linear techniques such as artificial neural network (ANN) [27-29], least squares support vector machine (LS-SVM) [30, 31], and Gaussian process regression (GPR) [10]. The third approach is to select variables (wavelengths) that are insensitive to light scattering, and only the selected variables are employed in the calibration model while others are removed [32, 33].

Although non-linear calibration models (ANN, GPR, LS-SVM) are capable of modeling the non-linearity directly, their predictive performances are not robust in the sense that small changes in data or model parameters can result in significant changes in model predictions. Thus, an effective way to improve the predictive accuracy and robustness of non-linear calibration methods simultaneously is required. Bagging (“**bootstrap aggregating**”) [34, 35] and its alternative of subagging (“**sub-sample aggregating**”) [36] have been demonstrated to work especially well for non-robust models, and will be studied for non-linear calibration methods in this thesis.

Usually the dimension of spectroscopic data is very high (about several hundred). However, not all of the variables contribute to multivariable calibration model. Some variables could be significantly useful while others may be nonsense or even detrimental. Therefore, selecting the dominant variables is helpful to improve the prediction performance of calibration models. However, traditional variable selection techniques, such as uninformative variable elimination (UVE) [37], are computationally expensive, and cannot deal with regression parameter estimation at the same time. Accordingly, two penalized regression algorithms with variable selection using LASSO (least absolute shrinkage and selection operator), penalized linear regression (PLR) [38] and penalized Gaussian process regression (PGPR) [39], will be investigated.

1.3 Significance of spectroscopic multivariate calibration

Usually, analyte properties of interest are directly measured by reference analytical methods (e.g. high performance liquid chromatography (HPLC) and acid-base titration), which are slow, expensive and off-line. However, spectroscopic techniques, such as NIR and Raman, possess the following advantages: (1) they are nondestructive and noninvasive; (2) special sample preparation is not required; (3) the analysis is fast and a large number of measurements can be obtained in a short period of time; and (4) they can be used for in-line and on-line process monitoring. However, spectra only give indirect, non-specific information about analyte properties. By multivariate calibration, it is possible to convert the relatively rapid, precise, but non-specific spectra into the information of analyte properties. Therefore, the significance of spectroscopy combined with multivariate calibration becomes indisputable.

Spectroscopic multivariate calibration is especially useful when analyte properties are difficult to be measured directly or should be determined by laborious off-line analytical methods. Furthermore, since many variables are measured in multivariate calibration, scientific understanding about samples can be obtained and accurate detection of analyte properties is provided. Consequently, spectroscopy combined with multivariate calibration has become more and more important in pharmaceutical, petrochemical and food industries to predict analyte properties to ensure product

quality and process performance [33, 40].

Spectroscopy combined with multivariate calibration can be applied to infer analyte properties (e.g. concentration) rapidly and accurately [40]. For example, in pharmaceutical industry, instead of the expensive, slow and complicated HPLC, spectroscopic multivariate calibration can be used to measure the content of active pharmaceutical ingredient (API) in drug to determine whether product quality satisfies the specification.

On-line process monitoring and quality control are significant in industries, since monitoring process through off-line measurements does not give real-time information and may be destructive to the product. For example, drying is an essential unit operation in pharmaceutical and other chemical industries, where the determination of drying end point is crucial since both over-drying and inadequate drying are detrimental to the product quality. Therefore, it is important to monitor moisture (or other solvent) content to find the optimal drying end point. However, the measurement by reference analytical method is off-line, and requires sampling of the wet cake, which may result in contamination to the product. In addition, time delay in off-line analysis will lead to over-drying. Consequently, the rapid and noninvasive on-line measurement of spectroscopic multivariate calibration is highly desirable [41]. This technique can also be applied to the monitoring of other processes, such as crystallization [42, 43] and polymerization [44, 45].

1.4 Contributions of this thesis

The primary focus of this thesis is to explore different chemometric techniques of pre-processing, non-linear calibration to handle the non-linearity of spectral data induced by external disturbances such as light scattering and baseline noise. Through these techniques, the prediction accuracy of calibration models is expected to be improved. In addition, model robustness of non-linear calibration techniques is improved through bagging/subbagging so that non-linear calibration models are less sensitive to small changes in data and/or model parameters. Furthermore, penalized regression algorithms of PLR and PGPR are studied, in which variable selection is solved when estimating model parameters. Finally, chemometric multivariate calibration techniques are applied to predict the distribution of single walled carbon nanotubes (SWCNTs) through ultraviolet-visible near-infrared (UV-vis-NIR) spectroscopy to validate their effectiveness. Since multivariate calibration cannot only relate spectroscopic data to analyte properties, but also account for the relationship between measured process variables (such as pressure, flow, and temperature) and analyte properties. Therefore, an interesting complement of the application of multivariate calibration is introduced in appendix A to design a soft sensor to monitor an industrial anaerobic wastewater treatment process through measured process variables.

Chapter 2 Literature Review and Research Methodology

The objective of this thesis is to study different chemometric multivariate calibration techniques to predict the analyte property through spectroscopic data. Therefore, these calibration techniques are explained in detail in this chapter.

Traditional multivariate calibration models rely on multiple linear regression (MLR). Therefore, the theory of MLR is elucidated first. Let the matrix $\mathbf{X}$ with dimensions $I \times J$ be spectral measurements in which the element x_{ij} denotes the recorded spectra of the i^{th} ($i = 1, \dots, I$) sample at the j^{th} ($j = 1, \dots, J$) wavelength (variable), and $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ij}]$. Likewise, let $\mathbf{y}$ with dimensions $I \times 1$ represent the vector of the response analyte property in which the element y_i corresponds to the i^{th} sample. A linear calibration model can be developed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (2.1)$$

where $\boldsymbol{\beta}$ is a $J \times 1$ vector of regression coefficients and $\mathbf{e}$ denotes the residuals.

The coefficient vector can be estimated by “least squares” method as

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.2)$$

Unfortunately, in the practice of multivariate calibration, MLR often fails for two reasons. First, the number of wavelengths J is usually larger than the number of samples I . Second, two or more variables in $\mathbf{X}$ are highly correlated, resulting in the colinearity problem [46]. In either case, $(\mathbf{X}'\mathbf{X})^{-1}$ is ill-conditioned and it is hard or

inconvenient to apply Eq. (2.2) to obtain the regression coefficients β . Principal component regression (PCR) [11] and partial least squares (PLS) [12, 13] are usual solutions to this ill-conditioned problem. Both PCR and PLS have been proposed to reduce the number of wavelengths by projecting the matrix $\mathbf{X}$ onto a lower dimensional score matrix $\mathbf{T}$ ($I \times H$), where $H \leq J$ and $\mathbf{T}$ still carries relevant information about $\mathbf{X}$.

The regulation method, *e.g.* Ridge regression (RR), is also a typical approach to deal with ill-conditioned data. In RR, a constant is added to the “ridge” of the covariance matrix to form the pseudoinverse, and the regression coefficients can be defined as

$$\beta = (\mathbf{X}'\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \quad (2.3)$$

where α is ridge constant, $0 < \alpha < 1$, and $\mathbf{I}$ is the identity matrix. The optimum value of α can be determined through cross-validation. However, RR is often used when it is concerned with the values of the regression coefficients themselves, rather than in prediction. Since this thesis mainly investigates the predictive models, RR is not interpreted in detail. Next, the theories of traditional multivariate calibration methods, PCR and PLS, are explained.

Principal Component Regression

Principal component regression (PCR) developed from principal component analysis (PCA). In PCA, $\mathbf{X}$ is decomposed into scores of $\mathbf{T}$ and loadings of $\mathbf{P}$ as

$$\mathbf{X} = \mathbf{T}\mathbf{P}' \quad (2.4)$$

$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_H]$ is a $I \times H$ matrix, and $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_H]$ is a $J \times H$ matrix, where H is the number of principal components. Usually, the algorithm of non-linear iterative partial least squares (NIPALS) is used to calculate $\mathbf{T}$ and $\mathbf{P}$ [13]. Firstly, $\mathbf{t}_1$ and $\mathbf{p}'_1$ are calculated. Then the residual ($\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}'_1$) of subtracting $\mathbf{t}_1\mathbf{p}'_1$ from $\mathbf{X}$ is obtained and used to calculate $\mathbf{t}_2$ and $\mathbf{p}'_2$, and so on. In this way, $\mathbf{T}$ and $\mathbf{P}'$ are calculated pair-by-pair. The procedures of NIPALS algorithm is as follows:

- (1) Take a vector $\mathbf{t}_h$ which is $\mathbf{x}_j$ from $\mathbf{X}$: $\mathbf{t}_h = \mathbf{x}_j$
- (2) Calculate $\mathbf{p}'_h$: $\mathbf{p}'_h = \mathbf{t}'_h\mathbf{X}/\mathbf{t}'_h\mathbf{t}_h$
- (3) Normalized $\mathbf{p}'_h$ to length 1: $\mathbf{p}'_{h\ new} = \mathbf{p}'_{h\ old} / \|\mathbf{p}'_{h\ old}\|$
- (4) Calculate $\mathbf{t}_h$: $\mathbf{t}_h = \mathbf{X}\mathbf{p}_h/\mathbf{p}'_h\mathbf{p}_h$
- (5) If $\mathbf{t}_h$ used in step 1 and $\mathbf{t}_h$ obtained from step 4 are the same, stop the algorithm.

If not, go to step 2.

After the first component is calculated, $\mathbf{X}$ in step 2 and 4 are taken place by the residual $\mathbf{E}_h = \mathbf{E}_{h-1} - \mathbf{t}_h\mathbf{p}'_h$.

In PCR, the representative information of $\mathbf{X}$ is transformed in the score matrix $\mathbf{T}$ as in PCA. Then $\mathbf{T}$ can be used in MLR to replace $\mathbf{X}$. Therefore, it can be rewritten as $\mathbf{y} = \mathbf{T}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\beta} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$.

Partial Least Squares

Besides the property of transforming the information of $\mathbf{X}$ to score matrix of $\mathbf{T}$

which is the same as in PCR, partial least squares (PLS) also possess some new features. The model of PLS consists of two relations: outer relation and inner relation.

1. The outer relation of $\mathbf{X}$ is the same as PCR as

$$\mathbf{X} = \mathbf{T}\mathbf{P}' \quad (2.5)$$

Similar to $\mathbf{X}$, $\mathbf{y}$ also can be decomposed into the matrixes of score $\mathbf{U}$ and loading $\mathbf{Q}$.

Therefore, the outer relation of $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{U}\mathbf{Q}' \quad (2.6)$$

$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_H]$ is a $I \times H$ matrix, and $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_H]$ is a $a \times H$ matrix, where a is the dimension of $\mathbf{y}$.

2. The inner relation describes the relationship between the score matrixes of $\mathbf{T}$ and $\mathbf{U}$ as

$$\mathbf{U} = \mathbf{T}\boldsymbol{\beta} \quad (2.7)$$

where $\boldsymbol{\beta}$ is the regression coefficients. Accordingly, from the outer and inner relations, the following relation can be obtained as $\mathbf{y} = \mathbf{T}\boldsymbol{\beta}\mathbf{Q}'$.

The NIPALS algorithm is also applied in PLS. Instead of being calculated separately, scores and loadings of $\mathbf{X}$ and $\mathbf{y}$ are calculated simultaneously. In order to improve the inner relation, scores of $\mathbf{X}$ and $\mathbf{y}$ are exchanged. Then the algorithm is modified as follows:

- (1) Find $\mathbf{u}_{start} = \mathbf{y}$;
- (2) $\mathbf{p}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$;
- (3) $\mathbf{p}'_{new} = \mathbf{p}'_{old}/\|\mathbf{p}'_{old}\|$;

$$(4) \mathbf{t} = \mathbf{Xp}/\mathbf{p}'\mathbf{p};$$

$$(5) \mathbf{q}' = \mathbf{t}'\mathbf{y}/\mathbf{t}'\mathbf{t};$$

$$(6) \mathbf{q}'_{new} = \mathbf{q}'_{old}/\|\mathbf{q}'_{old}\|;$$

$$(7) \mathbf{u} = \mathbf{yq}/\mathbf{q}'\mathbf{q};$$

(8) If $\mathbf{t}$ obtained in Step 4 is the same as that in the preceding iteration step, stop; else go to Step 2.

Both in PCR and PLS, the number of principle components or latent variables (the dimension of $\mathbf{T}$) are determined by cross-validation. Comparatively, PLS is more efficient than PCR in many practical cases in the sense that its prediction performance is better, since more variations in $\mathbf{X}$ are explained when the same number of principle components or latent variables are included. This is because when reducing the dimension, both $\mathbf{X}$ and $\mathbf{y}$ are considered in PLS, while only $\mathbf{X}$ is considered in PCR.

However, PCR and PLS only give satisfactory performance under linear situations. When external disturbances such as light scattering are present, non-linearity can be induced in spectral data. To address this issue, various chemometric strategies have been contrived to improve prediction accuracy and model robustness. This chapter mainly presents the development and theory of pre-processing techniques to account for non-linear variations, non-linear calibration methods, bagging/subbagging for robust calibration, and penalized regressions with variable selection using LASSO.

2.1 Pre-processing techniques to deal with external variations

External fluctuations, such as light scattering effect and baseline noise, introduce uncertain non-linear variations in spectroscopic data, which adversely affect the prediction accuracy of conventional linear calibration models. Pre-processing methods aim at removing the non-linearity in data, and then a linear model (PCR or PLS) can be applied. In this thesis, PLS is adopted.

Five pre-processing methods are reviewed, including the first (D1) and second derivatives (D2), standard normal variate (SNV), extended multiplicative signal correction (EMSC), and extended inverted signal correction (EISC). In addition, a unique linear calibration strategy, optical path length estimation and correction (OPLEC), is also investigated in this thesis. Different from the pre-processing technique, OPLEC involves the development of two calibration models. The principal of OPLEC will be explained in this section since it is also a linear calibration method.

2.1.1 The first and second derivatives

High- or low-frequency interferences, defined as noise and background, commonly exist in spectra. Here frequency means change from variable to variable. Light scattering effect resulting from sample-to-sample variations may lead to

interferences, which can be removed by taking derivative of the spectra with respect to the variable number or other relevant axis scale (wavelength, wavenumbers, etc.) [17].

The simplest form of derivative is the first derivative (D1). D1 is obtained by subtracting every variable of a sample from its immediate neighboring variable. From this subtraction, the same signal between the two variables is removed and the difference is retained. In this way, D1 eliminates offset and emphasizes higher-frequency signals. By repeating this procedure on D1, the second derivative (D2) can be calculated. D2 further intensifies higher-frequency features.

Since derivatives obtained from the above finite difference methods typically amplify high-frequency measurement noise, Savitzky-Golay algorithm is desired to attain smoother derivatives for calibration purposes [23]. As a low-pass filter, this algorithm smoothens data by fitting individual polynomials to windows around each variable in the spectra. Both the order of polynomial and the size of window (filter width) need to be selected. When the polynomial order is lower and the window is larger, more smoothing follows. Typically, the window should be on the order of or smaller than the nominal width of non-noise features which should not be smoothed.

2.1.2 Standard normal variate

Standard normal variate (SNV) is a weighted normalization algorithm, in which

the standard deviation of all variables for the given sample is calculated, and then the entire sample is normalized by this value. The model of SNV is

$$\mathbf{x}_i = a_i \mathbf{1} + b_i \mathbf{x}_{i,chem} \quad (2.8)$$

where $\mathbf{x}_{i,chem}$ denotes the theoretical spectra, and $\mathbf{1}$ is a unity vector. The coefficients, a_i and b_i , indicate the additive and multiplicative effects due to external disturbances. In SNV, a_i and b_i are the mean and standard deviation of the i^{th} sample with respect to the wavelength, and thus are calculated based on individual samples [24]. The corrected spectra are obtained by $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i \mathbf{1}) / b_i$, where

$$a_i = \frac{1}{J} \sum_{j=1}^J x_{ij} \quad (2.9)$$

$$b_i = \sqrt{\frac{1}{(J-1)} \sum_{j=1}^J (x_{ij} - a_i)^2} \quad (2.10)$$

2.1.3 Multiplicative signal correction and its extension

Multiplicative signal correction (MSC), originally called multiplicative scatter correction, is used to compensate for the additive and multiplicative effects in spectra. The conventional model of MSC is similar to SNV [14]:

$$\mathbf{x}_i = a_i \mathbf{1} + b_i \mathbf{x}_{i,chem} + \boldsymbol{\epsilon}_i \quad (2.11)$$

where $\boldsymbol{\epsilon}_i$ represents the residual. In MSC, it is assumed that $\mathbf{x}_{i,chem} \approx \mathbf{m} + \boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i$ is the unknown and irrelevant variation and $\mathbf{m} = [m_1, m_2, \dots, m_j]$ denotes

the reference spectra. The coefficients, a_i and b_i , can be estimated by ordinary least squares regression of $\mathbf{x}_i$ to $\mathbf{m}$, minimizing the sum of squared residuals [47]. After obtaining a_i and b_i , the corrected spectra is calculated as $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i \mathbf{1})/b_i$.

The applicability of MSC is restricted by the assumption that the chemical variations between theoretical and reference spectra can be ignored. Otherwise, the estimated coefficients may contain certain irrelevant and undesirable information [26]. More recently, extended multiplicative signal correction (EMSC) has been developed to address this issue [21]. The model of EMSC is given by

$$\mathbf{x}_i = a_i \mathbf{1} + b_i \mathbf{x}_{i,chem} + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2 + \boldsymbol{\epsilon}_i \quad (2.12)$$

where $\boldsymbol{\lambda}$ is the wavelength vector, d_i and e_i represent the smooth wavelength-dependent spectral variations from sample to sample. According to Beer-Lambert's law, the theoretical spectra $\mathbf{x}_{i,chem}$ are linear combination of absorbance contributions from all the K constituents:

$$\mathbf{x}_{i,chem} = \sum_{k=1}^K y_{ik} \mathbf{s}_k \quad (2.13)$$

where y_{ik} denotes the concentration of the k^{th} constituent in the i^{th} sample, $\mathbf{s}_k$ is the pure absorption spectra of the k^{th} constituent. In the ideal case, where the pure spectra are available, the model parameters can be estimated using least squares and then the corrected spectra are given by $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i \mathbf{1} - d_i \boldsymbol{\lambda} - e_i \boldsymbol{\lambda}^2)/b_i$. When this is not possible, the mean spectra of the entire dataset are typically used in place of $\mathbf{x}_{i,chem}$ in Eq. (2.11) for parameter estimation. The use of mean spectra, however, typically results in inferior prediction accuracy.

2.1.4 Inverted signal correction and its extension

Inverted signal correction (ISC) is based on the “reverse” expression of MSC. The model of ISC is [25]

$$\mathbf{m} = a_i \mathbf{1} + b_i \mathbf{x}_i + \epsilon_i \quad (2.14)$$

where the parameters, a_i and b_i , are estimated by least squares regression of $\mathbf{m}$ to $\mathbf{x}_i$ in ISC. Hence the corrected spectra is $\mathbf{x}_{i,corr} = a_i \mathbf{1} + b_i \mathbf{x}_i$.

As MSC has been extended to EMSC, ISC can be also extended to EISC in an attempt to separate physical light scattering effect and chemical absorbance information. EISC assumes that complex optical phenomena can be approximated by a polynomial extension [25]:

$$\mathbf{m} = a_i \mathbf{1} + b_i \mathbf{x}_i + c_i \mathbf{x}_i^2 + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2 + \epsilon_i \quad (2.15)$$

The parameters, a_i , b_i , c_i , d_i , e_i , can be estimated by least squares, and the corrected spectra are given by $\mathbf{x}_{i,corr} = a_i \mathbf{1} + b_i \mathbf{x}_i + c_i \mathbf{x}_i^2 + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2$.

2.1.5 Optical path length estimation and correction

The implementation of EMSC and EISC depends on the availability of pure spectra of all chemical constituents in the sample, which is difficult to be satisfied in practice. To solve this problem, Chen et al. [26] devised the technique of optical path

length estimation and correction (OPLEC) to reduce the detrimental effect of light scattering, which has been demonstrated to significantly improve predictive accuracy without using any pure spectra. However, OPLEC is different from the previous pre-processing techniques. In OPLEC, two calibration models are developed to eliminate light scattering.

OPLEC was devised using the same model as EMSC (Eq. (2.11)) [26], the difference being the parameter estimation algorithm. The parameters (a_i , d_i and e_i) can be removed by projecting the measured spectra $\mathbf{x}_i$ onto orthogonal complement of space spanned by the row vector of $\mathbf{D} = [\mathbf{1}; \boldsymbol{\lambda}; \boldsymbol{\lambda}^2]$:

$$\begin{aligned}\mathbf{z}_i &= \mathbf{x}_i(\mathbf{I} - \mathbf{D}^+\mathbf{D}) = (a_i\mathbf{1} + b_i\mathbf{x}_{i,chem} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2)(\mathbf{I} - \mathbf{D}^+\mathbf{D}) \\ &= \sum_{k=1}^K b_i y_{ik} \mathbf{d}_k\end{aligned}\tag{2.16}$$

where $\mathbf{z}_i$ is the standardized spectra, $\mathbf{d}_k = \mathbf{s}_k(\mathbf{I} - \mathbf{D}^+\mathbf{D})$, $\mathbf{I}$ is an identity matrix with the same dimension as $\mathbf{D}^+\mathbf{D}$, and $\mathbf{D}^+$ represents the pseudo-inverse of $\mathbf{D}$. Assuming that the first constituent is the target constituent in the mixture and $\sum_{k=1}^K y_{ik} = 1$, Eq. (2.15) can be re-written as

$$\begin{aligned}\mathbf{z}_i &= b_i y_{i1} \mathbf{d}_1 + b_i (1 - \sum_{\substack{k=1 \\ k \neq 2}}^K y_{ik}) \mathbf{d}_2 + \sum_{k=3}^K b_i y_{ik} \mathbf{d}_k \\ &= b_i y_{i1} \Delta \mathbf{d}_1 + b_i \mathbf{d}_2 + \sum_{k=3}^K b_i y_{ik} \Delta \mathbf{d}_k\end{aligned}\tag{2.17}$$

where $\Delta \mathbf{d}_k = \mathbf{d}_k - \mathbf{d}_2$.

Subsequently, the projected spectra of I samples, $\mathbf{Z} = [\mathbf{z}_1; \dots; \mathbf{z}_I]$, are partitioned

into $\mathbf{Z}_{base}$ and $\mathbf{Z}_{rest}$. A linear relationship exists between $\mathbf{Z}_{base}$ and $\mathbf{Z}_{rest}$ as $\mathbf{Z}_{rest} = \mathbf{A}\mathbf{Z}_{base}$, whereby the regression coefficients $\mathbf{A}$ are obtained through least squares method. Similarly, $\mathbf{b} = [b_1; \dots b_I]$ and $\mathbf{y}_1 = [y_{11}; \dots y_{I1}]$ are partitioned into $\mathbf{b}_{base}$ and $\mathbf{b}_{rest}$, $\mathbf{y}_{base}$ and $\mathbf{y}_{rest}$, respectively. According to Eq. (2.16), there is a linear relationship between $\mathbf{z}_i$ and b_i , and also between $\mathbf{z}_i$ and $b_i y_{i1}$. Therefore, the following equations can be obtained:

$$\mathbf{b}_{rest} = \mathbf{A}\mathbf{b}_{base} \quad (2.18)$$

$$\text{diag}(\mathbf{y}_{rest}) \mathbf{b}_{rest} = \mathbf{A} \text{diag}(\mathbf{y}_{base}) \mathbf{b}_{base} \quad (2.19)$$

where $\text{diag}(\mathbf{y}_{rest})$ is a diagonal matrix in which the corresponding elements are the elements of $\mathbf{y}_{rest}$. Because $\mathbf{y}_{rest}$ and $\mathbf{y}_{base}$ are known, $\mathbf{b}_{base}$ and $\mathbf{b}_{rest}$ can be obtained from Eq. (2.17) and (2.18). Re-combining $\mathbf{b}_{base}$ and $\mathbf{b}_{rest}$ generates the parameters $\mathbf{b}$. The detailed strategy for partitioning $\mathbf{Z}$ and estimating the parameters $\mathbf{b}$ can be found in the original paper [26].

Having obtained $\mathbf{b}$, two new regression vectors, β_1 and β_2 , can be estimated by the following two models $\text{diag}(\mathbf{y}_1)\mathbf{b} = [\mathbf{1}, \mathbf{Z}]\beta_1$ and $\mathbf{b} = [\mathbf{1}, \mathbf{Z}]\beta_2$ through multivariate linear regression (e.g. PLS). Then β_1 and β_2 are used to predict the concentration of a test sample (y_{test}) as follows:

$$\begin{aligned} b_{test} y_{test} &= [\mathbf{1}, \mathbf{z}_{test}] \beta_1, \quad b_{test} = [\mathbf{1}, \mathbf{z}_{test}] \beta_2 \\ y_{test} &= \frac{[\mathbf{1}, \mathbf{z}_{test}] \beta_1}{[\mathbf{1}, \mathbf{z}_{test}] \beta_2} \end{aligned} \quad (2.20)$$

2.2 Non-linear calibration techniques

Since external variations, such as light scattering effect and baseline noise, introduce unknown non-linearity to measured spectra, it is natural to adopt non-linear calibration techniques to directly model the non-linearity. An overview of three state-of-the-art non-linear calibration methods is presented, including artificial neural network (ANN), Gaussian processes regression (GPR), and least square support vector machine (LS-SVM).

2.2.1 Artificial neural network

As a flexible modeling tool, ANN has been found wide application in various areas, such as pattern recognition, signal processing, optimization and process control [48, 49]. In the 1990s, ANN was introduced in chemometrics community for spectroscopic calibration [27, 50, 51] and was demonstrated to achieve satisfactory predictive performance. A typical feed-forward ANN model consists of three layers (input, hidden and output layer), and each layer comprises multiple neurons. In the context of calibration modeling, the response y_i from the output layer is the analyte property to be predicted, and it can be expressed mathematically as [28]:

$$y_i = f \left[\sum_{n=1}^N \omega_n g \left(\sum_{j=1}^J (v_j x_{ij} + \phi_j) \right) + \varepsilon_n \right] \quad (2.21)$$

where N is the number of hidden-layer neurons, v_j represents the weight connecting

input- and hidden-layer neurons, ω_n indicates the weight connecting hidden- and output-layer neurons, ϕ_j and ε_n are biases in the hidden and output layers, respectively. The two “transfer functions”, $f(\cdot)$ and $g(\cdot)$, are typically taken as linear and sigmoid functions, respectively, since such a neural network is capable of approximating any function to arbitrary accuracy [49].

The parameters in ANN are typically estimated by the “back-propagation” algorithm, an iterative gradient algorithm designed to minimize modeling error by adjusting the weights in the direction of decent gradient [50]. In this study, a Bayesian regularized version of back-propagation algorithm is adopted. Bayesian back-propagation was initially introduced by MacKay [52, 53], based on a Gaussian approximation to the posterior distribution of parameters. The Bayesian approach can automatically determine the “effective” number of parameters, and thus is less sensitive to the pre-chosen number of neurons and less likely to over-fit the data when compared with traditional ANN. Therefore, the choice of the number of neurons does not appear to have significant impact on the results. Previous study has shown that the Bayesian approach to parameter estimation typically attains more robust and accurate ANN model [29].

2.2.2 Gaussian process regression

GPR was originally presented by O’Hagan [54] and it is regarded as an alternative

approach to ANN, because a large number of Bayesian regression models based on ANN can converge to a Gaussian process [55]. GPR has recently been applied to spectroscopic calibration modeling [10], with outstanding prediction accuracy being achieved. In a GPR model, the response analyte property y_i is modeled by a joint Gaussian distribution with zero mean:

$$\mathbf{y} = (y_1, \dots, y_I) \sim G(\mathbf{0}, \mathbf{C}) \quad (2.22)$$

where $\mathbf{C}$ with dimensions $I \times I$ is a covariance matrix whose elements are defined by a covariance function: $C_{ik} = C(\mathbf{x}_i, \mathbf{x}_k)$. The following covariance function, widely used in literatures, is adopted:

$$C(\mathbf{x}_i, \mathbf{x}_k) = a_0 + a_1 \sum_{j=1}^J x_{ij} x_{kj} + v_0 \exp\left(-\sum_{j=1}^J w_j (x_{ij} - x_{kj})^2\right) + \sigma^2 \delta_{ik} \quad (2.23)$$

where the first two terms represent the constant bias and linear correlation, the third term is similar to the form of a radial basis function, and the fourth term corresponds to the random error. Containing both linear and non-linear terms in the covariance function, GPR can effectively model both linear and non-linear datasets.

The model hyper-parameters, denoted by $\boldsymbol{\theta} = (a_0, a_1, v_0, w_1, \dots, w_J, \sigma^2)$, can be estimated by maximizing the following log-likelihood:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - \frac{I}{2} \log(2\pi) \quad (2.24)$$

To ensure that the covariance matrix is non-negative definite, the hyper-parameters also should be non-negative. Given the estimated parameters, for new data points $\mathbf{x}^*$, the predictions y^* are also Gaussian distributed with mean and variance as:

$$\begin{aligned}
E(y^*) &= \mathbf{k}^T(\mathbf{x}^*)\mathbf{C}^{-1}\mathbf{y} \\
\text{Var}(y^*) &= \mathcal{C}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T(\mathbf{x}^*)\mathbf{C}^{-1}\mathbf{k}(\mathbf{x}^*)
\end{aligned} \tag{2.25}$$

where $\mathbf{k}(\mathbf{x}^*) = [\mathcal{C}(\mathbf{x}^*, \mathbf{x}_1), \dots, \mathcal{C}(\mathbf{x}^*, \mathbf{x}_I)]^T$.

2.2.3 Least-square support vector machine

The technique of support vector machine (SVM) was initiated by Vapnik to solve pattern recognition problems [56], with advantages of unique global solution and handling both linear and non-linear problems. However, in SVM, three meta-parameters (related to error trade-off, error insensitivity and selected kernel function) need to be optimized by cross-validation, which is undesirable in terms of computation. In this regard, a variant of SVM, known as least square SVM (LS-SVM) is desired, since it has only two meta-parameters whilst still retaining the advantages of SVM. Therefore, LS-SVM is more widely adopted in calibration modeling [31].

The simplest LS-SVM is a linear regression model: $y_i = \mathbf{w}^T \mathbf{x}_i + \mathbf{e}_i$, where $\mathbf{w}$ is the vector of model parameters, and $\mathbf{e}_i$ is the regression error. As opposed to minimizing the sum of errors in conventional regressions, LS-SVM includes a regularization term, and aims to minimize the following cost function [31]:

$$Q = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \sum_{i=1}^I \|\mathbf{e}_i\|^2 \tag{2.26}$$

where the first part penalizes the magnitude of regression coefficients in order to avoid “over-fitting”. The meta-parameter γ decides the relative weight regression

error compared with the penalizing term. The model parameters can be estimated using the Lagrange method [57]. The extension of the above linear regression to non-linear model can be achieved by introducing non-linear kernel functions, for example the radial basis function (RBF).

In LS-SVM, meta-parameters that need to be cross-validated include γ and parameters in the kernel function (only one parameter in RBF kernel function). Note that reducing three meta-parameters in SVM to two in LS-SVM can significantly reduce the computation in cross-validation, since the increase in the number of meta-parameters results in exponential increase in the number of their combinations.

2.3 Ensemble modeling for robust calibration

Non-linear calibration methods have been demonstrated to achieve accurate prediction of analyte properties [10, 27, 50, 51]. However, their performances are not robust in the sense that models are sensitive to small changes in data or model parameters. Recent work has shown that the accuracy of non-robust methods can be significantly improved through ensemble modeling. The general idea of ensemble modeling is to construct multiple models, then combine their predictors into a single one using certain rules [58].

Bagging, short for ‘**bootstrap aggregating**’ and originally contrived by Breiman [34, 35], is one of the important ensemble modeling techniques. The bootstrap,

developed by Efron [59, 60], is a technique of forming different training sets by randomly selecting a fixed number of data points from the original training set *with replacement*. Many studies have confirmed the effectiveness of bagging in the reduction of prediction error [61-63]. It has also been demonstrated that bagging works especially well for non-robust models, in which small changes in data or model parameters can lead to significant changes in model prediction [34].

In regression field, bagging was initially applied to regression trees [34]. Subsequently, attention was focused on bagging neural networks [64-69]. Recently, bagging has been developed for other regression methods, such as partial least squares (PLS), multiple linear regression (MLR) [70], and Gaussian process regression (GPR) [71]. Over the years, bagging has been modified in several ways such as “nice” bagging [72], iterated bagging [73], subbagging (**sub**-sample **agg**regating, based on sub-sampling *without replacement*) [36], and trimmed bagging [74]. Among these modified methods, subbagging has been reported to provide similar performance to bagging with less computation, since it uses a subset of data for model development [36]. For the purpose of multivariate spectroscopic calibration, subbagging has already been applied to linear methods as PLS and MLR with variable selection [70]; yet its role in improving non-linear calibration methods (such as ANN and GPR) has been under-explored. Therefore, in this thesis, the application of bagging and subbagging for non-linear calibration methods (specifically ANN and GPR) will be investigated to obtain more accurate and robust predictions.

The principle of bagging is conceptually straightforward. In bagging, a number of models are developed from a re-sampling process on the original training (calibration) data *with replacement*. Suppose the original training data are $\mathbf{F} = \{\mathbf{x}_i, y_i\}$. A new re-sampled training set is constructed by randomly selecting I data points from $\mathbf{F}$ *with replacement*. This process is repeated R times to obtain R different training sets. Then, R calibration models can be built from the R re-sampled training sets. For a testing data point, R predicted values can be obtained from the R models, and then these R values are combined in a certain way to form the ultimate prediction. A simple averaging rule [34] has been recommended and is adopted in this thesis. Specifically, the final prediction $\hat{y}$ of ensemble model is calculated as

$$\hat{y} = \frac{1}{R} \sum_{r=1}^R \hat{y}(r) \quad (2.27)$$

where $\hat{y}(r)$ is the prediction of analyte property by the r^{th} model developed from the r^{th} training set. An alternative combination approach is the weighted averaging rule [69], which is more complicated than the simple averaging rule. In this rule, different weights are assigned to the models, and the final prediction $\hat{y}$ is defined as

$$\hat{y}_i = k_1 \hat{y}_i(1) + k_2 \hat{y}_i(2) + \cdots + k_R \hat{y}_i(R) \quad (2.28)$$

The simple averaging rule is sufficient to demonstrate the effectiveness of bagging/subbagging in the application of multivariate calibration, which is the mainly purpose of this thesis. Therefore, the simple averaging rule is just adopted.

The principle of subbagging is similar to bagging. In subbagging, R new calibration models are constructed by randomly sub-sampling P ($P \leq I$) data points from $\mathbf{F}$

without replacement R times. Subsequently, the R models are combined to make prediction.

In summary, the process of bagging/subbagging for non-linear calibration method includes two steps: (1) obtaining R calibration models from the R re-sampled/sub-sampled training datasets; (2) combining the resulting models to generate an ensemble model and making prediction. The fundamental reason for the effectiveness of bagging non-robust models is that the predictive capability of regression models is based on bias/variance trade-off, and variance of aggregated predictor is reduced and almost constant bias is maintained [75].

2.4 Penalized regression with variable selection

In multivariate calibration, not all variables have a large effect on the model, and some do not contain any relevant information. Therefore, informative variables should be selected. Instead of using all wavelengths to build a calibration model, variable selection methodology only chooses specific wavelengths that are informative for predicting the analyte property, but rejects spectral regions that are sensitive to external variations and non-informative to the analyte property [33]. Of course, variable selection cannot reduce the sum of squared errors (SSE) of modeling data. However, it can improve model prediction results in terms of accuracy and robustness using relevant spectral variables with least colinearity, redundancies and noise [76,

77]. A distinct advantage of variable selection is that the variance of estimated parameters decreases as fewer parameters are estimated from limited data. Uninformative variable elimination (UVE) [37] is a usual variable selection method. In this method, a leave-one-out procedure is used, and hundreds of calibration models need to be developed from randomly selected samples. Afterwards, the usefulness of each variable is evaluated through a selection criterion [76]. Therefore, this method involves complicated selection algorithm and requires high computation [33]. Furthermore, it is impossible to identify and remove all wavelengths that are influenced by external variations [78]. Hence, the development of heuristic selection methods has spurred. Öjelund [38] proposed penalized linear regression (PLR) with LASSO for variable selection. Recently, another penalization technique, penalized Gaussian process regression (PGPR), has also been put forward to select the correlated covariates [39]. The principles of PLR and PGPR are explained in details as follows.

2.4.1 Penalized linear regression with LASSO

In penalized linear regression with LASSO, the estimates $\hat{\boldsymbol{\eta}}$ can be obtained as [38, 79]

$$(\hat{\boldsymbol{\eta}}) = \arg \min \left(\frac{1}{I} \sum_{i=1}^I (y_i - \mathbf{x}_i^T \boldsymbol{\eta})^2 \right) \text{ subject to } \sum_{j=1}^J |\eta_j| \leq t \quad (2.29)$$

where $t \geq 0$ is a hyper-parameter. This constrained minimization problem can be

equivalent to an unconstrained problem as

$$(\hat{\boldsymbol{\eta}}) = \arg \min \left(\frac{1}{I} \sum_{i=1}^I (y_i - \mathbf{x}_i^T \boldsymbol{\eta})^2 + \kappa \sum_{j=1}^J |\eta_j| \right) \quad (2.30)$$

where the Lagrange parameter κ is selected by subjecting $\sum_{j=1}^J |\eta_j| \leq t$. The standard Newton-Raphson method is adopted to solve the minimization unconstrained problem of Eq. (2.28) [38].

At a certain value of κ , some parameter estimates which are not important to the calibration model are close to zero, while others are far from zero. The purpose of variable selection can be achieved by penalizing the close to zero parameters and setting them to zero, while other parameters are kept. After the estimation of $\boldsymbol{\eta}$, for new data points $\mathbf{x}^*$, the predictions y^* can be obtained by

$$y^* = \mathbf{x}^{*T} \boldsymbol{\eta} \quad (2.31)$$

2.4.2 Penalized Gaussian process regression with LASSO

In traditional GPR, the issue of variable selection is handled by automatic relevance determination (ARD) [55]. In ARD, when covariance structure is almost irrelevant to the analyte property, the predictor can be removed from GPR model. However, it is complicated to find an accurate “threshold” to remove the irrelevant predictors. Therefore, penalized regression technique is applied to GPR model to solve the variable selection problem [39].

The model of GPR used in [39] is the same as discussed in Section 2.2.2, but with

covariance function defined as

$$C(\mathbf{x}_i, \mathbf{x}_k) = v_0 \exp\left(-\frac{1}{2} \sum_{j=1}^J w_j (x_{ij} - x_{kj})^2\right) + \sigma^2 \delta_{ik} \quad (2.32)$$

In this thesis, since EMSC is first adopted to remove the non-linearity in the data, this non-linear squared exponential covariance function is not necessary, and the following linear covariance function is adopted as

$$C(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^J w_j x_{ij}^T x_{kj} + \sigma^2 \delta_{ik} \quad (2.33)$$

Instead of maximizing the log-likelihood in Eq. (2.23) in GPR, PGPR estimates the hyper-parameters by minimizing the next modified likelihood which is penalized by LASSO:

$$l_p(\boldsymbol{\theta}, \lambda_n) = -\frac{1}{I} l_i(\boldsymbol{\theta}) + \kappa \sum_{j=1}^J w_j \quad (2.34)$$

where $l_i(\boldsymbol{\theta}) = \log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - \frac{I}{2} \log(2\pi)$ as shown in Eq. (2.23).

In PGPR, since w_j is non-negative, a constrained optimization problem, which makes the optimization procedure more complicated, is needed. In order to simplify the optimization problem, the log values of hyper-parameters are used, and the complicated constrained optimization procedure is avoided. The purpose of variable selection can be achieved in the process of calculating the hyper-parameters of covariance function. When an optimal value of κ is found, some values of the hyper-parameters are small comparatively, while others are very large. Then the log

hyper-parameters with small values are set to zero, while others are kept and re-estimated. In this way, variable selection is implemented.

Chapter 3 Comparative Study on Multivariate Calibration Methods to Correct Light Scattering Effect

Various multivariate calibration techniques have been devised to handle light scattering effect. As a consequence of rapid development of scattering-corrected calibration models, the rigorous assessment and comparison of available techniques are required. In previous comparative studies, some chemometric methods, such as D1, D2, SNV, EMSC, EISC and ANN, have been investigated [19, 80-83]. However, several novel techniques, including OPLEC, LS-SVM and GPR, have emerged in recent years with very promising results being achieved. Therefore, this chapter provides an update of the new developments in this field. The relative advantage of pre-processing methods and non-linear calibration models are particularly interesting, since they originate from different rationales. Nine chemometric calibration methods (D1, D2, SNV, EMSC, EISC, OPLEC, ANN, GPR and LS-SVM) are considered in this chapter. Furthermore, the combination of pre-processing and non-linear calibration methods is explored in an attempt to achieve more accurate prediction than individual techniques. Two pre-processing techniques of SNV and EMSC are considered, which are followed by developing non-linear models to give six combinations: SNV+ANN, EMSC+ANN, SNV+GPR, EMSC+GPR, SNV+LS-SVM and EMSC+LS-SVM. The predictive accuracy, quantified by root mean square error

of prediction (RMSEP), is used to evaluate the performance of these investigated calibration methods. RMSEP is defined as

$$\sqrt{(1/I) \sum_{i=1}^I (y_i - \hat{y}_i)^2} \quad (3.1)$$

where y_i and $\hat{y}$ are the predicted and reference analyte property of the i^{th} sample, respectively.

3.1 Datasets

This comparative study is based on three NIR datasets. The first dataset was collected from the analysis of pharmaceutical tablets, which is publicly available at <http://www.models.life.ku.dk/Tablets>. The tablets were manufactured under diversely different conditions including laboratory, pilot, and full production scales, making the calibration task more challenging. This dataset consists of 310 samples with 404 wavenumbers in the range of 7400-10500cm⁻¹. The objective is to determine the active substance content of tablets. Two hundred samples were randomly selected as the training data for model development, and the remaining 110 samples were used for testing. Further details about this dataset are described in [40].

The second dataset is related to the transmittance spectra of wheat kernels [25]. In this dataset, 523 samples from three different locations were analyzed at 100 wavelengths in the range of 850-1050nm. The data are available at

http://www.models.life.ku.dk/wheat_kernels. The objective of this analysis is to predict the protein content of wheat kernels. This dataset is divided into 415 and 108 samples at random for training and testing, respectively.

The third dataset concerns about the gluten/starch powder mixture with 100 samples and 100 wavelengths [21]. For each of five mixtures with different weight ratios of gluten and starch (1/0, 0.75/0.25, 0.5/0.5, 0.25/0.75, 0/1), five samples were loosely filled into different glass cuvettes and measured in two consecutive replicates. Then the samples were compressed firmly and measured in two consecutive replicates again. In this study, the training data consist of 60 random samples and the rest 40 samples are used for testing.

3.2 Statistical evaluation of calibration methods

When dividing a limited dataset into training and testing data, the comparison results may be unreliable if training and/or testing data are not representative of the whole dataset. The usual method to address this issue is to randomly partition the dataset multiple times (e.g. 50 times), and then apply each calibration method to obtain 50 RMSEPs. The average RMSEP from each method can be used for the comparison purpose. Furthermore, a rigorous statistical hypothesis testing procedure, the paired t -test, is adopted to determine whether one method is statistically significantly better than the other. Specifically, suppose that r_m and s_m are

RMSEPs of two calibration methods tested on the m^{th} ($m = 1, \dots, M$) partition of the dataset, then t statistic is given by [84]

$$t = (\bar{r} - \bar{s}) \sqrt{\frac{M(M-1)}{\sum_{m=1}^M (\tilde{r}_m - \tilde{s}_m)^2}} \quad (3.2)$$

where $\bar{r}$ and $\bar{s}$ are the means of r_m and s_m , $\tilde{r}_m = r_m - \bar{r}$ and $\tilde{s}_m = s_m - \bar{s}$, and $M=50$ in this study. After calculating t statistic, the corresponding p -value can be obtained. If $p < 0.05$, it may be concluded that these two calibration methods have attained significantly different RMSEPs. The paired t -test is a widely used technique to conduct rigorous comparative studies [10, 19].

As mentioned in Chapter 2.1.3, pure spectra of individual chemical components are preferred as the reference spectra for the implementation of EMSC [26]. For “gluten/starch” dataset, the 3rd and 93rd samples are used as the pure spectra of gluten and starch, respectively [21]. However, for the other two datasets, no pure spectra are available, and thus the mean spectra of the entire training data are adopted as the reference spectra. In addition, whenever PLS is used in this chapter, either for calibration modeling or for estimating OPLEC parameters, the number of retained PLS components is always optimized through five-fold random-splitting cross-validation. Prior to developing non-linear calibration models (ANN, GPR and LS-SVM), PLS is first applied to reduce the dimension of spectral data, which is a usual method to reduce computational cost of parameter estimation in non-linear models [10, 50]. The number of LVs in this PLS-based “pre-processing” is also selected by five-fold random-splitting cross-validation. Subsequently, ANN and GPR

are developed from PLS scores, i.e. variable $\mathbf{x}$ denotes PLS scores other than the spectra. When carrying out LS-SVM, a Gaussian kernel function is used. Both meta-parameters γ (see Eq. (2.25)) and σ^2 (kernel parameter) are also optimized by five-fold random-splitting cross-validation.

Another important consideration in calibration is outlier detection. It is well known that including outliers in model development has adverse impact on prediction accuracy. Due to its importance, outlier detection has been extensively discussed in the literature; see e.g. [85-87] for an overview of current status and some recent development. In this work, the built-in outlier detection capability of PLS [88] is utilized and no obvious outliers are found in the datasets under investigation.

3.3 Software

All computation was carried out in Matlab 7.5. Matlab PLS toolbox version 5.2 (Eigenvector Research, Inc., Wenatchee, WA, USA) was used to perform PLS, D1, D2 and SNV. EMSC and EISC were performed by using EMSC toolbox version 1.3 (Eigenvector Research, Inc., Wenatchee, WA, USA). Matlab code for OPLEC is available in [89]. Matlab Neural Network toolbox was employed for the implementation of ANN. LS-SVM was carried out using a toolbox from Katholieke Universiteit Leuven (url: <http://www.esat.kuleuven.ac.be/sista/lssvmlab/>). Matlab code for Gaussian process was described in [55] and is available at

3.4 Results and discussion

Table 3.1 The number of selected components when using PLS in pre-processing or non-linear calibration techniques for (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture. The values are averaged over 50 repeated random partitions of the data.

(a)

Techniques	D1	D2	SNV	EMSC	EISC	OPLEC	ANN	GPR	LS-SVM
Number	2.7	8.8	6.5	4.1	4.3	6.0	6.5	6.6	6.4

(b)

Techniques	D1	D2	SNV	EMSC	EISC	OPLEC	ANN	GPR	LS-SVM
Number	11.1	9.3	12.1	9.4	9.2	12.0	12.0	12.1	12.0

(c)

Techniques	D1	D2	SNV	EMSC	EISC	OPLEC	ANN	GPR	LS-SVM
Number	8.2	6.7	9.5	6.6	6.8	6.0	9.3	9.6	9.3

Table 3.1 reports the number of the selected PLS components, averaged over 50 random partitions of the three datasets. Before constructing non-linear calibration models (ANN, GPR and LS-SVM), PLS is first used to reduce the dimension of spectral data. The number of retained PLS components is also reported in Table 3.1, where the spectral dimension is reduced from several hundred to less than 13, alleviating the computation for developing non-linear models.

Table 3.2 The p -values of paired t -test for different calibration techniques on (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.

(a)

p -value	D2	SNV	EMSC	EISC	OPLEC	ANN	GPR	LS-SVM
D1	$<10^{-3}$	0.55	$<10^{-3}$	0.15	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
D2		$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
SNV			$<10^{-3}$	0.25	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
EMSC				$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
EISC					$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
OPLEC						0.27	$<10^{-3}$	$<10^{-3}$
ANN							$<10^{-3}$	$<10^{-3}$
GPR								$<10^{-3}$

(b)

<i>p</i> -value	D2	SNV	EMSC	EISC	OPLEC	ANN	GPR	LS-SVM
D1	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
D2		$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
SNV			$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
EMSC				$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	0.27	0.03
EISC					0.78	0.71	$<10^{-3}$	0.78
OPLEC						0.88	$<10^{-3}$	0.63
ANN							$<10^{-3}$	$<10^{-3}$
GPR								$<10^{-3}$

(c)

<i>p</i> -value	D2	SNV	EMSC	EISC	OPLEC	ANN	GPR	LS-SVM
D1	$<10^{-3}$	0.04	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
D2		$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
SNV			$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
EMSC				$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	0.43	$<10^{-3}$
EISC					$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
OPLEC						$<10^{-3}$	10^{-3}	$<10^{-3}$
ANN							$<10^{-3}$	0.04
GPR								$<10^{-3}$

(a)

(b)

(c)

Figure 3.1 RMSEPs of different calibration techniques for (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.

Figure 3.1 gives RMSEPs averaged over 50 repeated random partitions of the three datasets for different calibration methods, and Table 3.2 shows the corresponding p -values of paired t -test. For “tablets” dataset, a general observation is that non-linear calibration models attain lower prediction error than linear methods. The only exception is that ANN gives the same RMSEP of 0.31 as OPLEC and their paired p -value is 0.27 (> 0.05), suggesting statistically insignificant difference between these two methods. Among non-linear calibration methods, GPR achieves the best predictive accuracy in terms of its RMSEP (0.27) and p -values between GPR and other methods being less than 0.001. LS-SVM also gives satisfactory RMSEP of 0.29. Among linear calibration methods, OPLEC attains the best performance (RMSEP = 0.31, $p < 0.05$ when compared with other linear methods); however, it does not outperform non-linear models. The second best linear method is EMSC (RMSEP = 0.34) followed by D1, SNV and EISC; the latter three methods have the same RMSEP (0.37) and $p > 0.05$ between them, indicating similar prediction performance. Finally, D2 gives the largest RMSEP of 0.44.

For “wheat kernels” dataset, although RMSEPs of these techniques are similar, linear methods of EMSC, EISC, OPLEC and non-linear methods of ANN, GPR, LS-SVM still attain lower errors than D1, D2, SNV, and these better performances are statistically significant according to the paired t -test. Among these superior methods, GPR and EMSC achieve the most accurate prediction. The p -value between GPR and

EMSC is greater than 0.05, indicating that they have similar predictive capability. RMSEP of 0.46 for EISC, OPLEC, ANN and 0.48 for LS-SVM suggest they also give satisfactory and similar predictions.

For “gluten/starch” dataset, EMSC, OPLEC and GPR give especially low prediction errors (RMSEP are 0.005, 0.003 and 0.005, respectively) and p -values are all less than 0.001, implying that they achieve excellent predictive accuracy over other calibration methods. EISC also attains an acceptable performance in view of its RMSEP (0.010). The results of other calibration methods (D1, D2, SNV and ANN, LS-SVM) are unsatisfactory. It should be noted that the pure spectra of chemical constituents are available for the application of EMSC and EISC for this dataset. If this is not the case in practice, their prediction accuracy is likely to decrease. Overall, the best linear method (OPLEC) outperforms the best non-linear model (GPR) on this dataset.

To assess the robustness of calibration methods, the standard deviations of RMSEPs are shown in Figure 3.2. For “tablets” dataset, the standard deviations of ANN and LS-SVM are remarkably higher than the other methods, implying that ANN and LS-SVM are sensitive to variations in spectral data. For “wheat kernels” dataset, EMSC, OPLEC, ANN and GPR appear to be more robust than other techniques. For “gluten/starch” dataset, EMSC and OPLEC achieve outstanding robustness in terms of their very small standard deviations. Overall, no single method has shown consistently better robustness than all other methods. The robustness of regression

models can be enhanced by ensemble modeling of combining predictions from multiple models, for example, bootstrap aggregating (bagging) [71, 90]. Further investigation on ensemble modeling for calibration is given in the next chapter.

Figure 3.3 shows the prediction performance when pre-processing and non-linear models are combined. For “tablets” dataset, SNV is preferred to EMSC as a pre-processing method to be combined with non-linear models. The combination of EMSC with the three non-linear calibration models gives significantly higher RMSEPs than individual methods. For “wheat kernels”, only SNV+LS-SVM gives better results than individual methods. Indeed, all combinations attain similar RMSEPs. It may be concluded that for this dataset, the use of SNV or EMSC as pre-processing for non-linear models does not introduce significant benefit. For “gluten/starch” dataset, although the combination of EMSC+ANN gives better result than ANN, it does not surpass EMSC. RMSEPs of SNV+GPR, SNV+LS-SVM and SNV+ANN are lower than SNV, but still higher than GPR, LS-SVM and ANN. The predictive performances of other combinations are all worse than using individual techniques.

(a)

(b)

(c)

Figure 3.2 The standard deviations of RMSEPs for different calibration techniques: (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.

(a)

(b)

(c)

Figure 3.3 RMSEPs of combining pre-processing and non-linear calibration models for (a) tablets; (b) wheat kernels; (c) gluten/starch power mixture.

3.5 Summary

This chapter compares linear and non-linear modeling techniques for the calibration of NIR spectroscopy in the presence of light scattering effect. Although none of the techniques is always the best on all datasets, OPLEC and GPR are found to be the most promising in terms of their low prediction error. Compared with traditional pre-processing approaches (D1, D2 and SNV), the more recently developed methods (EMSC, EISC and OPLEC) are more favorable. This is due to better modeling of light scattering effect (such as including the wavelength terms λ in the model) and more advanced parameter estimation strategy (such as that of OPLEC). Among the three non-linear models considered in this study, GPR is recommended since it consistently attained lower RMSEP than ANN and LS-SVM. Finally, the strategy of combining pre-processing and non-linear techniques does not always outperform individual techniques.

This work has been published in *Analytical Letters* (Wang et al. *Analytical Letters*, 2011, 44, 824–836); reuse in this chapter is under the permission of Taylor and Francis Ltd.

Chapter 4 Bagging for Robust Non-linear Multivariate Calibration

In the previous comparative study, non-linear calibration techniques have been found to give satisfactory prediction accuracy. However, their prediction performance may be sensitive to small changes in calibration data and/or model parameters in the sense that models based on different calibration data and/or initial parameters give different prediction errors as proved in Chapter 3. Therefore, bagging and subbagging for non-linear calibration methods (specifically ANN and GPR) to obtain more accurate and robust predictions are investigated. In addition, the effect of the amount of sub-sampled training data on the prediction performance of subbagging is examined, a topic that was not carefully examined in the literature of chemometric calibration. The results indicate that subbagging is sensitive to the amount of sub-sampled data, which needs to be determined by the computationally intensive cross-validation method. Therefore, it is suggested that bagging be preferred to subbagging.

Two publicly available near infrared (NIR) datasets form the basis of case study. When developing a calibration model, each dataset needs to be divided into training and testing data. It is crucial to ensure that the training/testing data are representative of the whole dataset in order to conduct reliable and meaningful evaluation. In Chapter 3, randomly splitting method is adopted to address this issue, since it is simple and data randomly sampled from a large set follow statistical distribution of

the entire dataset. However, it cannot prevent extrapolation problem and cannot guarantee that samples on the dataset boundary are included in training set [91]. Recently, a new partitioning strategy of SPXY algorithm (sample set partitioning based on joint $\mathbf{x}$ - $\mathbf{y}$ distances) [91], which is a development of the classic Kennard-Stone algorithm [92], has been devised. In this chapter, SPXY algorithm is applied to ensure that training/testing data are representative of the whole dataset.

4.1 Datasets

In order to validate the effectiveness of the proposed techniques of bagging/subbagging, several datasets are tested, including the datasets used in Chapter 3. Satisfactory improvements of model robustness and prediction accuracy are observed for almost all of them. Here two NIR spectral datasets among them are displayed to demonstrate the usefulness of bagging/subbagging. The first dataset is related to the transmittance spectra of wheat kernels, which is the same as used in Chapter 3 [25]. This dataset is divided into 415 and 108 samples for training and testing, respectively.

In order to show the wide applicability of bagging/subbagging, the second dataset, which is named “meat” and different from the datasets used in Chapter 3, is adopted. This dataset is publicly available at <http://lib.stat.cmu.edu/datasets/tecator>. The data were collected on a Tecator Infratec Food and Feed Analyzer, consisting of 215 samples with 100 wavelengths in the range of 850-1050 nm. The objective of analysis

is to predict the contents of moisture (water), fat and protein in finely chopped meat. In this study, the training data consist of 172 samples and the rest 43 samples are used for testing.

4.2 Statistical evaluation of bagging/subbagging

Root mean square error of prediction (RMSEP) is also used to evaluate the performance of investigated calibration methods. To better evaluate the advantage of bagging/subbagging, a relative RMSEP reduction is also calculated as [70]

$$\frac{\langle \text{RMSEP}_{\text{individual}} \rangle - \langle \text{RMSEP}_{(\text{su})\text{bagging}} \rangle}{\langle \text{RMSEP}_{\text{individual}} \rangle} \times 100\% \quad (4.1)$$

which indicates the improvement of (su)bagging procedure with respect to ANN and GPR.

When dividing a limited dataset into training and testing data, the training and testing data should be representative of the entire dataset. For this purpose, SPXY algorithm [91] is employed in this chapter to extract the representative training data, while the rest are used for testing. The actual partition based on SPXY is dependent on the initially selected training sample. Hence, the SPXY-partition is repeated 20 times with randomly different initialization, and the average RMSEP of the 20 repeats is reported to evaluate the prediction performance. Note that SPXY algorithm is based on a linear distance function; nevertheless its effectiveness for non-linear regression models has been reported in the literature [93].

For benchmark comparison, bagging/subbagging is also applied to linear PLS method, with the number of latent variables (LVs) being selected by five-fold random-splitting cross-validation. For the purpose of studying the influence of the amount of sub-sampled training data on the prediction performance of subbagging, different ratios of sub-sampled/whole training data (i.e. P/I) from 0.1 to 1 with a step of 0.1 is tried. In practice, this ratio should be selected automatically, and a four-fold random-splitting cross-validation is used in this study. For the purpose of bagging/subbagging, 20 models are developed for PLS, ANN and GPR. The selection of 20 models is supported by a preliminary study to be explained in the next section. Prior to developing non-linear calibration methods (ANN and GPR), PLS is applied to reduce the dimension of spectral data, which is the same as in Chapter 3.

All computation was carried out in Matlab 7.5. Matlab PLS toolbox version 5.2 (Eigenvector Research, Inc., Wenatchee, WA, USA) was used to perform PLS. Matlab Neural Network toolbox was employed for the implementation of ANN. Matlab code for Gaussian process was described in [55] and is available at <http://www.gaussianprocess.org/gpml/code/matlab/doc/>. The computation time was based on a Pentium 2.4 GHz desktop computer with 1 GB memory running Windows Vista.

4.3 Results and discussion

Figure 4.1 Prediction errors of single (a) GPR (b) ANN and (c) PLS model, and bagging (d) GPR (e) ANN and (f) PLS model on “wheat kernels” set.

Figure 4.1 illustrates the capability of bagging to improve model robustness on “wheat kernels” dataset. Figure 4.1(a), (b), and (c) display RMSEPs of 50 individual GPR, ANN and PLS models, respectively. The prediction performances of 50 models,

developed from the re-sampled training data with replacement, are quite different, indicating the instability of a single GPR, ANN or PLS model. Figure 4.1(d), (e), and (f) show that bagging improves calibration accuracy and robustness. It appears that the combination of 10 or more models results in satisfactory prediction. (The results, similar to Figure 4.1, are obtained for subbagging and thus not repeated here.) However, in order to obtain additional assurance, bagging/subbagging with 20 individual models are adopted subsequently.

Table 4.1 shows RMSEPs of individual calibration methods (PLS, GPR and ANN) and subbagging GPR, ANN and PLS, and the relative improvement of prediction accuracy by using subbagging with different sub-sampling ratios for “wheat kernels” data. All results (including RMSEP and its improvement, the number of LVs and computational time) are averaged over 20 repeated partitions of the dataset via SPXY algorithm. For this dataset, individual GPR and ANN give better performance than PLS, and their RMSEPs are 0.45 and 0.48, respectively. Then, the effect of different sub-sampling ratios of the training data on predictive performance is investigated. Table 4.1 shows that for the ratios of 0.1 to 0.6, RMSEPs of subbagging GPR are higher than individual GPR because the sub-sampled training data are not sufficient to cover the variation of the entire dataset. From the ratio of 0.7, RMSEP of subbagging GPR starts to be lower than individual GPR. At the ratios of 0.8 and 0.9, it attains the greatest improvement of 10.4%. Subbagging ANN gives the similar prediction performance to subbagging GPR, and also achieves satisfactory

improvement. However, subbagging PLS does not achieve better performance than PLS at any ratio. Figure 4.2 displays the RMSEP trend of subbagging GPR, ANN and PLS. It can be observed that in general, RMSEP of subbagging decreases with the increased sub-samples. The only exception is at “ratio = 1” for subbagging GPR and ANN, where RMSEP is higher than at the ratio of 0.9. In fact, “ratio = 1” essentially means no sub-sampling and the 20 models are developed from the same original data but with different initialization of parameter values. Note that for non-linear regression, parameter estimation is a non-linear optimization problem and usually results in different estimates. Combining these 20 models helps to reduce model sensitivity to the variation of parameters (but not to the variation of data). Therefore, subbagging at “ratio = 1” is not preferred.

Previous study indicated that subbagging is a useful variant of the original bagging algorithm with similar prediction performance but improved computational efficiency, because it uses less data [36]. However, the results in Table 4.1 suggest that the selection of sub-sampling ratio is crucial to the success of subbagging, and it should be accomplished by cross-validation, which consumes a large amount of time. In order to confirm this, Table 4.2 summarizes the prediction accuracy and computation time of various methods on “wheat kernels” dataset. Comparing Table 4.1 and 4.2, it is clear that if the optimal ratio is known *a priori*, subbagging gives similar accuracy but requires less computation than bagging. This advantage is especially remarkable for GPR since its computation increases in cubic order with the

increase of data [55]. However, the need of cross-validation dramatically increases the computation time of subbagging. For example, subbagging GPR and ANN by cross-validation took approximately 385.05 and 24.63 minutes, respectively, as compared with 76.51 and 0.60 minutes for bagging. Hence, in practice bagging is preferred to subbagging. For the other dataset of “meat”, only bagging is further investigated.

Figure 4.2 RMSEPs of subbagging GPR, ANN and PLS at different sub-sampling ratios on “wheat kernels” dataset.

The “meat” dataset is further used to validate the performance of bagging, whereby the results are summarized in Table 4.3. The results on “meat” dataset are similar to those on “wheat kernels”: the prediction accuracy of GPR and ANN is better than PLS; bagging GPR and ANN give superior accuracy to individual methods

in most cases; bagging PLS does not achieve significant improvement. In more details, for the moisture content, bagging attains excellent improvement of 16.9% and 11.1% on GPR and ANN, respectively. For the prediction of fat, bagging GPR shows outstanding prediction (relative improvement of 39.8%), while bagging ANN also performs satisfactorily in term of 7.2% improvement. For the protein, bagging GPR does not show any advantage; but the results of bagging ANN are favorable with 14.5% improvement.

4.5 Summary

This chapter explores the application of bagging/subbagging for non-linear calibration of NIR spectroscopy with the aim to improve prediction accuracy and model robustness. The results have confirmed the effectiveness of bagging non-linear models (GPR and ANN), while bagging linear PLS does not show significant advantage. Subbagging possesses similar prediction performance to bagging. However, in practice the sub-sampling ratio of subbagging needs to be optimized by cross-validation, which greatly increases the computation load. Therefore, bagging is recommended in practice. Clearly, by using multiple models, the computational cost at the model development stage inevitably multiplies. However, improved prediction accuracy as demonstrated in this chapter may well justify the additional computation, which is becoming an inexpensive resource with the rapid development of computers.

This work has been published in Chemometrics and Intelligent Laboratory Systems (Wang et al. Chemometrics and Intelligent Laboratory Systems, 2010, 105, 1–6); reuse in this chapter is under the permission of Elsevier Global Rights Department.

Table 4.1 Results of individual calibration methods (GPR, ANN, PLS) and those of subbagging with different ratios of sub-sampled training data on “wheat kernels” dataset.

$RMSEP_{GPR} = 0.48$; $RMSEP_{ANN} = 0.45$; $RMSEP_{PLS} = 0.49$.

Ratio		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Subbagging	RMSEP	0.76	0.74	0.72	0.68	0.59	0.51	0.44	0.43	0.43	0.48
_GPR	Improvement	-	-	-	-	-	-	8.3%	10.4%	10.4%	-
	No. of LVs	9.8	10.0	10.5	11.1	11.3	11.6	11.7	11.9	11.9	11.9
	Time (min)	2.07	2.88	3.48	4.86	7.59	10.74	26.19	38.22	44.73	64.11
Subbagging	RMSEP	0.48	0.46	0.45	0.43	0.43	0.42	0.42	0.42	0.42	0.43
_ANN	Improvement	-	-	-	4.4%	4.4%	6.7%	6.7%	6.7%	6.7%	4.4%
	No. of LVs	9.8	10.0	10.4	11.0	11.3	11.7	11.7	11.9	11.9	11.9
	Time (min)	0.51	0.52	0.50	0.50	0.53	0.53	0.55	0.56	0.59	0.66

Subbagging	RMSEP	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
_PLS	Improvement	-	-	-	-	-	-	-	-	-	-
	No. of LVs	9.8	10.0	10.5	11.0	11.3	11.6	11.8	11.9	11.9	11.9
	Time (min)	0.22	0.23	0.24	0.25	0.27	0.26	0.27	0.27	0.29	0.31

Table 4.2 A summary of results for individual calibration methods, bagging, and subbagging by cross-validation on “wheat kernels” dataset.

Methods	GPR	Subbagging	Bagging	ANN	Subbagging	Bagging	PLS	Subbagging	Bagging
		_GPR^(CV)	_GPR		_ANN^(CV)	_ANN		_PLS^(CV)	_PLS
RMSEP	0.48	0.43	0.41	0.45	0.42	0.42	0.49	0.49	0.49
Improvement	-	10.4%	14.6%	-	6.7%	6.7%	-	0%	0%
No. of LVs	12.0	11.9	18.0	12.0	11.8	18.9	11.9	11.7	18.6
Time (min)	2.82	385.05	76.51	0.03	24.63	0.60	0.02	7.47	0.30

Table 4.3 Summary of results for individual calibration methods (GPR, ANN, PLS) and bagging GPR, ANN, PLS on “meat” dataset: (a) Moisture; (b) Fat; (c) Protein.

(a)

Methods	GPR	Bagging_GPR	ANN	Bagging_ANN	PLS	Bagging_PLS
RMSEP	0.77	0.64	0.99	0.88	2.68	2.61
Improvement	-	16.9%	-	11.1%	-	2.6%
No. of LVs	16.2	14.9	16.0	19.0	17.1	19.5
Time (min)	0.30	9.96	0.03	0.72	0.01	0.25

(b)

Methods	GPR	Bagging_GPR	ANN	Bagging_ANN	PLS	Bagging_PLS
RMSEP	0.83	0.50	0.83	0.77	3.02	2.96
Improvement	-	39.8%	-	7.2%	-	2.0%
No. of LVs	16.2	14.7	16.6	19.3	16.5	18.4
Time (min)	0.30	9.15	0.03	0.78	0.01	0.24

(c)

Methods	GPR	Bagging_GPR	ANN	Bagging_ANN	PLS	Bagging_PLS
RMSEP	0.40	0.41	0.55	0.47	0.54	0.54
Improvement	-	-2.5%	-	14.5%	-	0%
No. of LVs	14.3	14.0	14.4	16.4	14.3	16.1
Time (min)	0.27	9.63	0.03	0.63	0.01	0.24

Chapter 5 Penalized Regression Methodology with Variable Selection

In Chapters 3 & 4, different chemometric techniques are investigated to improve the prediction accuracy of multivariate calibration, including pre-processing techniques, non-linear calibration methods, and bagging/subbagging for non-linear calibration methods. In these techniques, all variables are employed as predictors in the calibration model. However, not all variables provide useful information for the prediction of the analyte property in the sense that some variables may have no contribution or even negative contribution to the calibration model. Furthermore, when all variables are included in the calibration model, the dimension is very high (usually several hundred), leading to heavy computation load and high variance in parameters. Therefore, selecting informative variables becomes crucial, since it can not only improve prediction accuracy, but also reduce calculation burden. However, selecting variables is not always feasible, since traditional variable selection algorithms such as UVE require complex calculation procedure, requiring heavy computation as well. Therefore, the variable selection algorithm with less computation load needs to be developed. In this chapter, two penalization algorithms with variable selection by using LASSO, penalized linear regression (PLR) and penalized Gaussian process regression (PGPR), are investigated. In these two algorithms, variable selection is carried out in the process of estimating model parameters, making the computation procedure become more simple and applicable.

These two algorithms are similar in the way that LASSO is applied to penalize the

model parameters which are estimated by minimizing the target function in Eq. (2.28) and Eq. (2.32), respectively. In PLR, LASSO is used to penalize the linear regression parameters $\boldsymbol{\eta}$, while in PGPR LASSO is used to penalize the hyper-parameters $\boldsymbol{w}$. In these two penalization techniques, by limiting the absolute length, magnitude of some parameters will become small while others may become very large. The larger of the magnitude is, the more relevant the corresponding variable is to the analyte property. Small values indicate that the corresponding variables might be nuisance factors. By setting certain smaller regression parameters to zero, unrelated variables can be eliminated. After the regression parameters with small estimated magnitude are removed and large ones are kept, regression is repeated. This is the principle of penalized regression with variable selection, in which variable selection and parameter estimation are coped with simultaneously. In this chapter, in order to illustrate the effectiveness of variable selection in penalized regressions, the prediction performance of PLR, PLR with variable selection (PLR-VS), PGPR and PGPR with variable selection (PGPR-VS) are compared. The difference between penalized regression with variable selection and penalized regression is that in penalized regressions with variable selection, LASSO is first applied to select the important variables which have large magnitude, and then only the selected variables are applied in the regression algorithm, while in penalized regression, unrelated variables are not eliminated before applying the regression algorithm. Without variable selection, penalized regression requires more computation since more regression parameters are included in the calibration model.

5.1 Computation issues

This study is evaluated on “wheat kernels” NIR dataset as done in the previous chapters. This dataset is divided into the training set with 415 samples and testing set with 108 samples as in the original paper [25]. Since the spectra in this dataset are deeply influenced by light scattering effect, EMSC is first used to pre-process the raw data, and then PLR or PGPR is applied. For benchmark comparison, PLS pre-processed by EMSC is also investigated, in which the number of latent variables (LVs) are selected by five-fold random-splitting cross-validation. In PGPR, the log values of hyper-parameters are used to avoid the constrained optimization problem. In addition, when optimizing the hyper-parameters in PGPR, it can be easily trapped in local optima. Therefore, the optimization process is repeated five times, and the estimates with the smallest l_p in Eq. (2.32) are selected as the final results.

The Lagrange parameter κ is optimized by five-fold random-splitting cross-validation. When minimizing the target function to estimate model parameters, it is important to select the appropriate initial values of parameters in order to obtain stable convergence. In PLR, the minimum length least squares (MLLS) estimate is a good choice for the initial values of $\boldsymbol{\eta}$ [38]. The MLLS estimate is expressed as $\boldsymbol{\eta}_0 = (\mathbf{X}^T \mathbf{X})^- \mathbf{X}^T \mathbf{y}$, where $(\mathbf{X}^T \mathbf{X})^-$ is Moore-Penrose inverse. In PGPR, the initial log values of hyper-parameters w_j are set to zero. Since the response analyte is only related to a few predictors, the values of the significant nonzero parameters will deviate from zero very fast, while for the nuisance parameters, their value will stay at near zero.

5.2 Results and discussion

Figure 5.1 The regression parameters in EMSC-PLS. (X-axis represents the number of the parameters, while Y-axis is the value of the parameters.)

Through five-fold random-splitting cross-validation method, the Lagrange parameter κ in EMSC-PLR is optimized at 9.2×10^{-4} , and κ in EMSC-PGPR is determined to be 0.4528. For comparison, the regression parameters in PLS pre-processing by EMSC are displayed in Figure 5.1. It can be observed that the absolute values of very few parameters are close to zero. Figure 5.2 (a) displays the regression parameters η in EMSC-PLR. From the figure, it can be seen that the absolute values of some estimates are very large (from several hundred to several thousand), while some are very small (close to zero). In EMSC-PLR-VS, the estimates whose absolute values are less than 10^{-3} are set to zero as shown in Figure 5.2(b). The threshold of 10^{-3} is selected according to its influence on the balance of model prediction and computation load. If the threshold is $< 10^{-3}$, it needs more

Figure 5.2 The regression parameters η in (a) EMSC-PLR; (b) EMSC-PLR with variable selection. (X-axis represents the number of the parameters, while Y-axis is the value of the parameters.)

Figure 5.3 The hyper-parameters w in (a) EMSC-PGPR; (b) EMSC-PGPR with variable selection. (X-axis represents the number of the parameters, while Y-axis is the value of the parameters.)

computations but with almost the same prediction performance, while if the threshold is $> 10^{-3}$, prediction performance could become worse although with less computation.

After variable selection, the number of predictors are reduced from 100 to 45. The

hyper-parameters $\mathbf{w}$ in EMSC-PGPR are depicted in Figure 5.3(a). Different from the parameters $\boldsymbol{\eta}$ in EMSC-PLR, the values of $\mathbf{w}$ are much larger and few of them are close to zero. This may be because the log values of hyper-parameters are adopted to avoid the constrained optimization problem. However, some of them still have especially large values compared with others. The values of hyper-parameters, which is less than 10^6 , are set to 10^0 as shown in Figure 5.3(b). The selection principle of the threshold 10^6 is the same as EMSC-PLR. In EMSC-PGPR-VS, 46 predictors remain. From these figures, the effectiveness of penalized methods can be observed in terms of eliminating uninformative variables compared with the traditional regression technique of PLS.

The predictive performance of EMSC-PLR, EMSC-PLR-VS, EMSC-PGPR, and EMSC-PGPR-VS are compared with the baseline methods of PLS and EMSC-PLS, which are displayed in Table 5.1. From the table, it can be observed that light scattering effect in the dataset is largely removed by EMSC, since RMSEP is reduced from 0.63 of PLS to 0.41 of EMSC-PLS. EMSC-PLR achieves superior prediction performance than EMSC-PLS according to its RMSEP of 0.39. However, EMSC-PGPR just gives the same prediction performance since its prediction error is the same as EMSC-PLS. After eliminating some unrelated estimates, penalized regressions with variable selection (both EMSC-PLR-VS and EMSC-PGPR-VS) give slightly better prediction performance than EMSC-PLR and EMSC-PGPR, since their prediction errors are reduced 0.01. This indicates the success of penalized regressions with variable selection that they can achieve a little better prediction results with fewer predictors.

Table 5.1 RMSEPs of EMSC, EMSC-PLR, EMSC-PLR-VS, EMSC-PGPR, and EMSC-PGPR-VS.

Methods	PLS	EMSC-PLS	EMSC-PLR	EMSC-PLR-VS	EMSC-PGPR	EMSC-PGPR-VS
RMSEP	0.63	0.41	0.39	0.38	0.41	0.40

5.3 Summary

Traditionally, variable selection and parameter estimation have been usually treated separately, and the algorithms of variable selection are usually very fuzzy. In this chapter, two techniques of penalized regressions with variable selection (PLR-VS and PGPR-VS) that can deal with variable selection and parameter estimation simultaneously are investigated. After eliminating unrelated predictors, EMSC-PLR-VS gives better prediction results than the benchmark calibration method of EMSC-PLS. However, the prediction performance of EMSC-PGPR-VS is not superior to EMSC-PLS. The reason may be that the selected linear covariance function in Eq. (2.30) cannot account for the relationship between the spectra and analyte property accurately. Therefore, more suitable covariance functions need to be further studied in the future.

Chapter 6 Chemometric Determination of the Length Distribution of Single Walled Carbon Nanotubes

6.1 Introduction

Single walled carbon nanotubes (SWCNTs) are hollow cylinders with one atomic layer of graphene wall, lengths ranging from tens of nanometers to centimeters. SWCNTs have versatile applications including transistors, conductive films, sensors, probes, mechanical reinforcements, hydrogen storage and catalytic supports because of their excellent electrical, mechanical, thermal, optical and biological properties [94]. The macroscopic properties of SWCNTs depend on the underlying atomic characteristics, such as diameter, chirality and length [95]. Current synthesis methods do not ensure the uniformity of the structure and properties of SWCNTs, a fact that has hindered various potential applications. To overcome this barrier, significant research efforts have been dedicated to sorting SWCNTs according to diameter, metallicity and chirality [95-97]. Separation by length is especially useful because the electronic, optical and chemical properties of SWCNTs are significantly affected by their length [98-100]. Length sorting may be also necessary as a preliminary step to obtain high purity monodisperse samples [101, 102]. Furthermore, toxicity of SWCNTs is often associated with their length. Length-dependent SWCNT uptake by cells has been observed [103].

In recent years, Fagan *et al.* have observed length-dependent optical effects in SWCNTs [99]. In general, when a light beam encounters particles, such as nanotubes,

both scattering and absorption can consume energy, a phenomenon known as light extinction [104]. The law of conservation of energy requires that $\sigma_{ext} = \sigma_a + \sigma_s$, where σ_{ext} , σ_a , and σ_s are the cross sections of particles from extinction, absorption, and scattering, respectively [104]. Having the dimension of the area, cross sections reflect the probability for an event to occur. When a particle's radius (r) is comparable to the wavelength of light (λ), Mie theory [104] indicates that σ_s varies in the order of r^6 , while σ_a varies in the order of r^3 only. Therefore, for nano-scale particles, absorption is predominant over scattering [105]. This has been reported in both experimental and theoretical studies [105]. For instance, σ_s/σ_a of gold nanospheres with radius of 10 nm under 521 nm light is only 0.01 [106]. A recent study has showed that scattering from larger diameter multiple walled carbon nanotubes (MWCNTs) is minor ($< 8\%$) in light extinction over the wavelength range of 300-1300 nm [107]. Thus, it is reasonable to hypothesize that absorption spectroscopy may be an effective tool to measure the length of SWCNTs. Indeed, Fagan *et al.* have showed that optical absorption of SWCNTs has a nearly linear dependence on the mean (average) length [99]. The absorption peak intensity at optical resonance (absorbance at 984 nm) verses absorbance baseline (at 775 nm) has been found to scale almost linearly with the mean length of nanotubes up to 1 μm (note that the diameter of SWCNTs is around 1 nm only). A simple physical argument has been also proposed to relate the length dependence in SWCNT absorption to the localization of a bound exciton along the backbone of nanotubes, suggesting a quantitative link between the nanotube length and oscillator strength for quasi-one-dimensional ordered nanostructures [99]. In the Rayleigh regime, the bound exciton is localized on a length scale much smaller than the length of nanotube, implying that the imaginary

part of SWCNT dielectric response function is proportional to the nanotube length. Later, an empirical formula, given below, has been applied to predict the mean length of SWCNTs obtained from length sorting based on their optical absorbance [108]:

$$l(\text{nm}) = \left(\frac{\text{Absorbance (984 nm)}}{\text{Absorbance (775 nm)}} - 0.842 \right) \times 160.4 \text{ nm} \quad (6.1)$$

However, because of overlapping of optical absorbance from SWCNTs at different lengths, predicting length distribution has not been explored in this literature.

Multivariate chemometric calibration methods can utilize measured spectra to predict the corresponding analyte properties (SWCNT length distribution in this work). For this purpose, partial least squares (PLS) regression is a widely used linear calibration method. However, non-linear variations can be induced in spectra because of external disturbances, mainly baseline shift/noise in this study. This is usually the results of differences in sample homogeneity, temperature variation during analysis, and so on, giving rise to non-linear spectral data. In the literature, baseline and other noises can be addressed by using pre-processing methods and/or non-linear calibration models [109]. Typical pre-processing methods to remove non-linearity in spectral data include the first (D1) and second derivatives (D2) [17, 23], standard normal variate (SNV) [24], extended multiplicative signal correction (EMSC) [21], and extended inverted signal correction (EISC) [25]. Alternatively, non-linear calibration methods, such as artificial neural network (ANN) [27-29], and Gaussian process regression (GPR) [10], are capable of directly modeling the non-linearity in spectra.

Transmission electron microscope (TEM) and atomic force microscope (AFM) are the most widely used methods to quantify the length distribution of SWCNTs currently. However, sample preparation for TEM and AFM is often time-consuming

and tedious. Visual inspection and image processing can also introduce errors. In this chapter, the length distribution of sorted SWCNTs is first determined by AFM and used as the reference value for calibration. Their absorbance spectra are then measured in ultraviolet-visible near-infrared (UV-vis-NIR) absorption spectrometer. An existing method [108] to determine the mean length of SWCNTs from absorption spectra is extended in this work to predict the entire length distribution. In addition, advanced multivariate chemometric methods are adopted and compared for calibration purpose, in contrast to using an empirical formula of Eq. (6.1) in the previously reported work [108]. The results show that excellent prediction accuracy can be achieved by using chemometric techniques. In summary, absorption spectroscopy in conjunction with proper calibration technology is an effective analytical tool to complement the time-consuming, laborious and expensive AFM (or TEM) analysis for SWCNT research.

6.2 Dataset

SWCNTs (8, 10, and 12 mg, respectively) synthesized by CO decomposition on cobalt-molybdenum catalysts were dispersed in 10 mL of 2 wt% aqueous sodium deoxycholate (DOC) solution. Therefore, three concentrations (0.8 mg mL^{-1} , 1.0 mg mL^{-1} , and 1.2 mg mL^{-1}) of SWCNTs could be obtained. Then SWCNT suspensions were centrifuged to remove large nanotube bundles and other impurities. After centrifugation, supernatants of SWCNT suspensions containing individualized nanotubes were further sorted by length using the density gradient ultracentrifuge (DGU) method [108, 110, 111]. Afterwards, dispersion in the centrifugation tube was

extracted in ten individual fractions which are labeled from “1f” (the topmost layer) to “10f” (the bottom layer), respectively.

AFM was employed to characterize the length distribution of sorted SWCNT fractions. UV-vis-NIR spectra were recorded on a Varian Cary 5000 UV-vis-NIR spectrophotometer from 400 to 1350 nm with 1 nm increments. Therefore, 951 variables were obtained in this dataset. In total, length distributions of 21 length-sorted SWCNT samples were collected, which were from three different initial SWCNT concentrations: eight samples (3f-10f) from 0.8 mg mL⁻¹, seven samples (4f-10f) from 1.0 mg mL⁻¹, and six samples (5f-10f) from 1.2 mg mL⁻¹. The three concentrations correspond to three batches of centrifugation, thus referred to as “batches” hereafter, whilst the extracted layers of SWCNTs are termed “samples”. Some topmost layers from sorted samples obtained using higher initial SWCNT concentrations were discarded, because those sorted layers have a poor separation resolution due to the increase of the initial SWCNT concentration.

6.3 Chemometric prediction of the length distribution

Chemometric methods are applied to predict the length distribution of sorted SWCNTs from the corresponding absorption spectroscopy. Five pre-processing techniques (D1, D2, SNV, EMSC, and EISC) in conjunction with PLS and two non-linear calibration models (ANN and GPR) are evaluated to search for the optimal method. The results of PLS without pre-processing are reported as benchmark. The number of latent variables (LVs) in PLS is selected by five-fold random-splitting cross-validation. In addition, a simple “average prediction” method is used as the

bottom-line technique to test the usefulness of sophisticated calibration models, where the average of analyte properties of the calibration dataset is used as prediction. Similar to Chapters 3 & 4, PLS is also applied to reduce the dimension of the data prior to developing non-linear calibration methods of ANN and GPR.

The method of leave-one-out cross-validation (LOO-CV) is used to evaluate the chemometric techniques. In LOO-CV, one sample is chosen for prediction, while the remaining 20 samples are used as the reference data for developing the calibration model. This step is repeated until each sample has been used once for prediction. However, the left-out sample is related to other samples from the same centrifugation batch and is thus not independent. To resolve this issue, the method of leave-one-batch-out cross-validation (LOBO-CV) is also investigated, where samples from one batch are treated as the testing data while samples from the other two batches are used for developing the calibration model. This is repeated three times so that samples from each centrifugation batch have been used once for testing. LOO-CV and LOBO-CV rather than random splitting used in Chapter 3 and SPXY algorithm used in Chapter 4 are used in this chapter, since samples are so limited (only 21 samples) in this dataset that randomly splitting and SPXY algorithm are not reliable.

From the measured nanotube length by AFM, it is found that the length approximately follows lognormal distribution. Indeed, lognormal distribution is widely accepted to describe the distribution of size/length of particles [112-114]. A lognormal distribution is solely determined by two parameters: the mean (μ) and standard deviation (σ) of natural logarithm of a random variable (i.e. the length). Therefore, μ and σ are treated as analyte properties to be predicted by chemometric

calibration models.

Relative root mean square error (RRMSE) and coefficient of determination (R^2) are used to evaluate the performance of different calibration methods; RRMSE is defined as

$$\sqrt{(1/I) \sum_{i=1}^I ((y_i - \hat{y}_i)/y_i)^2} \quad (6.2)$$

where y_i is the actual analyte (either μ or σ) for the i^{th} sample determined by AFM, $\hat{y}_i$ is the chemometric prediction from spectral data, and I is the number of samples. RRMSE instead of RMSE is adopted in this chapter since from RRMSE it can be more obvious to see how much the predicted analyte deviates from the measured analyte. In previous chapters, comparison of the prediction performance of different calibration strategies is the main investigation purpose, which can be seen from their RMSEs. However, in this chapter, besides comparing the prediction performance of different calibration methods, it also would like to know the difference between the predicted and measure analyte property. As a result, RRMSE is preferred. R^2 is a selection criterion of regression models, which is very useful when it is unknown what type of model a dataset belongs to. In the previous chapters, R^2 is not given since in many literatures these calibration models have already been demonstrated to be suitable to account for the datasets. However, in this chapter, it is unknown whether these calibration models are suitable to the SWCNT data, thereby R^2 is investigated. RRMSE and R^2 on the parameters of log-normal distribution are indirect measures for prediction accuracy, since it is the distribution itself that is of ultimate interest. Another measure, Jensen-Shannon (J-S) divergence originally developed in the community of information theory and statistics [38], is used to directly quantify the

difference between the predicted and reference distributions. For continuous distributions with density functions $p(x)$ and $q(x)$, e.g. $p(x)$ being the reference (true) distribution and $q(x)$ being the predicted distribution, J-S divergence is defined as

$$D(p, q) = \frac{1}{2} \int \left[p(x) \log \frac{2p(x)}{p(x) + q(x)} + q(x) \log \frac{2q(x)}{p(x) + q(x)} \right] dx \quad (6.3)$$

J-S divergence is between zero and unity; zero means that two distributions are exactly the same whilst one means they are extremely different.

6.4 Results and discussion

The UV-vis-NIR spectra of SWCNT fractions extracted from the centrifugation tube are shown in Figure 6.1. Sorted SWCNTs show clear length dependence of the absorption spectra, similar to the previously reported findings [99, 108]. Longer SWCNTs display stronger optical feature. All absorption data are scaled to the same concentration at 775 nm, following the method proposed by Fagan et al. [99]. The 775 nm is selected as the reference for scaling because there is little contribution from any SWCNTs or graphitic carbon feature to the absorption around this wavelength. Such scaling helps alleviate the impact of SWCNT concentration on spectra, and thus, the scaled absorption spectra are mainly a function of the length of SWCNTs. The decrease of absorption intensity above 1200 nm comes from density medium. The histogram and fitted length distributions of nine selected nanotube samples obtained by AFM studies are presented in Figure 6.2. AFM results confirm that the length of these sorted SWCNTs follows a lognormal distribution. Afterwards, different chemometric calibration techniques are applied to predict the mean and standard

deviation of SWCNTs' log-length based on their scaled absorption spectra from 400 to 1350 nm.

Figure 6.1 Absorbance spectra of length-sorted SWCNT fractions after scaling at 775 nm. Note that “1f” and “2f”, corresponding to samples from the top of the centrifugation tube, contains few nanotubes, and thus their spectra are removed from analysis. The absorption spectra of the starting material before separation (dashed line and denoted “AP”) is shown as a reference.

Figure 6.2 The length distribution of sorted SWCNT samples obtained from AFM analysis. The red solid line represents the fitted log-normal distribution. The bars represent the histogram from AFM measure.

Five linear calibration methods (D1, D2, SNV, EMSC, and EISC) and two non-linear calibration methods (ANN and GPR) are evaluated to find the optimum calibration method. Table 6.1 shows the number of latent variables (LVs), RRMSE, R^2 and J-S divergence of different chemometric calibration techniques by using the method of LOO-CV. The values of J-S divergence in the table are the average of all testing samples. All methods are better than the baseline average prediction (labeled “Ave. Pred.” in the table). The measures of RRMSE and R^2 indicate that EMSC, EISC and GPR attain outstanding accuracy in predicting the log-mean length (GPR being the best according to its RRMSE). In addition, EMSC and EISC achieve superior performance for the log-standard deviation (EISC being slightly better). A general trend in Table 6.1 is that the prediction of the mean is more accurate than the standard deviation. This phenomenon suggests that prediction of the first-order information carried by the mean is easier than that of the second-order information carried by the

standard deviation, which is consistent with statistical intuitions. In addition, lower-order information usually has higher influence on the entire probability distribution. The overall prediction accuracy can be assessed by the J-S divergence, which suggests that EMSC, EISC and GPR are effective calibration methods for this dataset. The effectiveness of EISC as a representative pre-processing method is further illustrated in Figure 6.3, which clearly shows that baseline shift/noise has been largely removed, giving rise to improved calibration models.

Figure 6.3 The spectra before and after pre-processing by EISC.

Table 6.1 The LOO-CV calibration results.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	5.7	5.2	5.8	5.7	10.1	10.9	5.7	5.7	-
	RRMSE	20.5%	18.3%	36.0%	21.6%	7.3%	7.2%	22.3%	6.7%	48.8%
	R^2	0.78	0.78	0.05	0.75	0.95	0.95	0.77	0.93	-0.33
Standard Deviation	LVs	3.9	2.3	3.3	3.9	5.4	4.4	3.9	3.9	-
	RRMSE	23.9%	22.2%	21.1%	24.7%	13.3%	12.0%	35.0%	22.5%	57.8%
	R^2	0.41	0.37	0.34	0.40	0.83	0.85	0.01	0.42	-20.71
J-S Divergence		0.073	0.075	0.221	0.065	0.021	0.020	0.074	0.032	0.318

Using the predicted log-mean by GPR and log-standard deviation by EISC, the lognormal distribution of SWCNT length is recovered and displayed in Figure 6.4, overlaid with the histogram and fitted reference distribution from AFM measure. The J-S divergence values are also given for each sample. Visual inspection reveals that most distributions are well predicted. In addition, all J-S divergence measures, except for sample 20, are less than 0.1 and the majorities are fairly close to zero, indicating excellent calibration results.

Figure 6.4 LOO-CV calibration results where the log-mean is predicted by GPR and log-standard deviation by EISC. The red solid line represents the reference fitted distribution, while the blue dotted line indicates the predicted distribution. The bars represent the histogram from AFM measure.

Next, the results of LOBO-CV are reported. This is a more challenging scenario since the amount of data for calibration modeling is very limited. The results are summarized in Table 6.2(a), (b), and (c), corresponding to the testing samples from three batches with SWCNT concentration of 0.8, 1.0 and 1.2 mg mL⁻¹, respectively. Clearly, the prediction of the left-out batches is not as accurate as that of the left-out samples, and the prediction of log-standard deviation seems especially disappointing. In particular, some R^2 values become negative, suggesting that the sophisticated models are even worse than simply using the average analyte properties of testing data for prediction, which would give $R^2 = 0$. However, the average of testing data is

not available *a priori* when a calibration model is used in practice. Instead, only the average of *calibration* data can be used for “prediction”, and the results are given in the last column of Table 6.2. The R^2 values of this averaging method are also negative, especially for the log-standard deviation of the length distribution. More importantly, compared with this bottom-line method, it appears that GPR is still effective in predicting the log-mean of the length distribution, whilst EMSC and EISC provide satisfactory results on the log-standard deviation. The measure for accuracy of the overall distribution, J-S divergence, also indicates the improved results of calibration modeling. Nevertheless, whether the improvement over the baseline averaging method is practically useful should be carefully assessed. Figure 6.5 illustrates the predicted length distributions with LOBO-CV, where the log-mean is predicted by GPR and the log-standard deviation is predicted by EISC. There exists significant mismatch between the predicted and reference distribution for a few samples, the most pronounced being sample 1. However, for most samples, the overall location and shape characteristics of the length distribution have been properly captured, and the J-S divergence values are mostly within 0.1, indicating that calibration models are satisfactory in practice. In addition, the calibration accuracy is expected to be significantly improved if more batches of data can be collected.

Figure 6.5 LOBO-CV calibration results where the log-mean is predicted by GPR and log-standard deviation by EISC.

Table 6.2 The LOBO-CV calibration results.

(a) Testing samples from the concentration of 0.8 mg mL⁻¹.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	7	2	4	7	5	6	7	7	-
	RRMSE	58.6%	31.3%	54.0%	60.3%	36.1%	29.7%	59.8%	26.9%	48.8%
	R²	-0.55	0.40	-0.20	-0.64	0.41	0.59	-0.62	0.68	-0.33
Standard Deviation	LVs	3	2	3	3	4	4	3	3	-
	RRMSE	28.5%	28.1%	43.1%	48.1%	17.8%	19.8%	53.4%	54.0%	57.8%
	R²	-4.55	-4.34	-13.36	-16.91	-1.39	-1.85	-17.45	-18.94	-20.71
J-S Divergence		0.301	0.208	0.204	0.266	0.163	0.129	0.289	0.118	0.318

(b) Testing data: samples from the concentration of 1.0 mg mL⁻¹.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	4	5	3	4	4	4	4	4	-
	RRMSE	9.9%	41.2%	57.3%	9.3%	16.2%	21.8%	26.4%	8.3%	26.1%
	R²	0.86	-0.12	-2.16	0.82	0.58	0.33	-0.13	0.89	-0.21
Standard Deviation	LVs	5	4	4	5	4	4	5	5	-
	RRMSE	28.7%	30.8%	31.0%	29.6%	29.9%	31.1%	43.8%	43.2%	43.2%
	R²	-11.76	-13.71	-14.46	-12.34	-13.10	-14.13	-30.00	-11.76	-28.69
J-S Divergence		0.055	0.150	0.380	0.063	0.122	0.172	0.258	0.114	0.262

(c) Testing data: samples from the concentration of 1.2 mg mL⁻¹.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	5	8	7	5	7	7	5	5	-
	RRMSE	6.3%	11.2%	15.7%	7.4%	8.6%	10.4%	5.1%	6.1%	20.0%
	R²	0.90	0.74	0.57	0.87	0.89	0.82	0.94	0.91	-0.23
Standard Deviation	LVs	4	2	3	4	4	5	4	4	-
	RRMSE	26.6%	18.0%	18.6%	27.4%	19.7%	16.2%	31.9%	29.5%	32.0%
	R²	-2.38	-0.33	-0.47	-2.31	-0.71	-0.43	-3.25	-2.38	-3.25
J-S Divergence		0.040	0.057	0.091	0.043	0.041	0.051	0.035	0.035	0.165

From the results of LOO-CV and LOBO-CV, it appears that the prediction for mid-layer samples (5f-8f) at moderate “post-DGU” concentration (the actual concentration after DGU treatment and it can be inferred from the absorbance at 775 nm) are superior to samples from the top or bottom layers at either very low (absorbance at 775 nm < 0.01) or very high concentration (absorbance at 775 nm > 0.32). The relatively poor results at the low- and high-end post-DGU concentration may be the result of model extrapolation. In general, comfortable prediction performance has been achieved.

6.5 Summary

In this work, the length distribution is measured by AFM as a reference method, and it follows a lognormal distribution. Sorted SWCNT suspensions have length-dependent UV-vis-NIR absorption spectra, which provided the opportunity to calibrate spectroscopy against AFM-measured length distribution for rapid analysis. Using absorption spectral data, five linear calibration methods (D1, D2, SNV, EMSC, EISC) and two non-linear calibration methods (ANN and GPR) are evaluated in terms of prediction accuracy. Two chemometric calibration techniques (GPR and EISC) are able to predict the length distribution with satisfactory agreement with AFM measure. Two evaluating methods of LOO-CV and LOBO-CV are investigated. The prediction performance of LOBO-CV method is not as good as LOO-CV. This is largely due to the limited amount of data (only 13-15 samples) for calibration. Overall, the predicted distributions of most samples capture the major characteristics of the measured distributions, which is the foremost goal of this work.

It should be noted that there are several limitations of the current technique. Spectroscopy, once properly calibrated, is to complement but not completely replace the time-consuming and costly AFM measure. The calibration procedure itself requires centrifugation, sample preparation, and AFM analysis to obtain the reference data. Proper execution of the calibration stage is crucial to the success of prediction models. In addition, SWCNT dispersions with very high or low concentration as discussed above may induce errors in their length distribution prediction by absorption spectra, which is a well-known extrapolation problem encountered in all data-based chemometric techniques. Extrapolation could also be an issue for SWCNTs taken from different sources, in which situation a new calibration model may be required if samples have significantly different diameters or lengths.

This work has been completed by cooperating with Ms Si Rongmei. She is responsible for the experiments to obtain the dataset of UV-vis-NIR spectroscopy and SWCNT length from AFM, and scaling the spectra at 775nm. I am in charge of calibration model development and data analysis.

This work has been published in *Analytica Chimica Acta* (Wang, et al. *Analytica Chimica Acta*, 2011, 708, 28-36); reuse in this chapter is under the permission of Elsevier Global Rights Department.

Chapter 7 Conclusions and Future Work

7.1 Conclusions of current work

In this thesis, various chemometric multivariate calibration techniques and their applications of using the information in measured spectroscopic data to predict analyte properties are investigated. Firstly, in order to deal with the non-linearity induced by external fluctuations which traditional linear calibration techniques of PLS and PCR cannot handle, different chemometric strategies are investigated and compared to improve prediction accuracy. These techniques include linear techniques of D1, D2, SNV, EMSC, EISC, OPLEC, and non-linear calibration methods of ANN, GPR, and LS-SVM. Through comparison, it is found that non-linear calibration techniques could achieve superior predictive performance to linear approaches in most cases to deal with external disturbances. However, prediction results are not sufficiently robust in the sense that small changes in data or model parameters can lead to significant changes in model prediction.

As a result, the strategy of bagging/subbagging for non-linear calibration methods (specifically ANN and GPR) is studied to address this issue. Through two NIR spectroscopic datasets, it is proved that bagging/subbagging can attain more accurate and robust predictive performance. However, subbagging is sensitive to the amount of sub-sampled data, which needs to be determined by the computationally intensive cross-validation method. Therefore, bagging is preferred to subbagging in practice.

In the construction of calibration model, variable selection is also an important

way to improve prediction accuracy, since the beneficial variables that are related to the analyte property are retained while nonsense variables are eliminated. Therefore, penalized regression techniques with variable selection of PLR-VS and PGPR-VS are investigated. It is demonstrated that PLR gives better prediction result than the benchmark calibration method of PLS. However, the prediction performance of PGPR is not superior to PLS since unsuitable covariance function is employed. After eliminating unrelated predictors, PLR-VS and PGPR-VS still give the same prediction results as PLR and PGPR.

Finally, chemometric multivariate calibration techniques are applied to use UV-vis-NIR spectroscopy to predict the length distribution of single walled carbon nanotubes, which follows a lognormal distribution. GPR and EISC are found to obtain satisfactory prediction performance in the prediction of mean and standard deviation of lognormal distribution, respectively. In this way, spectroscopy combined with multivariate calibration provides a rapid and powerful analytical tool for SWCNT research.

7.2 Perspectives for future work

Traditional calibration techniques may be referred to as “supervised” regression, constructed by “labeled” samples (samples with both spectra and analyte properties obtained through reference methods). In the application of spectroscopy, a large number of spectral data are easy to be recorded, while the acquisition of analyte properties (e.g. concentration) is expensive. This is especially true in on-line applications when spectra can be recorded continuously, where a large number of

“unlabeled” samples are available. (Unlabeled samples have spectra but no analyte property). Hence, in order to utilize unlabeled samples to improve prediction accuracy, semi-supervised regression model [115] can be applied.

Semi-supervised learning originates from the field of machine learning, using both labeled and unlabeled data for model development [116]. It has been proved that incorporating unlabeled data can achieve considerable improvement in prediction accuracy [117]. However, until now most research on semi-supervised learning focuses on classification, and reports in regression are limited [115]. It is planned to introduce semi-supervised regression to calibration modeling in order to make use of the unlabeled spectral data to improve the predictive accuracy.

Co-training is a typical semi-supervised learning techniques, in which two models are separately trained on two attribute datasets in different and independent views [118]. When co-training is used in regression, an algorithm called ‘COREG’ has been developed [115]. Let $L = \{(x_1, y_1), \dots, (x_L, y_L)\}$ represent the labeled dataset, where L is the number of the labeled samples. Similarly, let $U = \{(x_1), \dots, (x_U)\}$ denote the unlabeled dataset, where U is the number of the unlabeled samples. In semi-supervised regression with co-training algorithm, the initial step is to develop two calibration models using labeled dataset L by adopting different model parameters. Afterwards, certain unlabeled samples, which are capable of reducing RMSEP of the calibration models, are selected. For each calibration model, the selected unlabeled samples can be labeled as $U_{label} = \{(x_1, \hat{y}_1), \dots, (x_{|U|}, \hat{y}_{|U|})\}$ based on the prediction of the analyte $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_{|U|}\}$, where $|U|$ is the number of the selected unlabeled samples. Thus two different augmented training datasets can be formed by incorporating U_{label} with the labeled set L for each model. Then, another

two calibration models can be constructed from the two augmented training datasets. Finally, predictions of the two calibration models can be averaged as the final prediction [115].

In addition, a Bayesian semi-supervised approach has been introduced, which relied crucially on a fully-specified probability model of a joint distribution $p(x, y | \theta)$ [119]. The joint distribution can be expressed as $p(x, y | \theta) = p(y|x, \theta)p(x|\theta)$, where $p(y|x, \theta)$ is a conditional distribution based on labeled data, $p(x|\theta)$ is a joint distribution based on unlabeled data, and θ is the model parameter to be optimized. Then, Bayesian model can be built by taking a convex combination of $\log p(y|x, \theta)$ and $\log p(x|\theta)$ as $\alpha \log p(y|x, \theta) + (1 - \alpha) \log p(x|\theta)$. By tuning an appropriate parameter α , the prediction can be made accurately.

Therefore, in the future, it is intended to explore the techniques of semi-supervised regression to make use of unlabeled spectroscopic data to improve the predictive accuracy of calibration models.

Appendix A Soft Sensor Design for an Industrial Anaerobic Wastewater Treatment Process

A.1 Introduction

This thesis mainly investigates spectroscopy in combination with chemometric multivariate calibration techniques of making use of measured spectra to predict the analyte property. However, multivariate calibration also has other applications. For example, it can relate measured process variables (such as pressure, flow, and temperature) to analyte properties [3]. This appendix utilizes the measured process variables of wastewater (WW) treatment to predict the analyte property, which is a promising complement of the application of chemometric multivariate calibration.

Over the last century, water pollution caused by industrial discharge can result in considerable impact on the environment, since it causes degradation of various ecosystems on which human life relies. Therefore, WW treatment of removing contaminants and producing disposable effluent with no harm to the surrounding environment become essential for sustainable human development. Through WW treatment system, deleterious substances require to be handled, such as suspended solids, oil and grease, organic content in terms of BOD or COD, pH, specific metals and organic compounds, nitrogen and phosphorus, and specific organisms [120].

Anaerobic-aerobic system, as a promising technology, has been applied in industrial WW treatment for many years because of its remarkable benefits, such as great potential of resource recovery, high overall treatment efficiency, less disposal of

sludge, low energy consumption, and so on [121, 122]. In this system, an anaerobic process is first applied on WW followed by an aerobic process. In the anaerobic process, complex wastes are degraded into CH_4 in absence of oxygen [123]. Highly polluted wastewater is treated and the level of chemical oxygen demand (COD) is significantly reduced during the anaerobic stage, where COD is an important measurement of the organic polluted compound amount in WW. However, in practical applications, despite of the high efficiency of anaerobic treatment, organic matters may not be completely stabilized because of the high organic strength of WW. In order to overcome the issues, the aerobic process is followed with anaerobic process to treat noxious anaerobic effluent which often contains ammonium ion (NH_4^+) and hydrogen sulphide (HS^-) to achieve higher degree of efficiency [124]. In aerobic process, organic wastes are converted into biomass and CO_2 in the condition of dissolved oxygen.

This work is mainly to investigate the anaerobic process, in which Expanded Granular Sludge Bed (EGSB) reactors are adopted. The advantages of EGSB include high up-flow velocity and organic loading rate, satisfactory suitability for dilute WW, expanded active sludge bed, good sludge-WW mixing, and so on [125]. In the anaerobic process, in order to ensure successful start-up, stable operation and high treatment efficiency, a proper environment for anaerobic bacteria need to be maintained, such as pH, temperature, flow rate, etc. Furthermore, influences of WW properties also need to be clarified. However, under the existing condition in the plant, the effect of these process parameters on effluent properties is only based on experience. It is difficult to quantitatively judge whether the existing process parameters can guarantee a successful environment and achieve successful treatment

performance. Therefore, this study is intended to identify the important process parameters which significantly affect process performance, wherein process performance is defined as being able to produce a treated effluent which meets the discharge limits imposed on the plant. The process performance can be evaluated by effluent total organic carbon (TOC), since it provides specific information about the type and origin of organic loads in WW and has steadily gained in importance in WW analysis in recent years. In addition, a software system is also proposed to establish and describe the relationship between process parameters and process performance through mathematical modeling. Hence, in this study, process parameters are treated as the measured input variables, while effluent TOC is the analyte property. Some nomenclatures which are used in this study are defined in Table A.1.

Table A.1 Some nomenclature.

TOC	Total Organic Carbon
COD	Chemical Oxygen Demand
BOD	Biological Oxygen Demand
VFA	Volatile Fatty Acids
TSS	Total Suspended Solids
TKN	Total Kjeldahl Nitrogen
EGSB	Expanded Granular Sludge Bed

A.2 The anaerobic wastewater treatment system

Figure A.1 The process of anaerobic wastewater treatment.

Figure A.1 depicts the anaerobic process for WW treatment. Firstly, raw WW originating from a chemical industry is collected in the influent header and flows into the equalization tank T-01. If the quantity of WW is too large, some of WW can be pumped into diversion tank T-001 first for buffering. When WW in T-01 is treated, WW in T-001 is returned into T-01 for treatment. T-130 is a tank similar to T-01, which takes over WW from a main customer of a specific chemical industry. In T-01 and T-130, WW is diluted in order to be equalized and cooled. After that the pre-treated WW enters the conditioning tank T-100A/B through the valve of HE-100A/B. In conditioning tanks T-100A and T-100B, pH, temperature and nutrient conditions (phosphorous and ferric chlorid) are appropriately adjusted for the subsistence of anaerobic bacteria in T-200A and T-200B. pH can be controlled by

dosing either caustic soda (NaOH) or hydrochloric acid (HCl), while temperature is controlled by diluting with sea water. Before bringing anaerobic biomass into the reactor and after the system is filled with WW, the air in the system needs to be eliminated by injecting nitrogen gas (N₂). After the seeding of T-200A or T-200B with biomass, the system is set on recycle for 24 hours with only the reactor feed pumps in operation but no raw WW intake. During this period, the conditioned WW is pumped into EGSB reactors of T-200A or T-200B, in which the anaerobic bacteria conversion process take place. If for some reasons pH and/or temperature are not within the specified range, the treated WW is recycled back into T-100A or T-100B for adjusting, then back to T-200A or T-200B for bacteria treatment again. It is on the analogy of this circulation until the treated water achieves the presetting standard that VFA in the effluent is below 5meq/l. Finally, biogas from the conditioning tanks of T-100A and T-100B are delivered to F-201 to flare, and the effluent treated WW is pumped from EGSB anaerobic system to aerobic system for further treatment.

A.3 Methodology: dynamic regression algorithms

Since anaerobic WW treatment is a dynamic process, static calibration techniques like PLS may not obtain satisfactory prediction results. Therefore, two dynamic regression algorithms of canonical variate analysis based state space (CVA-SS) and autoRegressive with eXogenous (ARX) models are introduced.

A.3.1 Canonical variate analysis based state space

Canonical variate analysis state space (CVA-SS) modeling was put forward by Larimore in 1983 [126]. As a powerful tool for linear dynamic system, it has been applied to various different fields, such as on-line adaptive control of an unstable aircraft wing flutter [127], high temperature short time pasteurization systems [128], the voltage response of methanol fuel cells [129], and so on. The most general form of state space model can be written as [126]

$$z_{t+1} = Fz_t + Gx_t + o_t \quad (\text{A.1})$$

$$y_t = Hz_t + Tx_t + Bw_t + \rho_t \quad (\text{A.2})$$

where z is the system states; F , G , H , T are the state, input, output, and transmission matrices, respectively; o and ρ are the process errors. In state space model, the first important thing is to approximate system states z . The most popular statistical method to approximate the real system states z are canonical variate analysis (CVA) and partial least squares (PLS) [126, 130]. The idea of CVA is to approximate the states from the process past $\mathbf{p}$ and future $\mathbf{f}$ which are defined as

$$\mathbf{p}_t = [y_{t-1}^T, y_{t-2}^T, \dots, x_{t-1}^T, x_{t-2}^T, \dots]^T \quad (\text{A.3})$$

$$\mathbf{f}_t = [y_t^T, y_{t+1}^T, \dots]^T \quad (\text{A.4})$$

Then linear combinations Z and D can be introduced as canonical variates that

$$Z = V\mathbf{p}_t \quad (\text{A.5})$$

$$D = W\mathbf{f}_t \quad (\text{A.6})$$

where V and W are the transformation matrices which can be calculated by singular value decomposition (SVD) as in [126]. Then the states can be found by $z_t = V_k \mathbf{p}_t$ where V_k is the first k components of V obtained by minimizing the prediction

error of process future $\mathbf{f}$. PLS is similar to CVA by projecting variations in measured process variables onto lower dimension orthogonal latent variables, which are considered as a fair approximation of system states [131]. However, CVA outperforms PLS since it provides a more accurate representation of the system using fewer identified parameters [132, 133]. Once the states are determined, the matrices of F , G , H , T , S and R can be calculated by traditional linear regression methods [126].

A.3.2 AutoRegressive with eXogenous model

The AutoRegressive model with eXogenous (ARX) inputs is a widely used dynamic representation which can be combined with kind of regression techniques, such as PLS, ANN and GPR. In this study, the performances of ARX-PLS, ARX-ANN, and ARX-GPR are compared.

The model of ARX-PLS can be expressed as [131]

$$y_t = \sum_{k_y=1}^{K_y} a_{k_y} y_{t-r} + \sum_{k_x=1}^{K_x} b_{k_x} \mathbf{x}_{t-r} + \mathbf{e} \quad (\text{A.7})$$

where K_y and K_x are the number of the past output and input required to model the present output. This model also can be rewritten in an alternative form as

$$\mathbf{y} = \boldsymbol{\varphi} \mathbf{B} + \mathbf{e} \quad (\text{A.8})$$

where $\boldsymbol{\varphi}$ composes of $\varphi_t = (y_{t-1}, \dots, y_{t-K_y}, x_{t-1}, \dots, x_{t-K_x})$. Then PLS is used to calculate the regression coefficients $\mathbf{B}$.

The model of ARX-ANN is based on ANN which can be written as

$$y_t = f \left[\sum_{n=1}^N \omega_n g(v\varphi_t + \phi) + \varepsilon_n \right] \quad (\text{A.9})$$

The model of ARX-GPR is similar to GPR, and the covariance function is defined as follows:

$$C(\varphi_i, \varphi_k) = a_0 + a_1 \sum_{j=1}^J \varphi_{ij} \varphi_{kj} + v_0 \exp \left(- \sum_{j=1}^J w_j (\varphi_{ij} - \varphi_{kj})^2 \right) + \sigma^2 \delta_{ik} \quad (\text{A.10})$$

Parameter estimation algorithms of ARX-ANN and ARX-GPR are the same as ANN and GPR.

A.4 Data review and pre-treatment

The data used in this study come from SembCorp industry. In the anaerobic system, 182 tags are recorded. Each tag represents one process parameter, such as WW flow into T-001, tank level of T-01, pH in T-100A, biogas flow from T-100B, the effluent TOC of T-200A/B, and so on. The first step of this work is to select one parameter as the analyte property which can represent water quality and is expected to be predicted through mathematical modeling. After investigation, effluent TOC measured in laboratory is chosen for the following reasons. Firstly, TOC is recognized as a reliable analytical technique to measure water quality in the way of providing specific information of the amount of natural organic matters in WW. Secondly, it has emerged as a rapid and accurate alternative to the classical BOD and COD tests which require much longer analytical time. Since the effluent TOC of T-200A and T-200B include almost the same information about the anaerobic process, the effluent TOC of

T-200A is chosen as the analyte property. Except the effluent TOC of T-200A, there are also other 181 recorded tags. Some of them are related to the effluent TOC of T-200A in the mathematical prediction model while others are not. The second work is to determine the useful tags which contribute to the effluent TOC of T-200A. Some parameters are eliminated for several reasons. Firstly, they are not related to effluent TOC, and changes of them have no influence on effluent TOC, such as tank level, flow rate and discharge pump valve opening of T-001, effluent COD, BOD, TSS, VFA, alkalinity, sulphide, phenol, cyanide of T-200A/B, and so on. For this reason, 34 tags are removed. Secondly, 110 tags of WW properties of pH, temperature, flow, TOC, COD, TSS, etc. from several specific customers is ruled out since they are out of system. Thirdly, 7 tags of COD, BOD, PO_4 , NH_3 , sulfate as SO_4 , sulphide, cyanide of T-01 are gotten rid of, since their sampling frequency is too low that only several data points are obtained every year. Therefore, most of the information included in these parameters is lacked. In addition, some measuring sensors of them become unreliable due to equipments aging, leading to the removal of 8 tags of recycled treated effluent to T-01 valve opening, recycled treated effluent to T-01 flow, seawater to HE-100A/B valve opening, biogas flow from T-100A, biogas flow from T-100B, influent to T-100B flow, T-100A to T-200A recirculation flow, and T-100B to T-200B recirculation flow. Through deliberate survey, 22 process parameters are retained as model inputs finally, which are listed in Table A.2. In this table, “lab” means the parameter is measured in laboratory but not on-line.

After selecting the analyte property of effluent TOC in T-200A and useful parameters contributing to the analyte property, the next work is to find a solution to match the 22 useful process parameters with effluent TOC. As mentioned in the

introduction, after the reactor of T-200A is seeded with biomass, the system is fitted in circulation for 24 hours with only reactor feed in operation but no raw WW intake. Therefore, the daily measured effluent TOC depends on the process parameters operated in the past 24 hours. Subsequently, it is considered using data in the past 24 hour of the 22 parameters to match with the daily measured effluent TOC in T-200A. However, since the sampling frequencies of effluent TOC and 22 useful parameters are different, it is difficult to directly relate these process parameters to the effluent TOC in T-200A. The effluent TOC in T-200A is measured around once per day, but the measurement frequencies of other process parameter are uneven. For the 10 parameters of T-01 tank level, T-01 outlet temperature, seawater to HE-100A/B flow, HE-100A/B WW outlet temperature, influent to T-100A/B flow, influent to T-100A/B v/v opening, biogas pressure, T-100A to T-200A recirculation v/v opening, T-100B to T-200B recirculation v/v opening, and EGSB effluent temperature, 20-60 data points are sampled per hour. For the 4 parameters of T-01 pH, T-100A pH, T-100B pH, and EGSB effluent pH, 0-20 samples are obtained per hour, where 0 means no sampling in some hours. For the 3 parameters of T-01 TOC, influent to T-100B v/v opening, and T-01 WW TOC (lab), the sampling frequency is 0-20 samples per day. For the 5 parameters of T-01 WW pH (lab), T-01 WW TSS (lab), T-01 WW VFA (lab), T-01 WW alkalinity (lab), and T-01 WW phenol (lab), only 0-10 samples are collected every month. In this condition, a standard frequency of one sample per hour (24 samples per day) is adopted to unify the 22 process parameters. According to this standard, samples which are not recorded on the hour are trimmed off, and only samples on the hour are retained. For the parameters which lack samples on the hour, linear interpolation is adopted to make up them. As a result, the dimension of the

Table A.2 The selected 22 tags and their corresponding process parameters which contribute to the mathematical prediction model.

Tag No.	Parameters	Tag No.	Parameters
1	T-01 tank level	12	T-100B pH
2	T-01 outlet temperature	13	T-100A to T-200A recirculation v/v opening
3	T-01 pH	14	T-100B to T-200B recirculation v/v opening
4	T-01 TOC	15	EGSB effluent temperature
5	Seawater to HE-100A/B flow	16	EGSB effluent pH
6	HE-100A/B WW outlet temperature	17	T-01 WW pH (lab) ^a
7	Influent to T-100A/B flow	18	T-01 WW TOC (lab)
8	Influent to T-100A/B v/v opening	19	T-01 WW TSS (lab)
9	Biogas pressure	20	T-01 WW VFA (lab)
10	Influent to T-100B v/v opening	21	T-01 WW alkalinity (lab)
11	T-100A pH	22	T-01 WW phenol (lab)

input variables is $22 \times 24 = 528$. In this way, the measured input variables x_t and output analyte property of effluent TOC y_t are matched.

Table A.3 The number of samples of the low frequency parameters from 2004 to 2009.

Parameters	No. of samples
T-01 WW COD (lab)	62
T-01 WW BOD (lab)	54
T-01 WW PO ₄ as P (lab)	1
T-01 WW NH ₃ as N (lab)	4
T-01 WW sulfate as SO ₄ ²⁻ (lab)	24
T-01 WW sulphide (lab)	9
T-01 WW cyanide (lab)	45

As mentioned before, seven tags are excluded since the sampling frequency is too low and only few data points are available. Table A.3 displays their number of samples from 2004 to 2009. These parameters are the influent properties of feeding in T-01 which are measured in laboratory, and could be directly contribute to effluent TOC. Pearson's correlation coefficient R of these influent properties and the effluent TOC in T-200A is an important index to characterize the relationship between two random variables (the influent feeding and effluent TOC) [134], which is defined as

$$R_p = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (\text{A.11})$$

where μ_X and μ_Y are the mean of influent properties X and effluent TOC Y ; σ_X and σ_Y are the standard deviations of X and Y , respectively. The value of correlation coefficient is between -1 and 1. The value of 1 indicates X and Y have a perfectly positive linear relationship that Y increases as X increases. The value of -1 indicates X and Y have a perfectly negative linear relationship that Y decreases as X increases. The value of 0 implies that there is no linear relationship between X and Y . Any parameter with the property that the absolute value of Pearson's R is greater than $1/N$ should be selected as the useful parameter of the prediction model, where N is the number of possible selected model parameters of 22 [135]. Hence, that the absolute value of R is > 0.05 implies the influent property has linear relationship with effluent TOC. Since Pearson's R can only indicate linear relationship, Spearman's R_s , an index of non-linear relationship, is also studied. The definition of Spearman's R_s is written as [136]

$$R_s = 1 - \frac{6 \sum_1^I (X_i - Y_i)}{I^3 - I} \quad (\text{A.12})$$

If Pearson's R is very small, but Spearman's R_s is large, the influent parameter also needs to be retained in the model, since there is a strong non-linear relationship between X and Y [134]. In order to see whether the results are significant, p -value is introduced, which is defined as the probability of obtaining a result when the null hypothesis is true [137]. If $p < 0.05$, then it can be concluded that the result is

significant, which means the process parameter is significantly related to effluent TOC.

The Pearson's and Spearman's correlation coefficients and the corresponding p -values of influent properties of T-01 and change of effluent TOC are calculated and are displayed in Table A.4. It can be observed that Pearson's R is greater than 0.05 and

Table A.4 Correlation coefficients of influent properties of T-01 and change of TOC.

Parameters	R	p -value of	R_s	p -value of
		R		R_s
T-01 WW pH (lab)	-0.14	0.00	-0.04	0.18
T-01 WW TSS (lab)	0.07	0.02	-0.11	0.00
T-01 WW VFA (lab)	-0.03	0.28	0.02	0.47
T-01 WW alkalinity (lab)	0.15	0.00	-0.23	0.00
T-01 WW phenol (lab)	-0.40	0.00	-0.47	0.00
T-01 WW COD (lab)	-0.32	0.01	0.002	0.99
T-01 WW BOD (lab)	-0.20	0.15	-0.17	0.23
T-01 WW NH ₃ as N (lab)	0.91	0.09	0.20	0.92
T-01 WW sulfate as SO ₄ ²⁻ (lab)	-0.21	0.33	-0.05	0.82
T-01 WW sulphide (lab)	-0.35	0.35	-0.17	0.68
T-01 WW cyanide (lab)	-0.15	0.34	-0.08	0.60

p -values are less than 0.05 for T-01 WW pH (lab), T-01 WW TSS (lab), T-01 WW alkalinity (lab), T-01 WW phenol (lab), T-01 WW COD (lab), indicating they are significantly linear to effluent TOC. Although p -values of T-01 WW BOD (lab), T-01 WW NH₃ as N (lab), T-01 WW sulfate as SO₄²⁻ (lab), T-01 WW sulphide (lab), T-01 WW cyanide (lab) is greater than 0.10, there are still more than 70% chance ($1-p$) that they are linear to effluent TOC. From Spearman's R_s and p -value it can be seen only T-01 WW TSS (lab), T-01 WW alkalinity (lab), and T-01 WW phenol (lab) have significant non-linear relationship with effluent TOC ($R_s > 0.05$ and $p < 0.05$), while others do not have non-linear contribution. Therefore, all these influent feedings except T-01 WW TSS (lab) may need to be included in the calibration model, since they have linear or non-linear contribution. However, only the first five parameters are present in the model, because too few data points (less than 100 during 2004-2009, Table A.3) are recorded for others, and cannot be used for modeling. Therefore, it is suggested that the influent properties in T-01 are measured in the future in order to obtain more accurate prediction model.

After matching of the input and output variables, the next step is to investigate different mathematical algorithms to describe the regression relation of the measured variables and the effluent TOC in T-200A. Five regression algorithms of PLS, ARX-PLS, ARX-ANN, ARX-GPR and CVA-SS are investigated, and their prediction performance are compared. The purpose of calibration model is to predict the future effluent TOC based on the history data in the past years. Up to now, data from 2004 to

2009 are available. Therefore, data from quarter 1 in 2004 to quarter 2 in 2009 are used as the training data, while data from quarter 3 in 2009 to quarter 4 in 2009 are used as the testing data to validate the model accuracy. Like in Chapter 6, relative root mean square error of prediction (RRMSEP) is used to evaluate the performance of these investigated calibration methods, since it can be more obvious to see how much the predicted data deviate from measured data from RRMSEP than RMSEP. In order for technicians in the plant to better understand the prediction performance, RRMSEP is used.

A.5 Prediction results of dynamic techniques

Figure A.2 The measured and predicted effluent TOC using PLS.

First, simple PLS is used to predict the effluent TOC. Figure A.2 displays the measured versus predicted effluent TOC by using PLS, in which the red solid line and

blue dash line indicate the measured and predicted data, respectively. Table A.5 gives RRMSEPs of PLS. It can be seen the prediction performance of PLS is not satisfactory, since the predicted data do not match the measured data very well in terms of its very high RRMSEP of 122.4%. This is because the anaerobic process is dynamic and the effluent TOC in T-200A is autocorrelated, which can be seen from Figure A.3 that the effluent TOC in the past few days have significant influence on T-200A TOC in current days. In Figure A.3, y-coordinate is the q^{th} autocorrelation coefficient defined as [138]:

$$r_q = c_q / c_0 \quad (\text{A.13})$$

where $c_q = \frac{1}{l} \sum_{t=1}^{l-q} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$, and $c_0 = \frac{1}{l} \sum_{t=1}^l (Y_t - \bar{Y})^2$. The x-coordinate indicate the time lags q .

Figure A.3 The autocorrelation of effluent TOC.

In order to solve this autocorrelation problem, dynamic autoregressive models, including ARX-PLS, ARX-ANN, ARX-GPR, and CVA-SS, are investigated to further improve the prediction accuracy.

In ARX-PLS, ARX-ANN, ARX-GPR, the effluent TOC in the past few days $y_{t-1,t-2,\dots,t-K_y}$ and the measured input variables x_t are juxtaposed, so the number of past effluent TOC K_y need to be determined. Similar to ANN and GPR in previous chapters, PLS is first applied to reduce the dimension of measured input variables prior to the development of ARX-ANN and ARX-GPR models. Table A.6 shows the influence of different past days on the prediction performance of ARX-PLS. It can be observed that as long as the effluent TOC in the past days is included in the model, RRMSEP will be reduced significantly compared with PLS, and the number of K_y does not have great influence on the prediction performance, since RRMSEP does not change obviously with K_y . In order to guarantee that enough information of effluent TOC in the past days is included in the model, $K_y = 10$ is selected, which means the effluent TOC in the past 10 days will be adopted in the calibration model. ARX-ANN and ARX-GPR also have a similar behavior, details of which are not stated here.

Figure A.4 shows the prediction performance of ARX-PLS, ARX-ANN, ARX-GPR, and CVA-SS. It can be seen that all the four dynamic models give satisfactory results. Their RRMSEPs are displayed in Table A.5. ARX-PLS gives the lowest RRMSEP and best prediction performance, while CVA-SS has the worst performance. Although ARX-ANN and ARX-GPR give the similar prediction results

to ARX-PLS, their computation time is longer. Therefore, ARX-PLS is considered as the optimal mathematical model to deal with this anaerobic system.

Figure A.4 The measured and predicted data of (a) ARX-PLS, (b) ARX-ANN, (c) ARX-GPR, and (d) CVA-SS.

Table A.5 The relative root mean square error of prediction (RRMSEP) of different techniques.

Techniques	PLS	ARX-PLS	CVA-SS	ARX-ANN	ARX-GPR
RRMSEP	122.4%	20.0%	27.1%	20.4%	20.3%

Table A.6 Effluence of the past n days on RRMSEP for ARX-PLS, ARX-ANN, and ARX-GPR.

n	2	4	6	8	10	12	14	16	18	20
ARX-PLS	19.8%	20.0%	19.7%	19.8%	19.8%	19.7%	19.8%	19.9%	20.0%	20.2%
ARX-ANN	20.3%	20.6%	19.8%	19.6%	20.1%	20.3%	19.8%	20.3%	19.8%	20.1%
ARX-GPR	19.9%	20.1%	20.2%	19.9%	20.1%	19.7%	20.2%	19.9%	20.1%	20.0%

A.6 Summary

The development of mathematical model between the process input and effluent output is very useful for WW treatment system, since the model can give guideline to discern whether the system work well or not. Four dynamic models of ARX-PLS, ARX-ANN, ARX-GPR, and CVA-SS are investigated and compared. According to their prediction performance, ARX-PLS is selected as the optimal software sensor to predict effluent TOC. However, it should be noted that there is also certain limitation for these autoregression techniques that the analyte property of effluent TOC in the past need to be measured timely to make prediction.

References

- [1] M. Kano, Y. Nakagawa, Process Systems Engineering Conference/16th European Symposium on Computer Aided Process Engineering (ESCAPE-16), Pergamon-Elsevier Science Ltd, Garmisch Partenkirchen, Germany, 2008, p. 12.
- [2] I. Miletic, S. Quinn, M. Dudzic, V. Vaculik, M. Champagne, J. Process Control 14 (2004) 821.
- [3] H. Martens, T. Naes, Trac-Trends Anal. Chem. 3 (1984) 204.
- [4] W.W. Blaser, R.A. Bredeweg, R.S. Harner, M.A. Lapack, A. Leugers, D.P. Martin, R.J. Pell, J. Workman, L.G. Wright, Anal. Chem. 67 (1995) R47.
- [5] J.W. Hall, B. McNeil, M.J. Rollins, I. Draper, B.G. Thompson, G. Macaloney, Appl. Spectrosc. 50 (1996) 102.
- [6] T. Togkalidou, M. Fujiwara, S. Patel, R.D. Braatz, J. Cryst. Growth 231 (2001) 534.
- [7] R. Bro, International Meeting on Analytical Horizons, Elsevier Science Bv, Ellecom, Netherlands, 2003, p. 185.
- [8] R.J. Hendershot, R. Vijay, B.J. Feist, C.M. Snively, J. Lauterbach, Meas. Sci. Technol. 16 (2005) 302.
- [9] R. Steponavicius, S.N. Thennadil, Anal. Chem. 81 (2009) 7713.
- [10] T. Chen, J. Morris, E. Martin, Chemom. Intell. Lab. Syst. 87 (2007) 59.
- [11] I.A. Cowe, J.W. McNicol, Appl. Spectrosc. 39 (1985) 257.
- [12] S. Wold, H. Martens, H. Wold, Lecture Notes in Mathematics 973 (1983) 286.

- [13] P. Geladi, B.R. Kowalski, *Anal. Chim. Acta* 185 (1986) 1.
- [14] P. Geladi, D. Macdougall, H. Martens, *Appl. Spectrosc.* 39 (1985) 491.
- [15] K. DeBraekeleer, F.C. Sanchez, P.A. Hailey, D.C.A. Sharp, A.J. Pettman, D.L. Massart, *J. Pharm. Biomed. Anal.* 17 (1998) 141.
- [16] E.D. Yalvac, M.B. Seasholtz, M.A. Beach, S.R. Crouch, *Appl. Spectrosc.* 51 (1997) 1565.
- [17] C.D. Brown, L. Vega-Montoto, P.D. Wentzell, *Appl. Spectrosc.* 54 (2000) 1055.
- [18] D.Z. Zhu, B.P. Ji, C.Y. Meng, B.L. Shi, Z.H. Tu, Z.S. Qing, *Chemom. Intell. Lab. Syst.* 90 (2008) 108.
- [19] S.N. Thennadil, E.B. Martin, *J. Chemometr.* 19 (2005) 77.
- [20] E. Gratton, *Biopolymers* 10 (1971) 2629.
- [21] H. Martens, J.P. Nielsen, S.B. Engelsen, *Anal. Chem.* 75 (2003) 394.
- [22] S.N. Thennadil, H. Martens, A. Kohler, *Appl. Spectrosc.* 60 (2006) 315.
- [23] A. Savitzky, M.J.E. Golay, *Anal. Chem.* 36 (1964) 1627.
- [24] R.J. Barnes, M.S. Dhanoa, S.J. Lister, *Appl. Spectrosc.* 43 (1989) 772.
- [25] D.K. Pedersen, H. Martens, J.P. Nielsen, S.B. Engelsen, *Appl. Spectrosc.* 56 (2002) 1206.
- [26] Z.P. Chen, J. Morris, E. Martin, *Anal. Chem.* 78 (2006) 7674.
- [27] J.R. Long, V.G. Gregoriou, P.J. Gemperline, *Anal. Chem.* 62 (1990) 1791.
- [28] Z.Y. Wang, J.N. Hwang, B.R. Kowalski, *Anal. Chem.* 67 (1995) 1497.
- [29] H.H. Thodberg, *IEEE Trans. Neural Netw.* 7 (1996) 56.

- [30] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Square Support Vector Machines, World Scientific, Singapore, 2002.
- [31] U. Thissen, B. Ustun, W.J. Melssen, L.M.C. Buydens, Anal. Chem. 76 (2004) 3099.
- [32] F. Wulfert, W.T. Kok, O.E. de Noord, A.K. Smilde, Chemom. Intell. Lab. Syst. 51 (2000) 189.
- [33] H. Swierenga, F. Wulfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, Anal. Chim. Acta 411 (2000) 121.
- [34] L. Breiman, Mach. Learn. 24 (1996) 123.
- [35] L. Breiman, Mach. Learn. 24 (1996) 49.
- [36] P. Buhlmann, B. Yu, Ann. Stat. 30 (2002) 927.
- [37] V. Centner, D.L. Massart, O.E. deNoord, S. deJong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
- [38] H. Ojelund, H. Madsen, P. Thyregod, J. Chemometr. 15 (2001) 497.
- [39] G. Yi, J.Q. Shi, T. Choi, Biometrics (2011).
- [40] M. Dyrby, S.B. Engelsen, L. Norgaard, M. Bruhn, L. Lundsberg-Nielsen, Appl. Spectrosc. 56 (2002) 579.
- [41] J. Burgbacher, J. Wiss, Org. Process Res. Dev. 12 (2008) 235.
- [42] S.W. Wong, C. Georgakis, G.D. Botsaris, K. Saranteas, R. Bakale, Ind. Eng. Chem. Res. 47 (2008) 5576.
- [43] H. Alatalo, J. Kohonen, H. Qu, H. Hatakka, S.P. Reinikainen, M.

- Louhi-Kultanen, J. Kallas, 10th Scandinavian Symposium on Chemometrics, Lappeenranta, Finland, 2007, p. 644.
- [44] O. Elizalde, J.R. Leiza, J.M. Asua, *Macromol. Symp.* 206 (2004) 135.
- [45] D.A.A. Hutsebaut, C. Vercaemst, C. Hugelier, F. Verpoort, *Appl. Spectrosc.* 56 (2002) 1314.
- [46] S. Wold, A. Ruhe, H. Wold, W.J. Dunn, *Siam J. Sci. Stat. Comput.* 5 (1984) 735.
- [47] J. Cornel, C. Lindenberg, M. Mazzotti, *Ind. Eng. Chem. Res.* 47 (2008) 4870.
- [48] A.K. Jain, J.C. Mao, K.M. Mohiuddin, *Computer* 29 (1996) 31.
- [49] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Walton Street, Oxford, 1995.
- [50] P.J. Gemperline, J.R. Long, V.G. Gregoriou, *Anal. Chem.* 63 (1991) 2313.
- [51] Y. Li, J. Wang, Z.R. Chen, X.T. Zhou, *Anal. Lett.* 34 (2001) 2203.
- [52] D.J.C. Mackay, *Neural Comput.* 4 (1992) 448.
- [53] D.J.C. Mackay, *Neural Comput.* 4 (1992) 415.
- [54] A. O'Hagan, *J. R. Stat. Soc. Ser. B-Methodol.* 40 (1978) 1.
- [55] C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT press, Cambridge, MA, 2006.
- [56] C. Cortes, V. Vapnik, *Mach. Learn.* 20 (1995) 273.
- [57] A. Borin, M.F. Ferrao, C. Mello, D.A. Maretto, R.J. Poppi, *Anal. Chim. Acta* 579 (2006) 25.
- [58] D.K. Agrafiotis, W. Cedeno, V.S. Lobanov, *J. Chem. Inf. Comput. Sci.* 42 (2002)

903.

- [59] B. Efron, Ann. Stat. 7 (1979) 1.
- [60] B. Efron, Biometrika 68 (1981) 589.
- [61] S. Dudoit, J. Fridlyand, NIPS Workshop on Machine Learning Techniques for Bioinformatics, Oxford Univ Press, Whistler, Canada, 2001, p. 1090.
- [62] E. Bauer, R. Kohavi, AAAI 1996 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, Kluwer Academic Publ, Oregon, Portland, 1996, p. 105.
- [63] T.G. Dietterich, Mach. Learn. 40 (2000) 139.
- [64] R. Gencay, M. Qi, IEEE Trans. Neural Netw. 12 (2001) 726.
- [65] P.R. Lajbcygier, J.T. Connor, Int. J. Neural Syst. 8 (1997) 457.
- [66] M. Sugimoto, S. Kikuchi, M. Arita, T. Soga, T. Nishioka, M. Tomita, Anal. Chem. 77 (2005) 78.
- [67] M. Taniguchi, V. Tresp, Neural Comput. 9 (1997) 1163.
- [68] D.H. Wolpert, Neural Netw. 5 (1992) 241.
- [69] J. Zhang, Neural Netw. 12 (1999) 927.
- [70] R.K.H. Galvao, M.C.U. Araujo, M.D. Martins, G.E. Jose, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Chemom. Intell. Lab. Syst. 81 (2006) 60.
- [71] T. Chen, J.H. Ren, Neurocomputing 72 (2009) 1605.
- [72] M. Skurichina, R.P.W. Duin, Pattern Recognit. 31 (1998) 909.
- [73] L. Breiman, Mach. Learn. 45 (2001) 261.

- [74] C. Croux, K. Joossens, A. Lemmens, *Comput. Stat. Data Anal.* 52 (2007) 362.
- [75] S. Borra, A. Di Ciaccio, *Meeting on Nonlinear Methods and Data Mining_(NMDM2000)*, Elsevier Science Bv, Rome, Italy, 2000, p. 407.
- [76] W.S. Cai, Y.K. Li, X.G. Shao, *Chemom. Intell. Lab. Syst.* 90 (2008) 188.
- [77] F. Liu, Y.H. Jiang, Y. He, *Anal. Chim. Acta*, 635 (2009) 45.
- [78] T. Chen, E. Martin, *International Chemometrics Research Meeting (ICRM 2006)*, John Wiley & Sons Ltd, Veldhoven, Netherlands, 2006, p. 198.
- [79] R. Tibshirani, *J. R. Stat. Soc. Ser. B-Methodol.* 58 (1996) 267.
- [80] L. Hadjiiski, P. Geladi, P. Hopke, *Chemom. Intell. Lab. Syst.* 49 (1999) 91.
- [81] V. Centner, J. Verdu-Andres, B. Walczak, D. Jouan-Rimbaud, F. Despagne, L. Pasti, R. Poppi, D.L. Massart, O.E. de Noord, *Appl. Spectrosc.* 54 (2000) 608.
- [82] U. Thissen, M. Pepers, B. Ustun, W.J. Melssen, L.M.C. Buydens, *Chemom. Intell. Lab. Syst.* 73 (2004) 169.
- [83] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, *Chemom. Intell. Lab. Syst.* 88 (2007) 183.
- [84] C.H. Coulden, *Methods of Statistical Analysis*, second ed., Wiley, New York, 1956.
- [85] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, *Technometrics* 47 (2005) 64.
- [86] M. Daszykowski, K. Kaczmarek, Y.V. Heyden, B. Walczak, *Chemom. Intell. Lab. Syst.* 85 (2007) 203.
- [87] M. Debruyne, M. Hubert, J. Van Horebeek, *Comput. Stat. Data Anal.* 54 (2010)

3007.

- [88] H. Martens, T. Naes, *Multivariate Calibration*, John Wiley and Sons Ltd., New York, 1992.
- [89] Z.P. Chen, J. Morris, *Analyst* 133 (2008) 914.
- [90] C. Tan, X. Qin, M.L. Li, *Anal. Lett.* 42 (2009) 1693.
- [91] R.K.H. Galvao, M.C.U. Araujo, G.E. Jose, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, *Talanta* 67 (2005) 736.
- [92] R.W. Kennard, L.A. Stone, *Technometrics* 11 (1969) 137.
- [93] B.L. Lei, J.Z. Li, H.X. Liu, X.J. Yao, *QSAR Comb. Sci.* 27 (2008) 850.
- [94] A. Jorio, G. Dresselhaus, M.S. Dresselhaus, *Carbon Nanotubes, Advanced Topics in the Synthesis, Structure, Properties and Applications*, Springer, Berlin, 2008.
- [95] M.C. Hersam, *Nat. Nanotechnol.* 3 (2008) 387.
- [96] S. Banerjee, T. Hemraj-Benny, S.S. Wong, *J. Nanosci. Nanotechnol.* 5 (2005) 841.
- [97] J. Liu, M.C. Hersam, *MRS Bull.* 35 (2010) 315.
- [98] Y. Matsuo, K. Tahara, E. Nakamura, *Org. Lett.* 5 (2003) 3181.
- [99] J.A. Fagan, J.R. Simpson, B.J. Bauer, S.H.D. Lacerda, M.L. Becker, J. Chun, K.B. Migler, A.R.H. Walker, E.K. Hobbie, *J. Am. Chem. Soc.* 129 (2007) 10607.
- [100] A. Rajan, M.S. Strano, D.A. Heller, T. Hertel, K. Schulten, *J. Phys. Chem. B* 114 (2010) 4938.

- [101]J.P. Casey, S.M. Bachilo, C.H. Moran, R.B. Weisman, ACS Nano 2 (2008) 1738.
- [102]M. Zheng, E.D. Semke, J. Am. Chem. Soc. 129 (2007) 6084.
- [103]M.L. Becker, J.A. Fagan, N.D. Gallant, B.J. Bauer, V. Bajpai, E.K. Hobbie, S.H. Lacerda, K.B. Migler, J.P. Jakupciak, Adv. Mater. 19 (2007) 939.
- [104]H.C.v.d. Hulst (Ed.), Light Scattering by Small Particles, Dover Publications, Inc New York, 1981.
- [105]M.A. van Dijk, A.L. Tchegotareva, M. Orrit, M. Lippitz, S. Berciaud, D. Lasne, L. Cognet, B. Lounis, PCCP Phys. Chem. Chem. Phys. 8 (2006) 3486.
- [106]P.K. Jain, K.S. Lee, I.H. El-Sayed, M.A. El-Sayed, J. Phys. Chem. B 110 (2006) 7238.
- [107]S. Sarkar, A.A. Gurjarpadhye, C.G. Rylander, M.N. Rylander, J. Biomed. Opt. 16 (2011).
- [108]J.A. Fagan, M.L. Becker, J.H. Chun, P.T. Nie, B.J. Bauer, J.R. Simpson, A. Hight-Walker, E.K. Hobbie, Langmuir 24 (2008) 13880.
- [109]K. Wang, G.Y. Chi, R. Lau, T. Chen, Anal. Lett. 44 (2011) 824.
- [110] J.A. Fagan, M.L. Becker, J. Chun, E.K. Hobbie, Adv. Mater. 20 (2008) 1609.
- [111]E.K. Hobbie, J.A. Fagan, J. Obrzut, S.D. Hudson, ACS Appl. Mater. Interfaces 1 (2009) 1561.
- [112]A. Matsuura, L.S. Fan, Aiche J. 30 (1984) 894.
- [113]M.I. Hossain, T. Chen, Y.H. Yang, R. Lau, Ind. Eng. Chem. Res. 48 (2009) 10136.

- [114]J.P.M. Syvitski, Principles, Methods and Application of Particle Size Analysis, Cambridge University Press, New York, 2007.
- [115]Z.H. Zhou, M. Li, IEEE Trans. Knowl. Data Eng. 19 (2007) 1479.
- [116]C. Zhu, Technical Report 1530, Dept. of Computer Sciences, Univ. of Wisconsin, Madison, http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf, 2006.
- [117]M. Belkin, P. Niyogi, Mach. Learn. 56 (2004) 209.
- [118]C.M. Bishop, J. Lasserre, Bayesian Statistics 8, Oxford University Press, 2007.
- [119]A. Blum, T. Mitchell, Proc. 11th Ann. Conf. Computational Learning Theory, ACM, New York, USA, 1998, p. 92.
- [120]W.J. Ng, Industrial Wastewater Treatment, World Scientific, Singapore, 2006.
- [121]B. Frostell, Starke 35 (1983) 185.
- [122]Y.J. Chan, M.F. Chong, C.L. Law, D.G. Hassell, Chem. Eng. J. 155 (2009) 1.
- [123] F.Y. Cakir, M.K. Stenstrom, Water Res. 39 (2005) 4197.
- [124]J.J. Heijnen, A. Mulder, R. Weltevrede, J. Hols, H. Vanleeuwen, Water Sci. Technol. 23 (1991) 1427.
- [125]L. Seghezzo, G. Zeeman, J.B. van Lier, H.V.M. Hamelers, G. Lettinga, Bioresour. Technol. 65 (1998) 175.
- [126]W.E. Larimore, American Control Conference, San Francisco, CA, USA, 1983, p. 445.
- [127]R.P. Peloubet, Jr., R.L. Haller, R.M. Bolding, Decision and Control, 1990, Proceedings of the 29th IEEE Conference on, Honolulu, HI , USA, 1990, p. 643.

- [128]A. Negiz, P. Ramanauskas, A. Cinar, J.E. Schlessner, D.J. Armstrong, Food Control, 9 (1998) 29.
- [129]A. Simoglou, P. Argyropoulos, E.B. Martin, K. Scott, A.J. Morris, W.M. Taama, Chem. Eng. Sci. 56 (2001) 6761.
- [130]W.E. Larimore, C.S.S. Ieee, Decision and Control, 1990, Proceedings of the 29th IEEE Conference on, Honolulu, HI, USA, 1990, p. 596.
- [131]A. Simoglou, E.B. Martin, A.J. Morris, American Control Conference, San Diego, CA , USA, 1999, p. 823.
- [132]A. Negiz, A. Cinar, Chemom. Intell. Lab. Syst. 38 (1997) 209.
- [133]A. Simoglou, Ph.D. thesis. University of Newcastle, Newcastle upon Tyne, 1999.
- [134]J.D. Jobson, Applied Multivariate Data Analysis, Springer Verlag, New York, 1991.
- [135]W.J. Ng, S.L. Ong, F. Hossain, Environ. Modell. Softw. 15 (2000) 199.
- [136]D.J. Best, D.E. Roberts, J. Appl. Stat. (1975) 377.
- [137]J.D. Gibbons, S. Chakraborti, Nonparametric Statistical Inference, fifth ed., CRC press, New York, 2010.
- [138]G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, third ed., Englewood Cliffs, N.J., Prentice Hall,1994.