Development of problem-oriented strategies for biomedical and raman spectroscopic signal analysis

Bai, Yanru

2019

Bai, Y. (2019). Development of problem-oriented strategies for biomedical and raman spectroscopic signal analysis. Singapore: Nanyang Technological University

<https://hdl.handle.net/10356/137051>

<https://doi.org/10.32657/10356/137051>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 19:42:40 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**DEVELOPMENT OF PROBLEM-ORIENTED
STRATEGIES FOR BIOMEDICAL AND RAMAN
SPECTROSCOPIC SIGNAL ANALYSIS**

BAI YANRU

School of Chemical and Biomedical Engineering

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

2019-08-20

.....

Date

Bai Yanru

.....

Bai Yanru

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

2019-08-20

.....

Date

.....

Assoc. Prof. Liu Quan

Authorship Attribution Statement

This dissertation contains material from two papers published in the following peer-reviewed journal where I was the first author.

Chapter 3 is published as **Y.R. Bai**, Q. Liu*, “Denoising Raman spectra by Wiener estimation with a numerical calibration dataset,” *Biomedical Optics Express* 11.1 (2020): 200-214. DOI: 10.1364/boe.11.000200.

The contributions of the co-authors are as follows:

- Prof. Liu Quan provided the initial project direction and edited the manuscript drafts.
- I co-designed the study with Prof. Liu Quan and prepared the manuscript drafts. All the data processing work is done by me at the School of Chemical and Biomedical Engineering, Nanyang Technological University.

Chapter 5 is published as **Y.R. Bai**, G. Huang, Y.H. Tu, A. Tan, Y.S. Hung, and Z.G. Zhang*, “Normalization of Pain-evoked Neural Responses Using Spontaneous EEG Improves the Performance of EEG-based Cross-individual Pain Prediction,” *Frontiers in Computational Neuroscience* 10 (2016): 31. DOI: 10.3389/fncom.2016.00031.

The contributions of the co-authors are as follows:

- Prof. Zhang Zhiguo provided the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by Prof.

Y.S. Hung.

- I co-designed the study with Prof. Zhang Zhiguo. All the data processing work is done by me at the School of Chemical and Biomedical Engineering, Nanyang Technological University.
- The EEG data was collected by Dr. Huang Gan at the University of Hong Kong.
- Dr. Tu Yiheng and Dr. Tan Ao provided guidance in the experiment design and data processing.

2019-08-20

Date

Bai Yanru

Bai Yanru

Acknowledgements

This dissertation would not have been completed without the help of many people. I would like to take this opportunity to thank all who have contributed to this dissertation for their numerous sacrifices, encouragement and immense support.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Liu Quan, who provides me with a great platform and environment so that I could perform my study and research comfortably throughout my PhD period. His insights and deep understanding towards research projects and his patience and kindness to students really set a good model for me to learn from. He has been instructive and supportive on every aspect of my PhD projects. I really benefit a lot from the inspiring discussions with him on every meeting. His meticulous reading and comments greatly help to enhance the quality and readability of the drafts. Thank you for your constant motivation and support which help me a great deal to achieve this feat.

Special appreciation also goes to my co-supervisor, Professor Zhang Zhiguo, who encouraged me to pursue my PhD degree and provided me with the opportunity to study at Nanyang Technological University. I am very honored to work with him for about three years in Hong Kong and Singapore. His insightful opinions and the invaluable guidance have always led me to the correct direction when I encountered problems and difficulties in my research projects.

I would like to give special acknowledgement to my TAC members, QE examiners,

the panel of my PhD Oral Defense for their insightful guidance and invaluable suggestions on my PhD projects. I am also thankful to all other faculty and staff in School of Chemical and Biomedical Engineering for the academic support they provided to me during my PhD period.

I am truly grateful to all the members and colleagues in my research group for their constant support, co-operation and the joyful experiences we shared in daily life. I would also like to thank my friends in Singapore for all their encouragement and motivations. Together, we have witnessed each other's growth and progress. The fulfilled and joyful time we spent together will be the precious memories of my life.

Most of all, my heartfelt thanks to my parents and my dearest sister. I am deeply indebted to them who always had their beliefs in me and gave their support for all the choices and decisions that I have made. It has been a long journey, but I finally made it! Thank you for the endless love and encouragement that always make me full of courage and strength to overcome any difficulties throughout my road to graduation.

Thanks for the wonderful life here.

Table of Contents

Acknowledgements.....	I
Table of Contents	III
Glossary of Acronyms.....	VI
List of Figures	IX
List of Tables.....	XVI
Summary	1
Chapter 1 Overview	6
1.1 Background and Motivations	6
1.2 Objectives of This Work.....	9
1.3 Organization of The Dissertation.....	10
Chapter 2 Introduction.....	13
2.1 General Procedure of Signal Processing.....	13
2.2 Multivariate Analysis	13
2.2.1 Classification Techniques.....	15
2.2.2 Regression/Prediction Techniques	18
2.3 Raman Spectroscopy and Spectral Data Analysis	21
2.3.1 Fundamentals of Raman Spectroscopy.....	21
2.3.2 Pre-processing Techniques in Raman spectroscopy	24
2.3.3 Raman Spectroscopy as a Quantitative Tool	32
2.4 Pain-related EEG Responses and Pain Prediction.....	33
2.4.1 Laser-Evoked Potentials.....	34
2.4.2 Challenge in Cross-Individual Classification/Prediction	36
Chapter 3 A Denoising Method for Raman Spectra Based on Wiener Estimation	

.....	39
3.1 Introduction.....	39
3.2 Methodology.....	43
3.2.1 Spectral Reconstruction by Wiener Estimation	44
3.2.2 Sample Preparation and Raman Measurements	50
3.2.3 Performance Assessment.....	52
3.3 Results and Discussion	53
3.3.1 Parameters Optimization	53
3.3.2 Denoising Performance	59
3.3.3 Discussion.....	64
3.4 Conclusions	72
Chapter 4 A Chemometric Technique for Target Species Quantification Using Raman Spectroscopy.....	73
4.1 Introduction.....	73
4.2 Methodology.....	77
4.2.1 Principle of Target Quantification Using Orthogonal Projection Jointly with Biochemical Component Analysis (OPBCA)	77
4.2.2 Quantification Assessment.....	80
4.2.3 Experimental Validation	80
4.2.4 Spectral Analysis.....	83
4.3 Results and Discussion	85
4.3.1 Dataset 1: Phantom Measurements	85
4.3.2 Dataset 2: Synthetic data of DOX and Leukemia Cells	89
4.3.3 Discussion.....	91
4.4 Application: Non-invasive Diagnosis of Cornea Infection	97
4.5 Conclusions	101

Chapter 5 A Strategy to Improve the Performance of EEG-based Cross-Individual Pain Prediction	102
5.1 Introduction.....	102
5.2 Methodology	105
5.2.1 Experimental Design	105
5.2.2 EEG Recording	107
5.2.3 EEG Data Analysis	107
5.2.4 Relationship Between sEEG and pEEG	108
5.2.5 Relationship Between Pain Perception and pEEG.....	108
5.2.6 Feature Normalization Based on sEEG	109
5.2.7 Binary Classification and Continuous Prediction	111
5.3 Results and Discussion	112
5.3.1 Relationship Between sEEG and pEEG	112
5.3.2 Relationship Between Pain Perception and pEEG.....	114
5.3.3 Feature Normalization Based on sEEG	115
5.3.4 Pain Prediction	117
5.3.5 Discussion.....	118
5.4 Conclusions	123
Chapter 6 Conclusions and Future Directions.....	125
6.1 Conclusions	125
6.2 Future Work	127
Author Publications.....	132
References.....	133

Glossary of Acronyms

BCIs: Brain-computer interfaces

BCs: Basic components

CPM: Cells per milliliter

DOX: Doxorubicin

EEG: Electroencephalography

EOC: Electrooculographic

EPs: Evoked potentials

F. s: Fusarium solani

fMRI: Functional magnetic resonance imaging

FT: Fourier transform

GC-MS: Gas chromatography-mass spectrometry

GN: Gaussian noise

IMDM: Iscove's modified dulbecco's medium

ISI: Inter-stimulus interval

LDA: Linear discriminant analysis

LEPs: Laser-evoked potentials

MA: Moving average (filtering)

MAPE: Mean absolute percent error

MCR-ALS: Multivariate curve resolution alternating least squares

MLR: Multivariate linear regression

MMSE: Minimum mean square error

MSP: Monosodium phosphate

nRMS_P: Normalized RMS of pEEG

N2P2: Laser evoked potential, N for negative, P for positive, 2 for 200 ms

NCD-WE: Numerical calibration dataset based wiener estimation method

Nd:YAP: Neodymium yttrium aluminiumperovskite

NRS: Numerical rating scale

OP: Orthogonal projection

OPBCA: Orthogonal projection and biochemical component analysis

PCA: Principal component analysis

PCR: Principal component regression

PCs: Principal components

pEEG: Pain-evoked EEG

PLS: Partial least squares

PN: Poisson noise

RMS: Root mean square

RMS_P: RMS of pEEG

RMS_s: RMS of sEEG

rRMSE: Relative root mean square error

SD: Standard deviation

sEEG: Spontaneous EEG

SERS: Surface enhanced Raman spectroscopy

SG: Savitzky-Golay (filtering)

SNR: Signal-to-noise ratio(s)

SNV: Standard normal variate

SVM: Support vector machine(s)

WE: Wiener estimation

WM: Wiener matrix

List of Figures

Figure 2-1. Workflow of signal processing.	13
Figure 2-2. Schematic diagram of a support vector machine (SVM) for a two-class sample set (Adopted from Fig. 7 in [33] with permission from the publisher). 18	
Figure 2-3. Schematic diagram of the operation in the PLS algorithm (Adopted from Fig. 2 in [40] with permission from the publisher).....	20
Figure 2-4. Vibrational and electronic transitions in Raman scattering (Modified from Fig. 1–Box 1 in [61] with permission from the publisher).....	22
Figure 2-5. Schematic diagram of Raman effect with UV, visible and near-IR excitation. S_0 refers to the ground electronic state; S_1 and S_2 refer to the excited electronic states (Adopted from Fig. 2 in [59] with permission from the publisher).	23
Figure 2-6. Interferences in a measured Raman spectrum include such as spikes, background and white noise (Adopted from Fig. 4 in [68] with permission from the publisher).....	25
Figure 2-7. Flowchart of spectral pre-processing in Raman spectroscopy.	25
Figure 2-8. (A) Laser-evoked potentials (LEPs) waveforms with different stimulus energy bins (S1: 1 – 2J; S2: 2 – 3J; S3: 3 – 4J; S4: 4 – 4.5J); (B) LEPs waveforms with different pain perception bins (R1: 0 – 3; R2: 3 – 5; R3: 5 – 7; R4: 7 – 10).	36
Figure 3-1. Composition of the numerical calibration dataset. (A) Clean basic components spectra each with a single Gaussian peak and a variable bandwidth	

and (B) Basic components spectra with additive Gaussian noise. (C) Basic components spectra with additive Poisson noise. "BW" stands for bandwidth.

..... 46

Figure 3-2. Wiener matrices obtained from numerical calibration dataset with different peak locations in spectra. (A) Calibration dataset with GN (upper) and Wiener matrix (lower) with the peak location at λ_1 . (B) Calibration dataset with GN (upper) and Wiener matrix (lower) with the peak location at λ_2 (Modified from Fig. 1 in [130] with permission from the publisher)..... 48

Figure 3-3. Relative RMSE (rRMSE) of denoised Raman spectra measured from the phantom sample using the NCD-WE method (calibration dataset with GN). Graphs from left to right correspond to P1, P2, P3 and P4 (Adopted from Fig. 2 in [130] with permission from the publisher)..... 56

Figure 3-4. Relative RMSE (rRMSE) of denoised Raman spectra measured from the phantom sample using the SG method. Graphs from left to right correspond to P1, P2, P3 and P4. Notice the presence of NAN (not a number) values at the right corner is due to the fact that the SG algorithm cannot be properly conducted when the window length is smaller than the polynomial order (Adopted from Fig. 3 in [130] with permission from the publisher). 56

Figure 3-5. Relative RMSE (rRMSE) of denoised Raman spectra measured from the phantom sample using the MA method. Graphs from left to right correspond to P1, P2, P3 and P4 (Adopted from Fig. 4 in [130] with permission from the publisher)..... 56

Figure 3-6. Relative RMSE (rRMSE) of denoised Raman spectra measured from

the fingernail sample using the NCD-WE method (Calibration dataset with GN).	
Graphs from left to right correspond to F1, F2, F3 and F4.	57
Figure 3-7. Relative RMSE (rRMSE) of denoised Raman spectra measured from the fingernail sample using the SG method. Graphs from left to right correspond to F1, F2, F3 and F4. Notice the presence of NAN (not a number) values at the right corner is due to the fact that the SG algorithm cannot be properly conducted when the window length is smaller than the polynomial order.....	57
Figure 3-8. Relative RMSE (rRMSE) of denoised Raman spectra measured from the fingernail sample using the MA method. Graphs from left to right correspond to F1, F2, F3 and F4.	58
Figure 3-9. Relative RMSE (rRMSE) of denoised Raman spectra measured from the leukemia cells using the NCD-WE method (Calibration dataset with GN). Graphs from left to right correspond to L1, L2, L3 and L4.	59
Figure 3-10. Relative RMSE (rRMSE) of denoised Raman spectra measured from the leukemia cells using the SG method. Graphs from left to right correspond to L1, L2, L3 and L4.	59
Figure 3-11. Relative RMSE (rRMSE) of denoised Raman spectra measured from the leukemia cells using the MA method. Graphs from left to right correspond to L1, L2, L3 and L4.....	59
Figure 3-12. Best denoised results of Raman spectra measured from the phantom sample using the NCD-WE method. (A) – (D) correspond to P1 – P4, respectively (Modified from Fig. 5 in [130] with permission from the publisher).	62

Figure 3-13. Best denoised results of Raman spectra measured from the fingernail sample using the NCD-WE method. (A) – (D) correspond to F1 – F4, respectively.	63
Figure 3-14. Best denoised results of Raman spectra measured from the leukemia cells using the NCD-WE method. (A) – (D) correspond to L1 – L4, respectively.	64
Figure 3-15. Effect of number of PCs on denoising performance for spectra with various SNR when applying the NCD-WE (GN) method to Raman spectra from the phantom sample. For each PC number, the rRMSE values at all window sizes are analysed to yield the mean and standard deviation (error bar) shown in the figure (Modified from Fig. 6 in [130] with permission from the publisher).	66
Figure 3-16. Effect of window length on denoising performance for spectra with various SNR when applying NCD-WE (GN) method to Raman spectra from the phantom sample. For each window size, the rRMSE values at all PC numbers are analysed to yield the mean and standard deviation (error bar) shown in the figure (Modified from Fig. 6 in [130] with permission from the publisher).	68
Figure 3-17. Comparison of the best denoised spectra reconstructed from measurement of the phantom sample using NCD-WE, SG and MA methods. (A) – (D) correspond to P1 – P4, respectively. The test spectrum refers to the noisy spectrum with various SNR without fluorescence background removal. The reference spectrum refers to the clean spectrum with fluorescence background removed, which is meant to be compared to the denoised result using each method (Modified from Fig. 7 in [130] with permission from the publisher)...	70

Figure 4-1. (A) Raman spectra from urea (i.e. target species) with concentrations of 0.033M, 0.067M, 0.133M, and 0.267M. The embedded plot displays the linear relationship between peak intensity and concentration. (B) Upper panel: Illustration of spectra of non-target species generated by mixing monosodium phosphate and potassium formate with different mixing ratios. Lower panel: Representative Raman spectra from monosodium phosphate and potassium formate solution. 86

Figure 4-2. (A) Average Raman spectra from mixtures that are made by mixing the target and non-target species, where the concentrations of target species are 0.033M, 0.067M, 0.133M, and 0.267M. (B) Estimated concentration vs. real concentration of the target species. The standard deviation is indicated by error bars. Note that circle, square, and triangle markers correspond to the estimation results from mixtures with the target species of the identical concentration but non-target species of different constitution. Details about the types of non-target species can be referred to the upper panel in Fig. 4-1 (B). The Dash-dotted line indicates the perfect regression result..... 88

Figure 4-3. (A–B) Average Raman spectra from the target species (DOX) and basic biochemical components. (C) Average Raman spectrum of the non-target species (leukemia cells). The optimal fitting to the spectrum using those of biochemical components as shown in (B) is presented in the red dashed line. 90

Figure 4-4. (A) Illustration of the average Raman spectra from synthetic mixtures that are created by adding up the contributions from the target species (DOX) and non-target species (leukemia cells), as well as Gaussian noise. The

concentration of the target species is varied from 0.2 to 2.0 with an increment of 0.2. (B) Estimated concentration vs. real concentration of the target species. The standard deviation is indicated by error bars. It should be noted that there is no unit for real and estimated concentrations as indicated by “-” since they refer to the ratio of the real or estimated concentration to the unit concentration. The dash-dotted line indicates the perfect regression result.	91
Figure 4-5. (A) Spectra of the target species (DOX) and non-target species (leukemia cells). (B) Illustration of the average Raman spectra from synthetic mixtures excluding the spectral range between 1420 cm^{-1} and 1530 cm^{-1} . (C) Estimated concentration vs. real concentration of the target species. The standard deviation was indicated by error bars. It should be noted that there is no unit for real and estimated concentrations as indicated by “-” since they refer to the ratio of the real or estimated concentration to the unit concentration. The dash-dotted line indicates the perfect regression result.	96
Figure 4-6. Comparison of spectra of F.s fungus and infected cornea, upper panel: original spectra; lower panel: OP-processed spectra.	100
Figure 4-7. Estimated content of F.s in the original and OP-processed spectra.	100
Figure 5-1. Correlation between the mean and SD of RMS_S and RMS_P at levels of (A) $NRS \leq 4$, and (B) $NRS > 4$. Red dots represent the mean or SD of RMS_S and RMS_P , which are averaged across all trials at each pain intensity level for each participant. Gray lines represent the best linear fit (Adopted from Fig. 1 in [119] with permission from the publisher).....	113

Figure 5-2. (A) Relationship between pain ratings and RMS_p (from one participant). Colored dots represent mean $\pm$ SD of RMS_p averaged across trials at each given level of pain perception. The red line represents the fitted global linear model, while the blue lines represent the fitted two-piecewise linear model. (B) Comparison of MSE (mean $\pm$ SD) of all participants between two fitting models (Adopted from Fig. 2 in [119] with permission from the publisher).	115
Figure 5-3. Effects of sEEG normalization on inter-individual variability of (A) binary classification thresholds, (B) slopes of linear regression models for high-pain trials, (C) intercepts of linear regression models for high-pain trials. Noted that mean values were removed from these parameters for illustration (Adopted from Fig. 3 in [119] with permission from the publisher).....	116
Figure 5-4. (A) Difference in pain ratings with the same stimulus energy between males and females. (B) Difference in RMS_p with same rating value between males and females (Modified from Fig. S1 in [119] with permission from the publisher).	122

List of Tables

Table 3-1. Parameters of experimental measurement and SNR of test Raman spectra (Adopted from Table 1 in [130] with permission from the publisher) ..	52
Table 3-2. Parameters tested for NCD-WE, SG and MA methods (Adopted from Table 2 in [130] with permission from the publisher)	54
Table 3-3. Best and average denoising performance (rRMSE) for the ranges of parameters for the NCD-WE (GN and PN), SG and MA methods (Adopted from Table 3 in [130] with permission from the publisher).	60
Table 5-1. Correlation between the mean and SD of RMS_S and RMS_P at each pain level	114
Table 5-2. Comparison of F-statistics obtained from the ANOVA F-test on RMS_P and $nRMS_P$	116
Table 5-3. Comparison of binary classifier thresholds and regression model parameters between using RMS_P and using $nRMS_P$	117
Table 5-4. Accuracy of binary classification and prediction error (MAE) of continuous prediction	117
Table 5-5. P-values of comparisons of pain ratings with the same stimulus energy between males and females	122
Table 5-6. P-values of comparisons of RMS_P with the same rating value between males and females	122

Summary

The dissertation is concerned with developing techniques for the analysis of biomedical signals and Raman spectroscopic data. Three studies are covered in this dissertation, involving two non-invasive and powerful analytical tools, i.e. Raman spectroscopy and electroencephalography (EEG), particularly pain-related EEG.

Raman Spectroscopy can detect the biochemical 'fingerprint' of a material and has been widely accepted as a mature analytical tool for the characterization and identification of molecules, cells and tissues in diverse biomedical applications. Despite its several advantages, such as being rapid and low-cost, non-invasive, non-destructive, and highly specific, Raman spectroscopy suffers from low signal intensity especially for biological samples. Thus, an effective method that can recover decent Raman spectra from Raman measurement with low signal-to-noise ratios (SNR) is highly demanded. On the other hand, the target species of interest is usually mixed with several other species in many circumstances. Quantitative analysis of target species is seriously influenced by the Raman contributions of non-target species in a complex mixture. Multivariate analysis has been extensively used as powerful assistance in the processing of Raman spectroscopic data due to its capability of coping with multidimensional datasets and exploring the complete spectral information simultaneously. There is an increasing demand for problem-oriented multivariate analysis methods to address challenges specific to a study. This

dissertation is focused on developing chemometric methods with regards to these two aspects as described next.

In the first study of this dissertation, with the aim of noise reduction, a method based on Wiener estimation using a numerical calibration dataset is proposed to recover Raman spectra with low SNRs. The method involves two stages, in the calibration stage, the Wiener matrix is created from the calibration dataset. Different from the traditional Wiener estimation based denoising approach, the calibration dataset in the proposed method is created numerically by considering the common characteristics of Raman spectra thus requires no experimental measurement and can be easily extended for analyzing Raman spectroscopic data from diverse categories of samples. Specifically, the calibration dataset includes clean spectra and noisy spectra. The clean spectra are made up of a series of basic components spectra and each contains a single peak numerically generated using a Gaussian function with various parameters, and the noisy spectra are obtained by adding additive noise into the clean spectra. Principle component analysis (PCA) is carried out on the noisy spectra. The Wiener matrix is created to associate the clean spectra with the PC scores obtained by PCA from the corresponding noisy spectra. Moreover, we simplify the process by making use of the observation that a spectrum with multiple peaks can be modelled as the linear summation of several spectra with single peak after proper spectral shifts. We found that the Wiener matrix obtained from a calibration dataset with multiple peaks in spectra is equivalent to the linear summation of all Wiener matrices obtained from a series of calibration

datasets each with a single peak in spectra after the peaks in the column vectors are shifted along the spectral dimension properly. In the reconstruction stage, dimension reduction using PCA is performed to the noisy Raman measurement first, followed by the reconstruction using the obtained Wiener matrix in the calibration stage. The method has been evaluated on Raman spectra of both phantom and biological samples (fingernail and leukemia cells). In general, the spectral shape and the main peaks of the spectrum are nicely reconstructed even in the low SNRs cases. Comparison with two common denoising methods, moving-average (MA) filtering and Savitzky-Golay (SG) filtering, is conducted from the perspective of denoising performance and the sensitivity to parameters choices. In terms of denoising performance, MA yields the worst performance in nearly all cases, and NCD-WE method is comparable to or even better than the SG method especially in lower SNRs cases. Regarding the sensitivity to parameters choices, results indicates that NCD-WE is less sensitive to the parameters choices within suggested range. It implies that a user does not have to make a perfect choice on the parameters of the NCD-WE method to obtain average denoising performance for a noisy spectrum with unknown SNR, which establishes a great advantage of our method over existing approaches, such as SG.

The second study is to address the challenge of quantifying the target species in a mixture with multiple species. A strategy for target quantification based on orthogonal projection and basic component analysis (OPBCA) is proposed, which can be potentially used in optical computing for fast data analysis. Two

steps are included in OPBCA method. In the first step, the contribution of the non-target species is estimated via biochemical component analysis based on the assumption that the non-target species can be approximated as the linear summation of the composing basic biochemical components. The second step of OPBCA is to estimate the concentration of target species by means of orthogonal projection. The effectiveness of the method has been demonstrated on both Raman phantom generated by mixing three chemicals and Raman spectra synthesized from the experimental measurement of doxorubicin (DOX) and leukemia cells. Quantification performance is evaluated by mean absolute percent error (MAPE), and an overall MAPE of 23.2% and 7.4% are obtained for the two datasets, respectively. Moreover, the proposed method requires neither iterations nor training step and is advantageous when the data for non-target species are very limited thus would yield inaccurate modelling in regression-based methods. Furthermore, our method has been extended to a preliminary study of rabbit cornea infection by *Fusarium solani* (F. s) fungus, and shows great potential in the application of non-invasive diagnosis of cornea infection.

Pain is an unpleasant experience related to substantive or potential tissue damage. An effective physiological pain assessment method is highly desired in pain clinical research and practice. EEG-based pain assessment has attracted a growing interest and demonstrated to be a useful alternative to self-report. The third study is related to EEG-based pain assessment, and we are aiming to improve the accuracy of pain prediction at cross-subject level. EEG data induced by nociceptive laser heat pulses with different laser stimulation energies are

collected, and the levels of pain perception reported by the participants using a numerical rating scale from 0 to 10 are also collected to provide the ground truth. We comprehensively investigate the relationship between pain-evoked EEG (pEEG) responses and spontaneous EEG (sEEG) activities, and come up with a hypothesis that one's spontaneous EEG can be used as an individual scale to normalize the pEEG to reduce the inter-individual variability based on the observation that the magnitude of pEEG is significantly correlated with that of sEEG of the same individual at each pain perception level. The relationship between the subjective pain ratings and pEEG responses is also investigated, and the results reveal a nonlinear relationship between the level of pain perception and pain-evoked EEG responses. Upon the observation, a two-stage prediction strategy is proposed, that is to classify the high-pain and low-pain trials first, followed by a continuous prediction of high-pain trials only. Variance of the parameters of both binary classification and linear regression models are significantly decreased, indicating that our normalization method can effectively reduce the inter-individual variability and lead to a higher accuracy of cross-subject pain prediction.

Chapter 1 Overview

1.1 Background and Motivations

Signal is information bearing. However, in most cases, the required information is hidden in the signal structure and may not be easily perceived, especially in some biomedical and clinical applications. Therefore, a proper signal processing procedure is an essential prerequisite prior to the meaningful interpretation and further analysis of the signal. The scientific study of signal processing implicates information theory. Generally, the motivation behind signal processing is to extract relevant information, enhance valuable signal components in noisy measurement (noise removal), extract features that are related to a certain characteristic, and reduce the complexity of classifying data into corresponding categories etc. [1]. This dissertation covers two non-invasive and powerful analytical tools in the field of biomedical applications, i.e. Raman spectroscopy and electroencephalography (EEG), particularly pain-related EEG.

The general background for the first and second studies is that, Raman spectroscopy offers distinct analytical capabilities applicable to a broad range of biomedical applications due to its capacity of highly specific, label-free and non-destructive characterization of biochemical materials. However, the Raman signal is inherently weak, which makes interesting Raman peaks susceptible to noise, especially in biological samples. The measured Raman spectrum is usually considered as the combination of the useful spectrum and additive noise. Noise

reduction and recovery of the clean Raman signal from Raman measurement is a fundamental task to obtain the valuable characteristics of the sample under study before any subsequent spectral analysis. Over the last few decades, various algorithms have been applied to reducing noise in spectroscopic data successfully, such as smoothing algorithms (e.g. moving average filtering [2] and Savitzky-Golay filtering [3, 4]) and frequency-domain algorithms (e.g. Fourier transform (FT) filtering [5] and wavelet transformation [6-9]), and so on. However, most of these existing denoising methods mentioned above demand significant experiences from users to ensure good and consistent performance for Raman spectra with different signal-to-noise ratios (SNR). Therefore, it is desirable to develop a denoising method for which the performance is less sensitive to parameter choices and still decent even for those parameter values in the margins of the usual ranges.

In particular for the second study, the relevant background is that, Raman spectroscopy is often considered as a powerful and well-established tool for quantitative analysis in many fields, such as food quality control [10, 11], pharmaceutical analysis [12-15], microorganism detection and environmental science [16, 17]. A common situation in the context of the quantitative analysis is that the target species is usually mixed with other non-target species. Moreover, spectral contributions from these non-target species are frequently not known beforehand. Therefore, it is of great importance to, first, accurately model the composition of non-target species and the estimation error caused by them, then remove the irrelevant information from non-target species and

find the clean contribution of the target species for quantitative analysis.

The general background for the third study is that, EEG-based pain assessment has attracted growing attention in recent years. In the basic research of pain, EEG activities elicited by nociceptive laser heat pulses are widely used to assess the neural processing of nociceptive pain [18-20]. Taking advantage of the positive relationship between the intensity of perceived pain and a variety of components (such as N2 and P2) in laser-evoked EEG responses [21-27], a technique for the prediction of the level of subjective pain perception using laser-evoked potentials (LEPs) has been successfully developed and proved effective with good predictive accuracy [25]. In this case, however, the classification/prediction model is carried out on a subject-specific paradigm. When it comes to cross-subject paradigm, i.e. to train a classifier/predictor from a set of subjects and then apply it to classify/predict specific characteristics of unknown subjects, the performance is significantly degraded. Challenges are present on the cross-subject approach of pain prediction because both humans' pain perception and nociceptive related EEG responses are characterized by enormous inter-individual variabilities, and mechanisms contributing to those variabilities are sophisticated, involving genetic, environmental, psychological, and cognitive factors [28-32]. Therefore, it is of high importance to develop a strategy to decrease the inter-individual variability of pain-related EEG responses from the perspective of EEG signal processing and improve the performance of cross-subject pain prediction.

1.2 Objectives of This Work

The main objectives of the dissertation can be summarized as follows:

In the first study, we aim to develop a denoising method for Raman spectra preprocessing using Wiener estimation based on the principle of spectral integration and reconstruction. A numerical calibration dataset will be created by considering the common characteristics of Raman spectra, such that the proposed method does not require any experimental measurement and is generally applicable to Raman spectra measured from samples of various categories. In addition, the method is designed to demand for less experience from users in terms of parameter selection compared to the currently popular methods for spectra with unknown noise levels.

In the second study, we attempt to address the challenge of target species quantification using Raman spectroscopy in the context that the target species is mixed with several other species, of which the exact spectral features are limited or even not available in advance. The proposed method is expected to, firstly, obtain the complete and good approximation of the subspace spanned by all possible non-target species by treating the contribution of non-target species as the linear summation of contributions from multiple basic biochemical components that compose these non-target species. Subsequently, the method will quantify the target content by eliminating the irrelevant information and estimate the possible contribution from non-target species.

In the third study, we set our sights on improving the performance of cross-

subject pain prediction using pain related laser-evoked potentials. From the perspective of EEG signal processing technology, we attempt to reduce the inter-individual variability of pain-evoked EEG responses by making use of one's spontaneous EEG. In addition, a two-stage cross-subject pain prediction paradigm, i.e. a binary classifier to discriminate between low-pain ($\text{NRS} \leq 4$) and high-pain ($\text{NRS} > 4$) followed by a linear prediction model to predict the pain ratings ($\text{NRS} > 4$) for high-pain trials, will be developed to further increase the accuracy of pain prediction.

1.3 Organization of The Dissertation

This dissertation is organized as six chapters which can be summarized as follows:

In **Chapter 1**, we introduce the relevant background and problems to be addressed in this dissertation. We describe the motivation and goals of the work. Finally, the organization of the dissertation is provided.

Chapter 2 starts with an overview of the general procedure of signal processing and the main phases involved. As one of the key phases in signal processing, the techniques of multivariate analysis and their use in biomedical applications are comprehensively reviewed. In this dissertation, we focus our attention on developing the problem-oriented signal processing techniques for three studies. The first study is about noise reduction in Raman spectral preprocessing, the second study relates to species quantification using Raman spectroscopy, and the third study is related to laser-evoked EEG responses and its use in cross-

subject pain prediction. For each study, we present an introductory review of the background, key technology and the current state of art reported in the related literature.

In **Chapter 3**, we present a denoising method for Raman spectroscopy based on Wiener estimation. A method for generating the Wiener matrix from a numerical calibration dataset is proposed, which takes into consideration of the common characteristics of Raman spectra such that the method does not require any experimental measurement. Validation of the proposed method is performed on the experimental Raman measurement from two types of samples, i.e. phantom sample (monosodium phosphate) and biological samples (human fingernail and leukemia cells) with various noise levels. Comparison between the proposed method and two commonly used denoising methods, moving-average (MA) filtering and Savitzky-Golay (SG) filtering demonstrates the advantage of our method in low sensitivity to parameter choices when dealing with spectra with various noise levels.

In **Chapter 4**, we discuss the quantification of the target content using Raman spectroscopy when the target species is mixed with multiple non-target species. A method involving orthogonal projection and biochemical component analysis, i.e. OPBCA, is proposed, which is validated on both Raman phantoms and spectra synthesized from the experimental measurement of doxorubicin (DOX) and leukemia cells. The effectiveness of the proposed method for the quantification of the target species and its potential application in non-invasive diagnosis of cornea infection are investigated.

Chapter 5 presents a strategy to improve the performance of cross-subject pain prediction using pain-evoked EEG responses. The comprehensive investigation of the relationship between pain-evoked EEG (pEEG) responses and spontaneous EEG (sEEG) activities as well as the relationship between subjective pain ratings and pEEG responses is conducted. The results reveal a strong inter-individual correlation between the magnitude of pEEG and sEEG, and confirm a nonlinear relationship between pEEG and the subjective pain ratings. Based on the observations, a normalization strategy to reduce the inter-individual variability in pain-evoked EEG responses is proposed and its effectiveness in improving the accuracy of cross-subject pain prediction is discussed in details.

Finally, **Chapter 6** concludes the major findings and contributions in this dissertation, including the denoising method for Raman spectroscopy, problem-oriented target species quantification, and normalization strategy to improve the performance of EEG-based cross-subject pain prediction. At last, we suggest the future directions of each individual study and reiterate their potential values in biomedical signal analysis.

Chapter 2 Introduction

2.1 General Procedure of Signal Processing

Signal is information bearing. Signal processing plays an important role in many biomedical applications. Generally, the objectives of signal processing are to enhance valuable signal components in noisy measurement, extract significant and useful information buried in the signal, reduce the complexity of a signal to facilitate the classification of data into categories. Fig. 2-1 illustrates the typical signal processing workflow, which contains five main phases, i.e. signal acquisition, signal pre-processing, feature extraction, analytical modelling and evaluation.

Figure 2-1. Workflow of signal processing.

Signal processing involves a wide variety of techniques applied to each phase. This section intends to introduce the basic concepts and procedure of signal processing and discuss some commonly used techniques and the applicability of each technique particularly in the fields of Raman spectroscopy and biomedical EEG.

2.2 Multivariate Analysis

Analytical modelling is a broad field involving many analysis in different

perspective, such as parameter analysis, which aims to examine the relationships between different parameters, the typical techniques include t-test, ANOVA, etc.; and those in simulation analysis include the bootstrap and Monte Carlo method etc., intend to figure out the possible outcomes and their probabilities.

In this dissertation, we are mainly concerned the multivariate analysis and its use in biomedical applications. Both spectroscopic data and biological signals can contain a large number of variables and measurement. To handle the large data set and fully exploit the information carried by the signals, multivariate analysis is highly desired [33]. Multivariate analysis refers to statistical analysis techniques that cope with multiple variables (more than one variable) at a time, taking into account the relationship between the variables [34]. For example, multivariate analysis techniques can be used to statistically investigate the correlation between the spectral changes in Raman spectrum and sample characteristics such as analyte concentration, and enable the quantitative analysis of various complex compounds [35] as well as the rapid determination of complex biological mixture components with minor separation or sample pre-treatment [36].

There are mainly two objectives as follows in multivariate data analysis [33]:

- 1). Discrimination and Classification. It refers to classifying the measured data set into two or more groups.
- 2). Regression. It refers to quantify two sets of variables with each other and help to understand the relationship between them, and how one changes

when another one varies.

Below, we summarize commonly used techniques for classification and regression.

2.2.1 Classification Techniques

In clinics, the main goal of classification is to differentiate normal from diseases and distinguish different pathological states [37]. For instance, in conjunction with classification techniques, EEG has been successfully used to diagnose patients with seizure disorders [38], and clinicians are capable of judging if patients are suffering from illness by monitoring their EEG recordings [39]. In statistics, classification aims to explore hidden patterns and tendencies among groups, visualize and categorize observations that share similar intrinsic characteristics into the same group, and assign a proper group for a new unknown observation [40]. In Raman spectroscopy, classification methodologies have been widely accepted in a broad range of applications [41-46]. For example, early detection and diagnosis of diseases are possible with the assistance of classification methodologies in clinical medicine. Classification methods also play important roles in EEG-related applications, such as identification of brain process, brain cancer detection [47], clinical and psychiatric studies [48], as well as brain-computer interface [49].

There are many classification techniques available, such as k-nearest neighbors (KNN), decision trees, random forest, neural networks, etc. Below we discuss two commonly used methods, i.e. linear discriminant analysis (LDA) and support

vector machines (SVM).

Linear Discriminant Analysis (LDA)

Discriminant analysis aims to find the “discriminant axes” that on one hand explains the variance in the original data in a compact manner, on the other another hand facilitates the differentiation of the original data into groups. When projected into the new obtained space constructed with the discriminant axes, the dimensionality of each new observation would be reduced, and meanwhile, the observations belonging to the same group would form a cluster and each cluster is more clearly distinguished with each other [50].

Assume χ_1 with size “ $m_1 \times n$ ” and χ_2 with size “ $m_2 \times n$ ” are matrices that contain samples from two different classes. The definitions of the within-class scatter S_W and between-class scatter S_B are given in Eq. (2-1) and Eq. (2-3), respectively [33].

$$S_W = \sum_{i=1}^N SC_i \quad (2-1)$$

where N is the total number of classes, and

$$SC_i = \sum_{j=1}^{M_i} (S_j - \mu_i)(S_j - \mu_i)^T \quad (2-2)$$

where S_j refers to one observation in the i^{th} class, and μ_i refers to the mean of all observations in the i^{th} class. M_i is the total number of observations in the i^{th} class.

$$S_B = \sum_{i=1}^N (\mu_i - \mu)(\mu_i - \mu)^T \quad (2-3)$$

where μ_i is the mean of observations of the i^{th} class and μ refers to the mean of all observations in all classes.

Linear discriminant is given by the projection direction vector ω that maximizes the objective function which is defined as the ratio of the between-class scatter to the within-class scatter [33, 51] as shown in Eq. (2-4).

$$J(\omega) = \frac{|\omega^T S_B \omega|}{|\omega^T S_W \omega|} \quad (2-4)$$

The intuition about maximizing $J(\omega)$ is to find a direction minimizing the denominator (i.e. within-class variance) of Eq. (2-4) meanwhile maximizing the numerator (i.e. projected class means) in this direction [51].

Support Vector Machines (SVM)

The target of support vector machines (SVM) is to separate the observations from different groups distinctly by a definite gap. The strategy is to find a decision boundary (hyperplane) between groups in a N-dimensional space (N is the number of variables) according to an optimization criterion [33].

The optimal hyperplane in SVM is obtained from the training data in the training phase. Taking a two-class data set (in a two-dimensional space) as an example, to classify the observations into two categories, there are various hyperplanes available. For example, the hyperplanes in green, orange and magenta colors represent three possible hyperplanes in Fig. 2-2. SVM aims to find the optimal hyperplane that maximizes the margin from the boundaries of the two groups, i.e. the maximum distance between the observations of the two groups, as shown in green color [33]. In the test phase, a new observation can be easily classified into a class when projected onto the hyperplane obtained in the training phase.

Figure 2-2. Schematic diagram of a support vector machine (SVM) for a two-class sample set (Adopted from Fig. 7 in [33] with permission from the publisher).

2.2.2 Regression/Prediction Techniques

Numerous regression modelling techniques are available to investigate the relationship between two sets of variables and quantify them with respect to each other. Traditionally, two matrices are involved in a regression method. One is matrix X ($m \times n$) with m observations and n variables, the other is matrix Y with dependent variables. Two important steps are involved in developing the model [33]. The first step is calibration, i.e. using these data matrices to learn the underlying regression model and determine the parameters utilized in the model; the next step is validation, i.e. apply the model obtained from the calibration stage to predict Y -values from new measurement. The estimation of new measurement is of great importance when it is dangerous, difficult, time consuming, or expensive to obtain them experimentally [52-54]. Some

commonly used regression methods are elaborated as follows [33].

Multiple Linear Regression (MLR)

In the case of multiple linear regression (MLR), a linear relation between dependent and independent variables is assumed, and the regression of dependent variable Y on a set of independent variables X is performed based on the least squares criterion [53, 55]. However, one of the problems encountered in spectroscopic applications when using MLR is ill-conditioned least squares. It happens because Raman measurement at different wavenumbers, i.e. the predictor variables, are highly correlated with each other. Principal component regression (PCR) and partial least squares (PLS) are two well-known methods that can solve this problem by projecting the original variables into a lower dimensional space, so that the obtained predictor variables are independent with each other.

Principal Component Regression (PCR)

In principal component regression (PCR), the issue of variable correlation mentioned above is overcome by applying PCA to the X -variables first to project the original variables to a lower dimensional space, then the Y -variables are regressed onto this latent X -variables [56, 57]. Although the introduction of PCA maximizes the variance in the X -variables, latent variables obtained from PCA do not guarantee an optimal prediction result, because the covariance between the predictor X -variables and response Y -variables is not taken into consideration.

Partial Least Squares (PLS)

Partial least squares (PLS) is one of the most widely used analytical algorithms for the estimation and quantification of given components in vibrational spectroscopy related applications. In the cases of MLR and PCR, dependent variables are usually assumed to be independent with each other, in practice, however, they are possibly correlated, because they are predicted from the same pool of independent variables. In PLS, both X and Y matrix are projected into the latent variable space, and the latent variables T and U , respectively, are obtained from singular value decomposition (SVD) with regard to the covariance between predictor variables and response variables, i.e. $X^T Y$, in contrast to $X^T X$ as in case of PCR [56, 57]. Coupled together, T and U maximize the covariance between the predictor variables and response variables and are further used to generate the linear regression model.

Figure 2-3. Schematic diagram of the operation in the PLS algorithm (Adopted from Fig. 2 in [40] with permission from the publisher).

Fig. 2-3 illustrates the operation of the PLS algorithm. Given a new observation, it will be first projected onto the latent variables space defined by T , then the prediction of the latent response variables is conducted based on the regression

model in the U-defined space. Finally, the actual response variables are obtained by transforming the predicted latent response variables from the U-defined space back to the original space.

2.3 Raman Spectroscopy and Spectral Data Analysis

2.3.1 Fundamentals of Raman Spectroscopy

Raman spectroscopy belongs to the category of vibrational spectroscopic techniques that allows to identify a molecule according to the inelastic scattering of incident light by vibrational modes of particular chemical bonds within the molecule. The process of energy exchange between light and molecules, known as Raman scattering or Raman effect [58, 59], was first experimentally discovered by C. V. Raman and K. S. Krishnan in 1928 [60]. When monochromatic light from a laser impinges a material, it can either be absorbed or scattered. The interaction between the light and a vibrating molecule results in spontaneous elastic light scattering, known as Rayleigh scattering (see Fig.2-4) [61]. This part of scattered light (from Rayleigh scattering) has the same frequency as that of the incident light. However, a small fraction of the incident light (approximately 1 out of 10^6 photons) can interact with the material and give rise to the vibration of molecules inside. This process of energy exchange between the incident light and scattering molecules is known as the Raman effect, as illustrated in Fig. 2-4 and 2-5 [59, 61]. The diagram in Fig. 2-5 demonstrates the transition in the Raman scattering that allow a molecule jump from its ground state to an excited vibrational state. This process is accompanied

by the absorption of incident photon and emission of Raman scattered photon simultaneously. Here, Stokes Raman scattering refers to energy transferred from the photon to the molecule, while anti-Stokes Raman scattering refers to energy transferred from the molecule to the photon. The frequency change of this scattered light should equal the vibrational frequency of the scattering molecules since light energy is proportional to frequency. Stokes process is more commonly used in traditional Raman spectroscopy because it is generally more intense than the anti-Stokes process.

One of the main drawbacks of Raman spectroscopy is the relatively low efficiency of the light scattering [1]. The low intensity of Raman signal often limits the speed of the method, and a long-time measurement is required to obtain a Raman signal with an adequate signal-to-noise ratio.

Figure 2-4. Vibrational and electronic transitions in Raman scattering (Modified from Fig. 1–Box 1 in [61] with permission from the publisher).

Figure 2-5. Schematic diagram of Raman effect with UV, visible and near-IR excitation. S_0 refers to the ground electronic state; S_1 and S_2 refer to the excited electronic states (Adopted from Fig. 2 in [59] with permission from the publisher).

Raman spectroscopy implies great potential for diverse applications in biological and biomedical schemes due to the most important characteristics and its remarkable advantages including being rapid, highly specific, non-invasive and non-destructive, culture independent, single cell level available, and so on. Furthermore, Raman spectroscopy differentiates itself from other well-established techniques by its advantages such as ease of use, low cost, as well as minimal sample preparation.

Rapidity is a critical consideration in many applications, for example, reliable and rapid detection of bacteria is unarguably of vital importance in many modern microbiological laboratories, as well as in agriculture, environmental science and clinical practice. The acquisition of Raman spectrum of single cells is possible within seconds, and Raman spectroscopy enables the culture independent detection of bacteria. These characteristics allow Raman spectroscopy being

successfully applied to the early and rapid detection of bacterial infections/disease and pathogen identification, etc. in medical and clinical aspects [62].

In addition, Raman spectroscopy has attracted an increasing attention in medical diagnostic applications due to its high biochemical specificity and the capability of combining with advanced optical technologies in the near-infrared or visible spectral range. Raman spectroscopy has been applied to the *in vivo* and *in vitro* diagnostics based on tissues or cells analysis, such as for the detection of brain cancer [63], breast cancer [45], etc. It also enables powerful biofluid diagnostics [64], such as glucose level monitoring in diabetes management [65] and malaria diagnosis [66, 67].

2.3.2 Pre-processing Techniques in Raman spectroscopy

Pre-processing Procedure

The major procedures for Raman spectra analysis can be classified into two parts: data pre-processing and analytical modeling. In this section, we only focus on the first part, and summarize the techniques that have been widely used in spectral pre-processing in Raman spectroscopy.

Cosmic spikes, background, Gaussian/Poisson noise, and other disturbances related to experimental parameters (Fig. 2-6) [68] are the main contributions that contaminate the Raman spectra. Accordingly, pre-processing is highly demanded for eliminating the effects of undesired signals such as cosmic rays, fluorescence background, Gaussian/Poisson noise, and unwanted signals from

substrate or cell media etc. to allow for further statistical analysis [33, 69].

Figure 2-6. Interferences in a measured Raman spectrum include such as spikes, background and white noise (Adopted from Fig. 4 in [68] with permission from the publisher).

As demonstrated in Fig. 2-7, the process of spectral pre-processing in Raman spectroscopy mainly consists of spike correction, baseline correction, smoothing, normalization, etc. The rest of this section explains in greater detail techniques in each pre-processing step.

Figure 2-7. Flowchart of spectral pre-processing in Raman spectroscopy.

Pre-processing Techniques

Spectral Axis Alignment

Alignment of spectra is important to ensure they share a common spectral axis, because every analysis technique assumes that all the spectra should have the same Raman shift axis. Intensities at a pre-defined common spectral axis can be calculated using the intensities at an existing spectral axis by means of local regression methods [33, 70].

Cosmic Ray/Spike Removal

The cosmic spike is usually removed from the raw spectrum by collecting multiple spectra (one or two extra spectra) in each experiment and comparing them pixel by pixel, then the greater count will be replaced by the smaller one if the difference exceeds the expected noise variance of the detector [71]. The idea takes advantage of the fact that spikes are usually sharper compared to genuine Raman bands. Spike affected regions are also repaired by methods based on local interpolation [72]. Although it is easy and conceptually simple to detect spikes via thresholding on the maximum intensity value, and it has been widely accepted in the spike removal software supplied by many commercial Raman spectroscopy manufacturers, it is not straightforward and it doubles the measurement time [71]. Therefore, several mathematical methods have been explored as alternatives, such as median filters, polynomial, PCA and wavelet transform [7, 71]. An algorithm has been developed recently to determine the spike position according to the evaluation of the second-order derivative [73]. In Raman spectroscopic imaging particularly, the upper bound spectrum method

and the refined nearest neighbor comparison method have been developed, which makes use of the information from unaffected neighbor pixels to correct the spike [71, 72, 74].

Baseline Correction

The origin of the background in Raman spectra include fluorescence, poor rejection of the laser light, and external stray light [75]. Fluorescence is the main source of background in Raman spectroscopy, especially with the extensive use of vibrational spectroscopy in the field of biological science and medicine in recent years. The strong intrinsic fluorescence from biological samples or impurities, solvent, and other components forms the main challenge in Raman spectroscopy, which is often a few orders of magnitude stronger than Raman-bands [33, 76]. Therefore, the step of baseline removal is very important before proceeding with further analysis to extract the useful information from Raman signal.

Many methods have been proposed for the elimination of fluorescence background, each of which follows a different mathematical procedure. In ref [77], a wavelet transform was applied to peak detection and peak width estimation. In refs [78, 79], morphological operations were utilized to approximate the baseline of Raman spectrum. The combination of peak detection and interpolation was proposed to remove the background contribution in a raw Raman spectrum [80]. Other methods involve polynomial fitting [81, 82] and frequency-domain filtering [83]. More details and comparison of these techniques can be found in [69, 84].

Although derivatives [85-87], wavelet transformation [8, 88, 89] and polynomial fitting [82, 87, 90, 91] are major algorithms for baseline correction in the analysis of Raman spectra, they have shortcomings. A derivative based algorithm has a risk of changing the original peak shape. The accuracy of polynomial fitting heavily depends on the users' experience. Wavelet transformation transforms the original signal into different frequency components and simply remove the varying components with low frequency. The operation may result in distortion in part of the spectrum and loss of useful information [77].

Smoothing/Denoising Techniques

Although baseline removal mainly takes care of eliminating unwanted disturbance of low frequency components or broad bands existing in the Raman spectra, some high frequency components, known as electronic noise, can also affect largely on the spectral intensity of Raman spectra. This type of noise includes shot noise, flicker noise and thermal noise, etc. The noise reduction protocol is required to smooth the spectra and increase the signal-to-noise ratio, and accentuate the true spectral features [92]. Lots of algorithms have been applied to reducing noise in Raman spectra, such as moving average filter, Savitzky-Golay (SG) Smoothing and frequency domain transform based algorithms.

- *Moving average filter*

As suggested by the name, moving average filter operates the smoothing by using a moving window of a given size, each data point in the output signal is the average of a number of points inside the window from the original signal, as

in Eq. (2-5) [93].

$$\hat{S}_i = \begin{cases} \frac{1}{2M+1} \sum_{j=-M}^M S_{i+j} & i > M \text{ and } i < N - M \\ 0 & \text{otherwise} \end{cases} \quad (2-5)$$

where S_{i+j} is the input signal with total number of N data points, and $\hat{S}_i$ is the output signal (window length is $2M+1$). Application of moving average filter includes removing high frequency noise with the minimal loss of signal components in the pass-band. However, it works well only for spectra with broad features [2] and the performance is highly dependent on the proper selection of window length.

- *Savitzky-Golay (SG) Smoothing*

Savitzky-Golay (SG) smoothing is one of the commonly used methods for reducing experimental noise in Raman spectra. SG uses a polynomial of degree n (Eq. 2-6) to fit the curve over a few adjacent pixels and replaces these pixels with polynomial curve based on local least squares polynomial approximation (Eq. 2-7) [75, 94]. SG can “smooth out” a noisy signal whose frequency span is large, and it is capable of preserving characteristics such as local maxima and minima and peak width. Despite the effectiveness of SG in reducing noise, there is potential risk of degrading the underlying Raman spectral features due to inappropriate selection of polynomial order and window size [4, 92, 95-97].

$$p(n) = \sum_{k=0}^N a_k n^k \quad (2-6)$$

$$\varepsilon_N = \sum_{n=-M}^M (p(n) - x[n])^2 = \sum_{n=-M}^M \left(\sum_{k=0}^N a_k n^k - x[n] \right)^2 \quad (2-7)$$

where $x[n]$ is a group of $2M+1$ samples of the signal centred at $n=0$, the

coefficients of a polynomial $p(n)$ are obtained to minimize the mean-squared approximation error ε_N for the group of input samples centred at $n=0$.

- *Wavelet transform*

In wavelet transform, spectral data are decomposed into components with various frequency components in the wavelet domain and the denoised spectrum is reconstructed after thresholding for noise removal. However, wavelet transform only supplies fixed and limited filter banks, which may not be the most appropriate ones for the spectrum under study [6]. It is essential to carefully choose the underlying parameters so that none of the important Raman bands are removed when using the wavelet transform algorithm for smoothing.

Normalization

Spectral intensities could vary among multiple measurement from the same sample, under different experimental conditions such as the levels of laser power and alignment. Normalization is a necessary step of pre-processing, which ensures that the intensity of a given Raman band of the same sample is at the same level across the measurement taken under slightly different experimental parameters or conditions. Numerous normalization techniques are available for dealing with the disparity in intensity levels in the literature [69, 98]. Here we summarize some commonly used techniques and give a general description for each. Note that one normalization method may be more appropriate than another in a particular problem to be solved.

1) Vector normalization

The “*norm*” of a spectrum is defined as the square root of the sum of the squared intensities of the spectrum as shown in Eq. (2-8). The normalized spectrum s_{n_i} is obtained by dividing each Raman intensity value s_i corresponding to a Raman shift by the “*norm*” following Eq. (2-9).

$$norm = \sqrt{s_1^2 + s_2^2 + \dots + s_N^2} \quad (2-8)$$

$$s_{n_i} = \frac{s_i}{norm}; i = 1, 2, \dots, N \quad (2-9)$$

2) Min-max normalization

In min-max normalization, the “*maximum*” and “*minimum*” values of the given spectrum are calculated according to Eq. (2-10) and Eq. (2-11), respectively. Then, each Raman intensity is normalized by subtracting the “*minimum*” value from the original intensity value and dividing by “*(maximum – minimum)*” value as in Eq. (2-12).

$$S_{max} = \max(s_1, s_2, \dots, s_N) \quad (2-10)$$

$$S_{min} = \min(s_1, s_2, \dots, s_N) \quad (2-11)$$

$$S_{n_i} = (s_i - s_{min}) / (s_{max} - s_{min}), i = 1, 2, \dots, N \quad (2-12)$$

3) Standard normal variate (SNV) normalization

Similarly to the min-max normalization, in standard normal variate (SNV) normalization, the “*minimum*” and “*(maximum – minimum)*” are replaced by “*mean*” and “*standard deviation (SD)*” as in Eq. (2-13) – Eq. (2-15):

$$mean = \frac{s_1 + s_2 + \dots + s_N}{N} \quad (2-13)$$

$$SD = \sqrt{\frac{(s_1 - \text{mean})^2 + (s_2 - \text{mean})^2 + \dots + (s_N - \text{mean})^2}{N-1}} \quad (2-14)$$

$$s_{n_i} = \frac{s_i - \text{mean}}{SD}; i = 1, 2, \dots, N \quad (2-15)$$

4) Peak normalization

In peak normalization, the normalized spectrum is obtained by dividing each Raman intensity of the spectrum s_i by a reference “ P ”. Here “ P ” refers to the intensity value of a peak corresponding to a particular Raman band.

$$s_{n_i} = s_i / P; i = 1, 2, \dots, N \quad (2-16)$$

With the help of advanced algorithms, numerous comprehensive studies have been reported in the literature [5, 69, 84] on the pre-processing of Raman spectra. However, the art of the pre-processing step is yet to be standardized. In most circumstances, the appropriate pre-processing steps are usually problem dependent, and the performance highly relies on the observations and prior experience of the user.

2.3.3 Raman Spectroscopy as a Quantitative Tool

The well-resolved spectral features of Raman spectroscopy provide good specificity for not only qualitative analysis, but also quantitative analysis, to either determine analyte concentrations or make diagnosis. Although Raman spectroscopy is more often considered as a useful tool for qualitative analysis, its applications in sample composition quantification have been reported in the technical literature for more than half a century [99]. The power of Raman spectroscopy as an alternative technique for quantitative analysis has been demonstrated in various fields, such as food quality control [10, 11], drug

analysis [12-14, 40, 100-102], microorganism detection and environmental science [16, 17] etc. Linear regression modelling is well adopted in the quantitative analysis of Raman spectroscopy, where two assumptions are usually associated with the modelling.

A) The Raman spectrum of a mixture of biochemicals is simply the linear superposition of the Raman spectra of all biochemicals in the mixture;

B) There is linear relationship between signal contribution and concentration for each biochemical.

Commonly used regression methods have been summarized in section 2.2.2.

2.4 Pain-related EEG Responses and Pain Prediction

Biological signal (also known as biosignal or biomedical signal) refers to any biological quantity or magnitude exhibiting variation in the time (1D), space (2D) or time-space (3D) domain that potentially provides information on the status of a biological system. As one of the most well-known 1D biological signal, Electroencephalogram (EEG) is the recording of electrical activity through electrode sensors placed on the scalp [48]. Due to the non-invasive characteristic and relatively low cost of data collection, EEG has been accepted to be a reliable tool in various biomedical applications and clinical diagnosis. For example, EEG has played an important role in the diagnosis of epilepsy, sleep disorders, tumors and other focal brain disorders [103, 104]. Evoked potentials (EPs) is a derivative of EEG. Distinguished from spontaneous EEG, EPs refers to brain electrical potentials in a specific pattern that are time-locked to the

presentation of a stimulus [105]. EPs are more often used in cognitive science [106], mood recognition [107] and psychophysiological research [108] etc. In this study, the possibility to utilize pain related laser-evoked potentials (LEPs) for pain prediction at the cross-subject level is explored.

2.4.1 Laser-Evoked Potentials

Pain is an unpleasant experience related to substantive or potential tissue damage [109, 110]. According to the kind of damage, pain can be categorized into nociceptive pain caused by tissue damage, neuropathic pain caused by nerve damage, and psychogenic pain affected by psychological factors. The dissertation is focused exclusively on the study of nociceptive pain. In clinic practice, the gold standard to determine the presence/absence, and the degree of nociceptive pain perception is based on self-report [111, 112], but it may fail in certain patient populations, e.g., patients who suffer from consciousness disorders or are in coma [113]. Therefore, an effective physiological pain assessment method is highly desired in pain related clinical research and practice. Among a variety of pain-related neural signatures, such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), laser-evoked potentials (LEPs), elicited by nociceptive laser heat pulses usually to the dorsal hand [18], has attracted a growing interest to study the central and peripheral processing of nociceptive sensory input. Because the laser is regarded as the most suitable way to exclusively activate nerve fibers transmitting nociceptive signals [114, 115].

The relationship between LEPs features and nociceptive pain perception has been well recognized in many studies [21-27]. One of the most fascinating findings indicates that the amplitude of LEPs is highly correlated with the subjective pain rating [116], which implies the potential of LEPs served as a measure for subjective pain. LEPs includes several deflections that are time-locked to the start of the laser stimulus [117]. The predominant component of LEPs is a negative-positive biphasic deflection recorded over the midline of the scalp at around 160 ms and 390 ms (referred to as N2 and P2 components, respectively) after the delivery of laser stimulation to the subjects' hand dorsum [18]. Fig. 2-8 (A) displays the LEPs waveforms under different stimulus energy bin. The whole range of stimulus energies was divided into four bins (S1: 1 – 2 J; S2: 2 – 3 J; S3: 3 – 4 J; S4: 4 – 4.5 J), and the LEPs waveform at each bin was obtained by averaging all trials within the same bin. It's evident that N2/P2 amplitudes are positively correlated with the stimulus intensity. Fig. 2-8 (B) presents the LEPs waveforms at different levels of pain perception (measured as the rating values). Similarly, the whole range of rating values was divided into four bins from low to high level (R1: 0 – 3; R2: 3 – 5; R3: 5 – 7; R4: 7 – 10). Positive correlation between N2/P2 amplitudes and the level of pain perception can also be observed.

With the assistance of multivariate pattern analysis, such as classification and prediction algorithms, the laser-evoked potentials (LEPs) has been widely adopted to predict pain perception [118, 119], which has achieved good predictive accuracy. Some commonly used pattern analysis methods have been

described in sections 2.2.1 and 2.2.2.

Figure 2-8. (A) Laser-evoked potentials (LEPs) waveforms with different stimulus energy bins (S1: 1 – 2J; S2: 2 – 3J; S3: 3 – 4J; S4: 4 – 4.5J); (B) LEPs waveforms with different pain perception bins (R1: 0 – 3; R2: 3 – 5; R3: 5 – 7; R4: 7 – 10).

2.4.2 Challenge in Cross-Individual Classification/Prediction

Classification/prediction methods contribute in myriad EEG related clinical studies and realistic practice, e.g. subject-specific emotion recognition [120], diagnosis of epilepsy [38] etc. However, in most of these cases, the classifier/predictor is built individually for each subject based on the subject-

specific paradigm. The obvious disadvantage is that it requires a large amount of EEG data collected from each subject for a reliable training of the classifiers. This causes burdens to the subject, and more importantly, the data are not available in some circumstance. In recent years, an increasing attention has been paid to the feasibility of EEG in clinical diagnosis and EEG-based brain-computer interfaces (BCIs) for the purpose of communication and rehabilitation, however, it is suggested in several studies [121, 122] that a more practical system requires cross-subjects strategy, that is, to train a classifier/predictor on a group of subjects and then apply it to classify/predict specific characteristics of unknown subjects.

However, performance of classification/regression techniques, e.g. pain prediction from laser evoked potentials, at the cross-subject level is largely limited by individual differences. To our best knowledge, few studies have been done on the cross-subjects approach of pain prediction, simply because of the challenge that both human's pain perception and nociceptive related EEG responses are characterized by enormous inter-individual variabilities, and the mechanisms contributing to those variabilities are sophisticated, involving genetic, environmental psychological, and cognitive factors [28-32]. The default assumption in many learning scenarios is that training and test datasets are independently and identically drawn from the same distribution. When the distributions on the training (from the source domain) and test (from the target domain) sets do not match, e.g. when there are individual differences in EEG responses due to structural and functional differences of the brain, we are facing

the issue of sample selection bias or covariate shift. Many cross-domain learning techniques [1, 2] have been proposed to cope with the issue of inconsistency in data distribution. These methods reweight the samples from the training datasets by using the unlabeled data from the test datasets to ensure that the statistical distribution of samples from both the source and target domains are matched with each other.

Chapter 3 A Denoising Method for Raman Spectra Based on Wiener Estimation

3.1 Introduction

Raman spectroscopy has been well accepted as a non-destructive analytical tool for chemical and biological analysis in the last few decades. It is a powerful technique for both qualitative and quantitative characterization of materials including biological samples. However, the Raman signal is inherently weak, making interesting Raman peaks susceptible to noise, especially in biological samples. There are two major sources of noise: the camera and signal itself (shot noise) [4]. As one of the most important preprocessing steps, noise reduction is usually performed before any subsequent Raman spectral analysis. The reduction of noise and recovery of the clean Raman signal from Raman measurement with low signal-to-noise ratio (SNR) is a fundamental task to obtain the valuable characteristics of the sample under study.

Smoothing algorithms are commonly used to reduce noise; however, spectral information might be lost in the smoothing process because smoothing removes noise by essentially reducing the spectral resolution of original measurement [2]. Among various smoothing methods, moving-average (MA) filtering is one of the simplest smoothing algorithms where the value of each spectral point is replaced by the average value of all points in a predefined spectral window centred on it. However, it works well only for those spectra with very broad

features [2], and the performance is highly dependent on the proper selection of window length. Another commonly used smoothing method for reducing noise in Raman spectra is the Savitzky-Golay (SG) algorithm [3, 4]. SG smoothing is conducted by fitting each segment of the original Raman spectrum within a specified window to a polynomial function [123]. The main performance index of the SG filter is determined by the polynomial order and window length [95]. Although this method is useful in noise reduction, it may also suffer from the degradation of the underlying Raman spectral features [4, 95-97]. Particularly, a short window length results in a poor smoothing performance, while a long window length may degrade the spectral resolution due to distortion of the weak spectral features. Therefore, a prudent selection of the polynomial order and window length is of great importance to ensure a tradeoff between denoising performance and spectral resolution, which highly depends on a user's experience. Another category of smoothing methods such as Fourier transform (FT) filtering [5] and wavelet transformation [6-9] are conducted in the frequency domain. These methods decompose a noisy Raman spectrum into high- and low-frequency components in the frequency domain, making use of the fact that the high-frequency components are regarded to be contributed mainly by noise, and they are truncated with a low-pass filter for the denoising purpose. However, over-filtering or under-filtering are the main drawbacks to FT and wavelet filtering when it is not easy to separate the frequency components of Raman spectrum and noise [124]. In addition, FT or wavelet filtering usually introduces some subjectivity into the analysis because their performance

heavily depends on the choices of several complex parameters [5, 124], for instance, the composition of wavelet filters and threshold values in wavelet filtering. In sum, most existing denoising methods such as those mentioned above demand for significant experiences from users to ensure good and consistent performance for Raman spectra with different signal-to-noise ratios (SNR). Therefore, it is desirable to develop a denoising method for which the performance is less sensitive to parameter choices and still decent even for those parameter values in the margins of the usual ranges.

In an early study [125], the clean Raman spectra have been successfully reconstructed from the original Raman measurement of phantoms and bacteria samples with extremely low SNR using the principle of spectral integration followed by Wiener estimation (WE). Based on the minimum mean square error (MMSE) criterion, Wiener estimation is capable of estimating the corresponding high-dimensional data from low-dimensional measurement. However, one significant disadvantage of this method is that a dataset taken from calibration samples similar to the test samples in Raman characteristics is required in Wiener estimation, which greatly limits the application of this method. To get rid of this limitation, later we came up with a method to recover a Raman spectrum using a universal calibration dataset that consisted of Raman spectra measured from each basic biochemical component in the test samples instead of the actual samples [126]. For example, a calibration dataset composed of Raman spectra of proteins, lipids and carbohydrates has been created to recover the Raman spectra of K562 cells and Caki-2 cells. The use of the universal

calibration dataset has showed reconstruction performance comparable to that using the conventional calibration dataset measured from the actual biological samples. The Raman spectra of cells and biological fluid are successfully reconstructed using the universal calibration set based WE without sacrificing spectral resolution. Because of common biochemical components in most biological samples, the proposed universal calibration dataset, which requires only a limited number of Raman measurement, is applicable to any samples with similar Raman property and biochemical composition.

To further simplify the procedure and expand the universality of the WE based denoising method, in this study, we attempt to create a numerical calibration dataset by considering the common characteristics of Raman spectra such that the method does not require any experimental measurement, and is generally applicable to Raman spectra measured from any samples. For the purpose of denoising, the numerical calibration dataset includes both clean Raman spectra and noisy Raman spectra, in which the noisy spectra are created by adding random noise to the clean spectra. Principle component analysis (PCA) is carried out on the noisy spectra, which partially plays a role of denoising while reducing dimensionality. The Wiener matrix is created from the numerical calibration dataset, which associates the clean Raman spectra with the PC scores obtained by PCA from the corresponding noisy Raman spectra.

Experimental Raman measurement of two types of samples, i.e. the phantom sample (monosodium phosphate) and biological samples (human fingernail and leukemia cells) with various noise levels are used to evaluate the method.

Raman spectra with low SNR are obtained by exciting samples sequentially with a series of laser power values during measurement, and a reference spectrum with high SNR is also obtained for comparison using a large laser power and long exposure time with multiple accumulations. Performance comparison in terms of relative root mean square error (rRMSE) between the denoised spectrum and reference spectrum is conducted with the other two commonly used denoising methods, i.e. Savitzky-Golay filtering (SG) and moving-average filtering (MA). The results indicate that the proposed method can recover Raman spectra from noisy measurement with different noise levels, and the method is demonstrated to be significantly less sensitive to the selection of involving parameters than the common SG method. The proposed method provides denoising performance comparable to the SG method and surpasses the MA method especially in low SNR cases.

3.2 Methodology

In this study, we define the SNR as in Eq. (3-1) to evaluate the level of noise in a spectrum:

$$\text{SNR} = E/\sigma \quad (3-1)$$

$$E = \sqrt{\frac{1}{N} \sum_{i=1}^N S_{Ref}(\lambda_i)^2} \quad (3-2)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N [S_{Noisy}(\lambda_i) - S_{Ref}(\lambda_i)]^2} \quad (3-3)$$

where E refers to the root-mean-square intensity of the reference spectrum, which is supposed to be free from noise, calculated according to Eq. (3-2), and

all $S_{Ref}(\lambda_i)$ refers to the spectral intensity at wavenumber λ_i ($i = 1, \dots, N$) in the reference spectrum. σ is the root-mean-square-error of the noisy Raman spectrum relative to the reference Raman spectrum calculated according to Eq. (3-3), where $S_{Noisy}(\lambda_i)$ is the intensity values of the noisy Raman spectrum at the wavenumber λ_i . The fluorescence background is not removed from the reference and noisy Raman spectra prior to the SNR calculation; however, their intensity values are scaled such that the maximum and minimum values in each spectrum fall in the same range.

3.2.1 Spectral Reconstruction by Wiener Estimation

Denoising of Raman measurement using Wiener estimation involves two stages. The first stage is to create a Wiener matrix from the calibration dataset, i.e. the calibration stage, and the second stage is to reconstruct a denoised spectrum from noisy Raman measurement in the test dataset using the obtained Wiener matrix, i.e. the reconstruction stage.

Stage 1: Calibration

In this work, the calibration dataset is designed to include two sets of spectra, i.e. clean spectra and noisy spectra. The set of clean spectra is made up of a series of basic components (BCs) spectra and each component spectrum contains a single peak numerically generated using a Gaussian function as shown in Eq. (3-4). The wavenumber range and increment in the basic components spectra should be identical to those of the measured spectra in the test dataset.

$$I = I_{max} \times e^{-\frac{(\lambda-\lambda_0)^2}{2\sigma^2}} \quad (3-4)$$

where I_{max} is the maximum peak intensity with two possible values, i.e. 1 and 2. λ refers to wavenumber (cm^{-1}), the range of which depends on that of the spectra to be reconstructed in the test dataset. λ_0 refers to the central peak position. σ is related to bandwidth varied within the interval $10\text{-}30 \text{ cm}^{-1}$ with an increment of 5 cm^{-1} . Fig. 3-1 (A) shows some examples of BCs (Gaussian peaks) each with a centre location of 1200 cm^{-1} and a wavenumber range of 600 cm^{-1} to 1800 cm^{-1} but a variable bandwidth.

In addition, the set of noisy spectra in the calibration dataset is obtained by adding one of two kinds of random noises to each clean spectrum as displayed in Fig. 3-1 (B – C): 1) Gaussian noise (GN) spectrum generated using the function ‘awgn’ in MATLAB, yielding an SNR range of 1.51 to 1.77 for the noisy spectra in the calibration dataset according to the definition in Eq. (3-1); 2) Poisson noise (PN) spectrum generated using the function ‘poissrnd’ function in MATLAB, where the original clean spectrum after scaling with respect to intensity is used as the rate parameter to specify the Poisson distribution at each data point. The scaling factor is set to be 10^2 so that the SNR range of noisy spectra in the calibration dataset is varied from 1.26 to 3.09 according to the definition in Eq. (3-1). Although Gaussian noise is commonly used as random noise model, the Poisson noise model has been confirmed to be more suitable in Raman spectroscopy [127]. Therefore, we considered both cases to be comprehensive. In addition, we have verified that no significant difference exists in the denoising performance of our method as long as the SNR of the noisy spectra in the

calibration dataset is lower than those of the spectra in the test dataset.

Figure 3-1. Composition of the numerical calibration dataset. (A) Clean basic components spectra each with a single Gaussian peak and a variable bandwidth and (B) Basic components spectra with additive Gaussian noise. (C) Basic components spectra with additive Poisson noise. “BW” stands for bandwidth.

Then, principal-component analysis (PCA) is performed for the noisy spectra in the calibration dataset according to Eq. (3-5). The process compresses the spectra to reduce data dimension.

$$D_c = FS_{c,L} \quad (3-5)$$

where $S_{c,L}$ is an $(m \times n)$ matrix of the noisy Raman spectra with low SNR, m represents the number of spectra, and n represents the number of discrete wavenumbers in the Raman spectrum. Subscript c refers to the calibration dataset and L refers to low SNR. The $(n \times l)$ matrix F is the transmission spectra of principal components (PCs) based filters, and l is the number of PCs. A truncated $(m \times l)$ matrix D_c , denoted as the PC score matrix, is thus obtained by mapping the original spectra matrix of n variables to a new space of l variables. Matrix D_c will be used for Wiener estimation in the reconstruction later.

Subsequently, Wiener matrix W is created from the calibration dataset

according to Eq. (3-6) [128].

$$W = E[S_{c,H}D_c^T]\{E[D_cD_c^T]\}^{-1} \quad (3-6)$$

where $S_{c,H}$ presents the set of clean spectra with high SNR in the calibration dataset, and D_c is the PC score matrix obtained earlier according to Eq. (3-5). $E[\cdot]$ function refers to ensemble average, and the superscripts T and -1 represent matrix transpose and matrix inverse, respectively. The Wiener matrix W associates the clean Raman spectra with the PC scores obtained from the corresponding noisy Raman spectra by PCA, which partially contributes to the denoising power of our method.

As described above, the established numerical calibration dataset is made up of a series of clean and noisy spectra each with a single Gaussian peak at the same peak location. One of the general accepted assumptions in the quantitative analysis of Raman spectroscopy is that the spectrum of sample is the linear summation of spectra of all basic components present in the sample [129]. Simplified the assumption to a linear model, a spectrum with multiple peaks can be modelled as the linear summation of several spectra with single peak after proper spectral shifts. It can be predicted that Wiener matrices obtained from calibration datasets in which the spectra possess different central peak locations are identical in column vectors except the differences in central peak location. Therefore, the Wiener matrix obtained from a calibration dataset with multiple peaks in spectra is equivalent to the linear summation of all Wiener matrices obtained from a series of calibration datasets each with a single peak in spectra after the peaks in the column vectors of the Wiener matrices are spectrally

shifted along the spectral dimension properly. Fig. 3-2 demonstrates two Wiener matrices obtained from calibration datasets with λ_1 and λ_2 as the central peak location, respectively. The upper panel shows the clean and noisy spectra in the calibration dataset, and the obtained Wiener matrix according to Eq. (3-6) is presented column-wise in the lower panel, where each curve corresponds to one column vector of the Wiener matrix.

Figure 3-2. Wiener matrices obtained from numerical calibration dataset with different peak locations in spectra. (A) Calibration dataset with GN (upper) and Wiener matrix (lower) with the peak location at λ_1 . (B) Calibration dataset with GN (upper) and Wiener matrix (lower) with the peak location at λ_2 (Modified from Fig. 1 in [130] with permission from the publisher).

Stage 2: Reconstruction

In the reconstruction stage, the corresponding denoised Raman spectra are reconstructed from Raman spectra measured with low SNR in the test dataset using the Wiener matrix obtained in the calibration stage. This stage involves

two steps. First, dimension reduction using PCA is performed to the low-SNR Raman measurement to yield the PC score matrix, according to Eq. (3-7).

$$D_t = FS_{t,L} \quad (3-7)$$

where $S_{t,L}$ is a $(m \times n)$ matrix of the low-SNR Raman spectra, m represents the number of spectra, and n represents the number of discrete wavenumbers in the Raman spectrum. Subscript t refers to the test dataset and L refers to low SNR. F is the transmission spectra of PCs based filters, and it should be noted that the number of PCs is the same as that of the calibration dataset. PC score matrix D_t will be used for Wiener estimation in the following step.

Then, the reconstruction is performed step by step using the Wiener matrix obtained from the calibration dataset after spectral shifting with a shifting increment of 1 cm^{-1} . Take the central location λ_j as an example, the reconstructed spectrum $\hat{S}_{\lambda_j}$ is obtained according to Eq. (3-8), then only values falling in a spectral window with a length L and a central location of λ_j are retained, all values outside the window are set to 0. Issues regarding the determination of a proper window length L will be discussed in the subsequent section below.

$$\hat{S}_{\lambda_j} = W_{\lambda_j} D_t \quad (3-8)$$

where D_t is the PC scores of the test dataset calculated according to Eq. (3-7). W_{λ_j} refers to Wiener matrix obtained from the calibration dataset with a central location at λ_j . The final reconstructed spectrum is achieved by averaging all $\hat{S}_{\lambda_j}$ as shown in Eq. (3-9)

$$\hat{S} = \frac{1}{N} \sum_{j=1}^N \hat{S}_{\lambda_j} = \frac{1}{N} \sum_{j=1}^N W_{\lambda_j} D_t \quad (3-9)$$

where $\hat{S}$ is the reconstructed spectrum and N refers to the total number of shifted Wiener matrix. N should be equal to the range of wavenumber in Raman spectra divided by the shifting increment (1 cm^{-1} in this work).

3.2.2 Sample Preparation and Raman Measurements

In this study, Raman measurement from three samples, a phantom, human fingernail and leukemia cells, were utilized to validate the effectiveness of the numerical calibration dataset based wiener estimation method (NCD-WE) for noise reduction. The results of the proposed method were compared with two traditional denoising techniques, i.e. Savitzky-Golay (SG) filtering and moving-average (MA) filtering. Parameter determination of the three methods will be discussed in the first part of “Results” section below.

The phantom was made by dissolving monosodium phosphate (20233-1 KG, Affymetrix, USA) in distilled water to create a solution with a concentration of 1M. The Raman spectra of phantom were measured over a wavenumber range from 600 to 1800 cm^{-1} with an increment of 2 cm^{-1} , using a micro-Raman system (InVia, Renishaw, U.K.) coupled to a microscope (Alpha 300, WITec, Germany) with a 20× objective lens. The excitation wavelength was 785 nm (HPNIR785, Renishaw, UK). The maximum laser power reaching the samples was approximately 218 mW. Spectra with four levels of SNR were collected. This was achieved using 10%, 1%, 0.1%, and 0.05% of the maximum laser power, respectively, and the exposure time was kept at 10s. In addition, a reference

Raman spectrum (with high SNR) was collected from the sample using 50% of the maximum laser power with an exposure time of 20 s and 3 accumulations.

Raman measurements were taken from a human fingernail sample *ex vivo* using the same system as above with laser light focusing on the surface of the fingernail sample. Spectra with four levels of SNR were collected. This was achieved by using 1%, 0.5%, 0.01%, and 0.05% of the maximum laser power, respectively, and the exposure time was 10s. A reference Raman spectrum (with high SNR) was also collected from the sample using 50% of the maximum laser power with an exposure time of 20 s and 3 accumulations.

Leukemia cells (CCL-243, USA) were cultured in Iscove's Modified Dulbecco's Medium (IMDM) added with 10% fetal bovine serum and incubated under a 37°C with 5% CO₂ circumstance. It should be noted that the cells were washed twice and immersed in PBS before Raman measurement to reduce fluorescence background originated from culture media. Raman measurement of cells was conducted using the same system as above. Spectra with four levels of SNR were collected using 10%, 5%, 1%, and 0.5% of the maximum laser power, respectively, and the exposure time was 10s. A reference Raman spectrum (with high SNR) was also collected from the sample using 50% of the maximum laser power with an exposure time of 60 s and 3 accumulations. The state of the cells was fine according to our visual observation and spectral evaluation. In fact, the integrity of cells has been confirmed in the literature when exposed to 115 mW laser power at 785 nm for more than one hour [131].

Table 3-1. Parameters of experimental measurement and SNR of test Raman spectra
(Adopted from Table 1 in [130] with permission from the publisher)

Sample	Test	SNR	Laser power (percent of maximum)	Exposure time (s)	Accumulation
Phantom	Reference	N.A.	50%	20	3
	P1	25.77	10%	10	1
	P2	11.77	1%	10	1
	P3	3.57	0.1%	10	1
	P4	2.51	0.05%	10	1
Fingernail	Reference	N.A.	50%	20	3
	F1	17.31	1%	10	1
	F2	16.16	0.5%	10	1
	F3	6.30	0.01%	10	1
	F4	4.60	0.05%	10	1
Leukemia Cells	Reference	N.A.	50%	60	3
	L1	12.49	10%	10	1
	L2	11.14	5%	10	1
	L3	7.98	1%	10	1
	L4	4.51	0.5%	10	1

The parameters of all experimental measurement and the resulting SNRs of measured spectra are summarized in Table 3-1. The noisy spectra in the test dataset with four noise levels are denoted as P1 – P4 for the phantom sample, F1 – F4 for fingernail, and L1 – L4 for leukemia cells, respectively, hereinafter.

3.2.3 Performance Assessment

The performance of noise reduction is evaluated according to the relative root mean square error (rRMSE) as defined in Eq. (3-10).

$$rRMSE = \sqrt{\frac{\sum_{i=1}^N [\hat{S}(\lambda_i) - S_{ref}(\lambda_i)]^2}{N \times \{ \max[S_{ref}(\lambda_i)] \}^2}} \quad (3-10)$$

where $\hat{S}$ is the reconstructed spectrum from the Raman measurement with low SNR in the test dataset. S_{ref} refers to the reference Raman spectrum with high SNR. λ_i ($i = 1, \dots, N$) refers to the i -th wavenumber, and the function $\max[\cdot]$

provides the maximum value in the entire spectrum.

It should be noted that the intensity values of both the reconstructed spectra and the reference spectrum went through z-score normalization to have a mean of 0 and a standard deviation of 1 before calculating rRMSE. The purpose of normalization is to minimize the difference in spectral intensities due to varied measurement parameters. In addition, fluorescent background was also removed using a wavelet algorithm [77] to facilitate comparison and eliminate the influence of the background.

3.3 Results and Discussion

3.3.1 Parameters Optimization

In addition to NCD-WE, another two widely utilized denoising methods in Raman spectroscopy, Savitzky-Golay filtering (SG) and moving-average filtering (MA), were evaluated on the Raman measurement in low SNR for comparison. Because the performance of a method depends on both the parameters involved and the SNR of the test spectra, we analysed the performance of each method according to the rRMSE of the denoised spectrum relative to the reference spectrum, using a range of values for every parameter for each method. In particular, the parameters involved in NCD-WE include the number of PCs and window length. The SG method also involves two parameters, i.e. polynomial order and window length. It is important to note that SG smoothing algorithm performs properly only if the window length is odd and the polynomial order is designated a value smaller than the window length [96]. In

an earlier study [96], a polynomial order varied from 1 to 11 and a window length varied from 3 to 25 were adopted to test the performance of SG smoothing filter for denoising the Raman spectra of non-structural protein 1, a biomarker for flavivirus origin diseases. Results demonstrated that an optimal polynomial order of 3 and a window length of 13 provided the best smoothing effect, meanwhile, preserved most characteristics in the original Raman spectra. Inspired by the observations in this study, a polynomial order ranging from 3 to 9, and a window length ranging from 5 to 95 were considered to ensure that at least one combination of these parameters would yield the optimal performance in denoising spectra with various SNR. For the MA method, window length is the only adjustable parameter [132]. The range and increment of each parameter in every method are summarized in Table 3-2, which cover most commonly used values.

Table 3-2. Parameters tested for NCD-WE, SG and MA methods (Adopted from Table 2 in [130] with permission from the publisher)

	NCD-WE		SG		MA
	Number of PCs	Window Length	Order	Window Length*	Window Length
Range	3 – 9	1 – 20	3 – 9	5 – 95	1 – 20
Increment	1	2	1	6	1

*Note that the window length for the SG method has to be an odd number.

Fig. 3-3 and 3-4 display the rRMSE distribution of denoised Raman spectra measured from the phantom sample using the NCD-WE method for the ranges of number of PCs and window length and that using the SG method for the ranges of order and window length, respectively, as listed in Table 3-2. The graphs from left to right correspond to P1, P2, P3 and P4, respectively, in both

figures. Here, we only demonstrate the rRMSE distribution of NCD-WE method with GN added in the calibration dataset. As for PN, the best and average performance in terms of rRMSE are summarized in Table 3-3. The window length appears to be an important parameter. The best denoising performance was achieved when the window length values of 5, 7, 11 and 19 were used, respectively, for P1, P2, P3, and P4. However, the method is relatively insensitive to the number of PCs for denoising Raman spectra measured from the phantom sample, especially in the range of 5 to 9. This can be clearly seen from the observation that the maximum rRMSE value is not 86% greater than the minimum rRMSE value at each SNR. In contrast, the performance of SG relies highly on window length especially for lower SNR, e.g. P3 and P4. Only a lower order combined with a longer window length can guarantee a reliable result for P3 and P4, while a short window length (less than 30) works very poorly. There is an obvious trend that a longer window length yields a better result for lower SNR. The maximum rRMSE value can be anywhere from 180% to 406% greater than the minimum rRMSE value across all SNR.

Fig. 3-5 displays the rRMSE of denoised Raman spectra measured from the phantom sample using the MA method for a range of window lengths. The lower the SNR, the longer the window length required to yield better denoising performance.

Figure 3-3. Relative RMSE (rRMSE) of denoised Raman spectra measured from the phantom sample using the NCD-WE method (calibration dataset with GN). Graphs from left to right correspond to P1, P2, P3 and P4 (Adopted from Fig. 2 in [130] with permission from the publisher).

Figure 3-4. Relative RMSE (rRMSE) of denoised Raman spectra measured from the phantom sample using the SG method. Graphs from left to right correspond to P1, P2, P3 and P4. Notice the presence of NAN (not a number) values at the right corner is due to the fact that the SG algorithm cannot be properly conducted when the window length is smaller than the polynomial order (Adopted from Fig. 3 in [130] with permission from the publisher).

Figure 3-5. Relative RMSE (rRMSE) of denoised Raman spectra measured from the phantom sample using the MA method. Graphs from left to right correspond to P1, P2, P3 and P4 (Adopted from Fig. 4 in [130] with permission from the publisher).

Fig. 3-6 and 3-7 display the rRMSE distribution of denoised Raman spectra measured from the human fingernail sample using the NCD-WE method for the ranges of number of PCs and window length and that using the SG method for the ranges of order and window length, respectively, as listed in Table 3-2. The

graphs from left to right correspond to F1, F2, F3 and F4, respectively, in both figures. In general, the performance of NCD-WE is relatively consistent in the entire range of selected parameters for F1 – F4, which can be seen from the smaller range of rRMSE values compared to that for the SG method in Fig. 3-7 except in case of F2. The overall rRMSE distribution of the SG method in Fig. 3-7 is similar to that in Fig. 3-4 for the phantom sample, where the two parameters interacted with each other resulting a noticeable pattern. For instance, a lower order accompanied with a longer window length offers better results for F3 and F4, and the denoising performance is poor when the window length is shorter than 23 for both.

Figure 3-6. Relative RMSE (rRMSE) of denoised Raman spectra measured from the fingernail sample using the NCD-WE method (Calibration dataset with GN). Graphs from left to right correspond to F1, F2, F3 and F4.

Figure 3-7. Relative RMSE (rRMSE) of denoised Raman spectra measured from the fingernail sample using the SG method. Graphs from left to right correspond to F1, F2, F3 and F4. Notice the presence of NAN (not a number) values at the right corner is due to the fact that the SG algorithm cannot be properly conducted when the window length is smaller than the polynomial order.

Fig. 3-8 displays the rRMSE of denoised Raman spectra measured from the fingernail sample using the MA method for a range of window lengths. The trend is quite similar to that in Fig. 3-5 for the phantom sample, where a longer window length is required to yield better denoising performance for the lower SNR.

Figure 3-8. Relative RMSE (rRMSE) of denoised Raman spectra measured from the fingernail sample using the MA method. Graphs from left to right correspond to F1, F2, F3 and F4.

Fig. 3-9 and 3-10 display the rRMSE distribution of denoised Raman spectra measured from the leukemia cells using the NCD-WE method for the ranges of number of PCs and window length and that using the SG method for the ranges of order and window length, respectively, as listed in Table 3-2. The graphs from left to right correspond to L1, L2, L3 and L4, respectively, in both figures. Relatively consistent performance is obtained for the NCD-WE method in the entire range of the selected parameters. As for the SG method shown in Fig. 3-10, the interaction between the two parameters are obvious for L3 and L4, in which the pattern is similar to that in Fig. 3-4 for the phantom sample and Fig. 3-7 for the fingernail sample, and the denoising performance is poor when the window length is shorter than 23 for low SNR cases. Besides, the range of rRMSE value is significantly greater than that of the NCD-WE method at every SNR case.

Fig. 3-11 displays the rRMSE of denoised Raman spectra measured from the

leukemia cells using the MA method for a range of window lengths. Similarly, a short window is more suitable for dealing with spectra with high SNR, while a long window yields better performance in low SNR cases.

Figure 3-9. Relative RMSE (rRMSE) of denoised Raman spectra measured from the leukemia cells using the NCD-WE method (Calibration dataset with GN). Graphs from left to right correspond to L1, L2, L3 and L4.

Figure 3-10. Relative RMSE (rRMSE) of denoised Raman spectra measured from the leukemia cells using the SG method. Graphs from left to right correspond to L1, L2, L3 and L4.

Figure 3-11. Relative RMSE (rRMSE) of denoised Raman spectra measured from the leukemia cells using the MA method. Graphs from left to right correspond to L1, L2, L3 and L4.

3.3.2 Denoising Performance

The best and average denoising performance based on rRMSE of the NCD-WE, SG and MA methods are summarized in Table 3-3. In general, the MA method

yields the worst performance in nearly all cases except for P1 and P2 for the phantom sample. The SG method with optimal parameters works better than the NCD-WE method for denoising spectra with higher SNR. However, the NCD-WE method is comparable or even better than the SG method in denoising spectra with lower SNR. From the average rRMSE data we can see that the performance of NCD-WE is relatively consistent because its average rRMSE spans a smaller range from SNR 1 to SNR 4 compared to the SG method. Comparing NCD-WE (GN) and NCD-WE (PN), the performance of NCD-WE (PN) is slightly better for the phantom sample and leukemia cells in higher SNR cases (e.g. P1, P2 and L2) but worse in lower SNR cases (e.g. P3, P4, F3 and L4).

Table 3-3. Best and average denoising performance (rRMSE) for the ranges of parameters for the NCD-WE (GN and PN), SG and MA methods (Adopted from Table 3 in [130] with permission from the publisher).

	rRMSE ($\times 10^{-3}$)	Best				Average			
		NCD-WE (GN)	NCD-WE (PN)	SG	MA	NCD-WE (GN)	NCD-WE (PN)	SG	MA
Phantom	P1	0.72	0.64	0.39	0.57	1.00	0.81	0.79	1.50
	P2	0.81	0.72	0.54	0.67	1.00	0.93	0.89	1.50
	P3	1.50	1.60	1.40	2.30	1.70	1.80	2.20	2.70
	P4	2.30	2.60	2.00	3.30	2.60	3.10	3.70	3.90
Fingernail	F1	1.10	1.10	1.00	1.30	1.40	1.70	1.70	2.20
	F2	1.10	1.10	0.86	1.00	1.40	1.90	1.40	1.90
	F3	1.20	1.50	1.30	1.80	1.60	2.20	2.20	2.60
	F4	1.50	1.40	1.60	2.50	1.90	2.10	2.50	3.00
Leukemia Cells	L1	2.02	2.09	1.79	2.41	3.00	2.90	2.70	3.50
	L2	2.44	1.89	1.79	2.26	3.10	2.60	2.90	3.30
	L3	2.58	2.60	2.61	3.93	3.20	3.80	4.00	4.80
	L4	2.33	2.46	2.87	3.89	3.00	3.30	4.30	5.00

Fig. 3-12 presents the best denoised results of Raman spectra measured from the phantom sample using the NCD-WE (GN) method. Fig. 3-12 (A) – (D) correspond to P1 – P4 respectively. In the upper panel, the blue line shows the

low-SNR test spectrum, and the red dot line shows the denoised spectrum. In the lower panel, the grey line and the red dot line display the reference spectrum and the denoised spectrum in both of which fluorescence background has been removed using a wavelet algorithm [77] for easy comparison. It can be observed that the spectral shape and the positions of main peaks are accurately recovered using the NCD-WE method, even in the very low SNR case (P4), where the peaks are overwhelmed by noise in the original test spectrum. Another observation is that the major peak intensity values are slightly underestimated, but this problem could be ameliorated by performing normalization on the denoised spectrum. In addition, some artificial peaks are observed in the relatively flat regions in cases of P3 and P4.

Figure 3-12. Best denoised results of Raman spectra measured from the phantom sample using the NCD-WE method. (A) – (D) correspond to P1 – P4, respectively (Modified from Fig. 5 in [130] with permission from the publisher).

Similarly, Fig. 3-13 and 3-14 present the best denoised results of Raman spectra measured from the fingernail sample and leukemia cells, respectively, using the NCD-WE method. The spectral shape and the main peaks of the spectrum are nicely reconstructed, except the very narrow peak at 1000 cm^{-1} peak in the spectra of leukemia cells. Only some small spectral features are not accurately recovered.

Figure 3-13. Best denoised results of Raman spectra measured from the fingernail sample using the NCD-WE method. (A) – (D) correspond to F1 – F4, respectively.

Figure 3-14. Best denoised results of Raman spectra measured from the leukemia cells using the NCD-WE method. (A) – (D) correspond to L1 – L4, respectively.

3.3.3 Discussion

In this study, we develop a denoising method using Wiener estimation based on a calibration dataset created numerically rather than experimentally, to overcome the disadvantages of the latter approach in cost, complexity and applicability because experimental measurement of real samples can be costly and the Wiener matrix obtained in this manner is specific to calibration samples. The clean spectra in the numerical calibration dataset is composed of a series of basic component spectra each with a single Gaussian peak possessing a different width. The noisy spectra in the numerical calibration dataset are obtained by

adding random noise to the clean spectra. Two types of commonly used noises are taken into consideration, i.e. Gaussian noise (GN) and Poisson noise (PN). An actual spectrum with multiple peaks is treated as the linear summation of several basic components' spectra each with a single Gaussian peak after proper shifting in the central location. A similar treatment in which the Raman spectrum of a sample is the linear summation of the spectra of all basic components has been applied earlier in cell and tissue spectra studies [45, 129, 133]. Furthermore, we prove that the Wiener matrix obtained from calibration spectra with multiple peaks is equivalent to the linear summation of all Wiener matrices obtained from those calibration spectra each with a single peak after proper peak shifting along the spectral dimension. This finding significantly reduces the complexity and time of generating Wiener matrices.

Guideline on The Parameter Selection

Our method involves two parameters that may affect the denoising performance. One is the number of principal components in PCA, and the other is the window size used in the reconstruction stage. The number of principal components determines the percentage of variance explained by the components. Since PCA plays a role in not only dimension reduction but also denoising, the components that capture more variance do not necessarily yield better denoising performance. Similarly, an appropriate choice of the window length would ensure the denoising performance while preventing weak peaks from being smoothed out.

Figure 3-15. Effect of number of PCs on denoising performance for spectra with various SNR when applying the NCD-WE (GN) method to Raman spectra from the phantom sample. For each PC number, the rRMSE values at all window sizes are analysed to yield the mean and standard deviation (error bar) shown in the figure (Modified from Fig. 6 in [130] with permission from the publisher).

We systematically analysed the denoising performance in terms of rRMSE by varying the two parameters one at a time. The number of PCs ranging from 3 to 9 with an increment of 1 was utilized to test our method. Since the set of noisy spectra in the numerical calibration dataset contains 10 spectra, including five basic components each with a different Gaussian bandwidth and two variants for every basic component, the maximum number of PCs obtained from the PCA of the calibration dataset is 9. The smallest number of PCs was set to be 3 because the first three components account for more than 80% of the total variance. Taking the spectra of the phantom sample reconstructed using GN based calibration dataset as an example, Fig. 3-15 demonstrates the effect of PCs number on the denoising performance for each SNR case, where the error

bar indicates the standard deviation of rRMSE when various window size is used in the reconstruction. It can be seen that performance of 3 PCs is not as good as that of other number of PCs, which may be due to the loss of useful spectral information in the principal component analysis. In general, 5 – 9 PCs yield equally good denoising performance regardless of the SNR of Raman spectra, so that the number of PCs can be selected freely in this range for denoising Raman spectra with unknown noise level. The same trend can be observed from the data for human fingernail and leukemia cells (results now shown for conciseness).

For selected number of PCs, a series of window length varied from 1 to 20 were used in the reconstruction stage. Similar to Fig. 3-15, Fig. 3-16 demonstrates the effect of window length on the denoising performance for each SNR case, where the error bar indicates the standard deviation of rRMSE when various number of PCs are used. It can be seen that a window length of 5 – 10 is more suitable in high SNR cases, e.g. P1 and P2, while the denoising performance is less dependent on the selection of window length for low SNR cases generally. Overall, a window length within 5 – 16 can provide satisfied results for all SNR (P1 - P4) without significant difference. This rule of thumb is also applicable to human fingernail and leukemia cells.

One implication regarding the selection of two parameters is that a user does not have to make a perfect choice on the parameters of the NCD-WE method to obtain average denoising performance for a noisy spectrum with unknown SNR, which establishes a great advantage of our method over existing approaches,

such as SG.

Figure 3-16. Effect of window length on denoising performance for spectra with various SNR when applying NCD-WE (GN) method to Raman spectra from the phantom sample. For each window size, the rRMSE values at all PC numbers are analysed to yield the mean and standard deviation (error bar) shown in the figure (Modified from Fig. 6 in [130] with permission from the publisher).

Comparison Between GN and PN Model

Many studies have been carried out on noise modelling in Raman spectral data. While Gaussian noise model has been widely used [4, 69, 134, 135], Sompel D Van de *et al.* reported in an earlier study [127] that noise in Raman spectra follows closer to Poisson distribution than to Gaussian distribution. They developed a new version of 'hybrid reference spectrum and principle component analysis' algorithm by incorporating a Poisson noise model. Results indicated that both noise models yield comparable quantification performance for measured data, nevertheless, the Poisson noise model outperformed the Gaussian noise model consistently for the simulated data. For comparison, we

take into account both the Gaussian noise and Poisson noise models when creating the calibration dataset in this study. In Table 3-3, no significant difference was observed between the two noise models in terms of either the best or the average rRMSE values, except that the Poisson model often slightly outperformed in higher SNR cases.

Comparison of the NCD-WE Method with Two Common Denoising Methods

Comparison is carried out among NCD-WE, SG and MA methods from the perspectives of denoising performance and spectral shape recovery. The best denoised spectra (according to rRMSE) reconstructed from the measurement of the phantom sample and leukemia cells using NCD-WE, SG and MA methods are displayed in Fig. 3-17 and Fig. 3-18, respectively, where (A) – (D) correspond to P1 – P4 and L1 – L4. In each graph, the black line shows the noisy Raman measurement without any treatment; the red bold line is the reference spectrum; the magenta line and blue line represent NCD-WE with the GN model and PN models-based calibration datasets, respectively; the orange line and green line are denoised spectra reconstructed using SG and MA methods, respectively. It should be mentioned that the fluorescence background of the reference spectrum and denoised spectra has been removed using a wavelet algorithm [77] for easy comparison.

Figure 3-17. Comparison of the best denoised spectra reconstructed from measurement of the phantom sample using NCD-WE, SG and MA methods. (A) – (D) correspond to P1 – P4, respectively. The test spectrum refers to the noisy spectrum with various SNR without fluorescence background removal. The reference spectrum refers to the clean spectrum with fluorescence background removed, which is meant to be compared to the denoised result using each method (Modified from Fig. 7 in [130] with permission from the publisher).

It can be observed that in the high SNR cases, all methods yield good results and the performance degrades as SNR of the original spectrum decreases. In low SNR cases (P3, P4 and L3, L4), NCD-WE and SG are still capable of recovering the main peaks of the original spectrum accurately, although noises are not completely removed in some relatively flat regions. In contrast, MA works very poorly in these cases; and some spectral regions with large noises are mistakenly retained as peaks, more importantly, there are artificial shifts in the main peaks between the recovered spectrum and the reference spectrum, despite the fact

that MA shows advantage over the other two methods in recovering very narrow peak, e.g. the peak of leukemia cells spectra at 1000 cm^{-1} , in high SNR cases (L1 and L2).

Figure 3-18. Comparison of the best denoised spectra reconstructed from measurement of leukemia cells using NCD-WE, SG and MA methods. (A) – (D) correspond to L1 – L4, respectively. The test spectrum refers to the noisy spectrum with various SNR without fluorescence background removal. The reference spectrum refers to the clean spectrum with fluorescence background removed, which is meant to be compared to the denoised result using each method (Modified from Fig. 8 in [130] with permission from the publisher).

The computational time values for denoising a single spectrum using the NCD-WE, SG and MA methods are 0.82s , $2.8 \times 10^{-3}\text{s}$ and $7 \times 10^{-4}\text{s}$, respectively. The computational efficiency of the NCD-WE method could be improved by performing the reconstruction using the Wiener matrix (after spectral shifting) with an adaptive shifting increment according to the bandwidth of Raman peaks.

3.4 Conclusions

Raman spectroscopy is widely accepted as a non-destructive analytical tool for biochemical and biological analysis in the last few decades. However, the Raman signal is inherently weak, making interesting Raman peaks susceptible to noise, especially in biological samples. Most existing denoising methods such as Savitzky-Golay filtering (SG) demand for significant experiences from users to ensure good and consistent performance for Raman spectra with different signal-to-noise ratios (SNR). In this study, we proposed a method based on the principle of spectral integration followed by Wiener estimation using a numerical calibration dataset by considering the common characteristics of Raman spectra such that the method does not require any experimental measurement for calibration as required in the previous Wiener estimation based denoising method. This method was tested on the Raman measurement of three types of samples, including a phantom sample, human fingernail and leukemia cells, with various noise levels. Compared to two commonly used denoising methods, i.e., moving-average (MA) filtering and Savitzky-Golay (SG) filtering, the performance of NCD-WE method is significantly less sensitive to the choices of parameters. Moreover, this method provides comparable or even better denoising performance in the cases with low signal-to-noise ratios.

Chapter 4 A Chemometric Technique for Target Species Quantification Using Raman Spectroscopy

4.1 Introduction

Raman spectroscopy has been widely accepted as a useful tool for characterization and identification of a large variety of biochemical and biological materials due to its advantages such as non-invasiveness and rapidity. In the past few decades, in addition to the composition determination and qualitative analysis of the material, there is an increasing interest in using Raman spectroscopy for quantitative analysis [99].

The power of Raman spectroscopy as an alternative technology for quantitative analysis has been demonstrated in various fields, such as food quality control [10, 11], pharmaceutical analysis [12-15], microorganism detection and environmental science [16, 17] etc. Nedeljković *et al.* applied Raman spectroscopy in conjunction with the partial least squares (PLS) regression to quantify butter adulteration with margarine [11]. Raman spectroscopy coupled with gas chromatography-mass spectrometry (GC-MS) method has shown potential in the quantitative analysis of essential oil extracted from *Thymus daenensis* [136]. Different approaches based on Raman spectroscopy and chemometric techniques were compared by Dörfer *et al.* in the context of

quantitative mineral analysis [137]. Zhu *et al.* explored the use of Raman spectroscopy in measuring the relative concentration of a given species in the mixture of oral bacteria [16]. Diverse applications of Raman spectroscopy in both chemical and physical quantification of solid-state pharmaceutical have been reviewed [12]. More recently, Raman spectroscopy has attracted growing attention in the application of drug release monitoring. The local concentration of microbicide drugs was quantified with acceptable accuracy by means of confocal Raman spectroscopy in fluids, drug delivery gels, as well as tissues [14]. The process of intracellular drug releasing has been tracked on a graphene oxide-based nanoplatfrom using surface enhanced Raman spectroscopy (SERS) [15], and the SERS spectrum was demonstrated as a potential indicator for monitoring and quantifying the content of drug. These previous studies demonstrate the capability of Raman spectroscopy, with the help of chemometrics methods, in serving as a valuable tool for quantitative analysis, highlighting its advantages in terms of non-invasive, non-destructive and label-free measurement and low requirement in sample preparation.

In many applications, the target species is mixed with several others. Moreover, the spectral contributions from these non-target species are frequently not known beforehand. As a result, spectral preprocessing is usually required prior to quantitative analysis to remove irrelevant information and possible estimation error caused by interference from the non-target species. Blind source separation method such as PCA and ICA are widely adopted for mixture analysis. They are able to separate the mixture into orthogonal or independent

components, but they are computationally expensive for large data sets, and not capable of estimating the accurate coefficients of each component. While in our study, the quantification capability is also desired in addition to separation. On the other hand, suppose the exact composition of the mixture is known, multiple Linear Regression (MLR), principal component regression (PCR) and partial least squares regression (PLSR) are optional methods for the quantification purpose. They are performed based on least squares criterion by treating each component equally in the linear model without priority on the target species. Orthogonal projection is a common method used in spectral data preprocessing, which has been shown effective in non-analyte information removal [138-140], and pure signal identification in a mixture dataset [141]. By mapping the data in the original spectral space to another space containing all the spectral contributions from unwanted sources and subtracting these unwanted contributions from the data, orthogonal projection is capable of extracting the unique characteristics of the target species. Compared to MLR, PCR and PLSR, orthogonal projection does not require iterations thus is simpler and more straightforward. Moreover, there is no need of a training step. However, the performance of target quantification using orthogonal projection highly depends on the accuracy of estimated contributions from non-target species. Most often, only the target species is known for sure while the exact spectral features of other non-target species are limited or even not available at all in practice, which induces a great challenge in the estimation of contributions from non-target species.

To address this challenge, we propose a method to approximate the subspace spanned by all possible non-target species by treating the contribution of non-target species as the linear summation of contributions from multiple basic biochemical components that compose these non-target species. It is inspired by results from the previous studies [45, 126, 129, 133] which demonstrated the Raman spectrum of biological cells can be approximated by the linear summation of spectra contributed by the composing basic biochemical components. While each type of biological cells has its own unique Raman fingerprints, the underlying basic biochemical components, such as nucleic acids, carbohydrates, proteins and lipids, are well known. We extend this method to the estimation of the unknown non-target species, which is then used in orthogonal projection to find the clean contribution of the target species for quantitative analysis.

In this study, the proposed method, i.e. OPBCA, is validated on both Raman phantoms and spectra synthesized from the experimental measurement of doxorubicin (DOX) and leukemia cells. In both cases, the non-target species are accurately estimated with high accuracy, and our results demonstrate the effectiveness of the proposed method for the quantification of the target species. Moreover, the new method does not require a training step and is advantageous when the data for non-target species are very limited thus would yield inaccurate modelling in regression-based methods.

4.2 Methodology

4.2.1 Principle of Target Quantification Using Orthogonal Projection Jointly with Biochemical Component Analysis (OPBCA)

Estimation of Non-Target Species

It is assumed that each measured Raman spectrum of a mixture is the linear summation of contributions from the target species and non-target species in the mixture, and the spectral information of the non-target species are unknown.

The model can be mathematically expressed as follows:

$$X = c_T \cdot T + U + n \quad (4-1)$$

where X represents the measured spectrum, T represents the Raman contribution from the target species at a unit concentration, and c_T is the concentration of the target species. U represents the Raman contribution from the non-target species, and n is the noise term. Note that n will be omitted for simplicity in the following derivation because it is small compared to T and U in typical Raman measurement.

Our OPBCA method for target quantification includes two steps. The first step is to estimate the non-target species in the mixture via biochemical component analysis. To simplify the principle of biochemical component analysis, the non-target species U is assumed to be the linear summation of contributions from two components, which is true in the phantom study as described below, Eq. (4-1) becomes

$$X = c_T \cdot T + a_1 \cdot f_1 + a_2 \cdot f_2 \quad (4-2)$$

where f_1 and f_2 represent the Raman contribution from two components each at a unit concentration, and a_1 and a_2 are the concentrations of the two components.

To estimate a_1 , the orthogonal projection factors V_1 , V_2 , and V_3 are calculated first as demonstrated in Eq. (4-3) – (4-5).

$$V_1 = T \quad (4-3)$$

$$V_2 = OP_{V_1}(f_2) \quad (4-4)$$

$$V_3 = OP_{V_2}(OP_{V_1}(f_1)) \quad (4-5)$$

Then the orthogonal projection of X is performed onto V_1 , V_2 , and V_3 sequentially as shown in Eq. (4-6) – (4-8).

$$X_1 = OP_{V_1}(X) = a_1 \cdot OP_{V_1}(f_1) + a_2 \cdot V_2 \quad (4-6)$$

$$X_2 = OP_{V_2}(X_1) = a_1 \cdot OP_{V_2}(OP_{V_1}(f_1)) = a_1 \cdot V_3 \quad (4-7)$$

$$\hat{a}_1 = \frac{X_2}{V_3} \quad (4-8)$$

Similarly, $\hat{a}_2$ can be estimated in the same way using Eq. (4-6) – (4-8). Afterwards, the non-target species $\hat{U}$ can be estimated as the linear summation of the component spectra with the estimated concentrations, i.e.

$$\hat{U} = \hat{a}_1 \cdot f_1 + \hat{a}_2 \cdot f_2 \quad (4-9)$$

It should be emphasized that the estimation method described above can be easily extended to the situation in which the non-target species can be approximated as the mixture of more than two components. This will be applicable to the synthetic data from doxorubicin and cells in the subsequent

sections, in which the leukemia cells are viewed as non-target species while doxorubicin is the target species. The Raman spectrum of leukemia cells can be modeled as the linear summation of seven basic component spectra as demonstrated previously [129], i.e.

$$X = c_T \cdot T + U = c_T \cdot T + \sum_{i=1}^7 a_i \cdot f_i \quad (4-10)$$

Quantification of Target Species

Gram-Schmidt process is a method to orthogonalize a set of vectors in an inner product space. Using this method, a vector w_k can be derived for a desired vector v_k in the set $V = \{v_1, v_2, \dots, v_k\}$ by Gram-Schmidt process such that it is orthogonal to other vectors in the set:

$$w_k = v_k - \sum_{j=1}^{k-1} P_{v_j}(v_k) \quad (4-11)$$

where the projection operator $P_u(v)$ is defined as $P_u(v) = \frac{\langle u, v \rangle}{\langle u, u \rangle} \cdot u$, and

orthogonal projection operator is defined as $OP_u(v) = v - \frac{\langle u, v \rangle}{\langle u, u \rangle} \cdot u$. Therefore,

it is straightforward to know that

$$OP_u(u) = OP_v(v) = 0 \quad (4-12)$$

Since the contribution of the non-target species $\hat{U}$ is estimated earlier as in Eq. (4-9), the second step of OPBCA is to perform the orthogonal projection of both sides of Eq. (4-1) onto $\hat{U}$, i.e.

$$OP_{\hat{U}}(X) = c_T \cdot OP_{\hat{U}}(T) + OP_{\hat{U}}(U) \quad (4-13)$$

in which the 2nd term approaches zero according to Eq. (4-12), when the estimation of the non-target species $\hat{U}$ is sufficiently accurate. So, the target concentration is:

$$\hat{c}_T = \frac{OP_{\hat{U}}(X)}{OP_{\hat{U}}(T)} \quad (4-14)$$

4.2.2 Quantification Assessment

The accuracy of target quantification is expressed in terms of the mean absolute percent error (MAPE),

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{c_T^i - \hat{c}_T^i}{c_T^i} \right| \times 100\% \quad (4-15)$$

where N is the total number of test samples, c_T^i and $\hat{c}_T^i$ are, respectively, the actual and estimated values of the target concentration in the i^{th} test sample. The MAPE provides a measure of the quantification accuracy across multiple test samples.

4.2.3 Experimental Validation

To determine the feasibility of the proposed method for quantifying the concentration of a target species in a mixture, we tested the method on two sets of Raman measurement. The first set was acquired from phantoms composed of mixture of various chemicals known as Raman scatterers, while the second set was synthesized from the Raman spectra experimentally measured from doxorubicin (DOX) and leukemia cells.

Dataset 1: Phantom Measurement

The phantoms were made by mixing equal volumes of three chemicals in solutions, including urea (V3171, Promega corporation, USA), monosodium phosphate (20233-1 KG, Affymetrix, USA), and potassium formate (294454-500

G, Sigma-Aldrich, USA). Urea served as the target species, while the mixture of monosodium phosphate and potassium formate solutions served as non-target species. The concentrations of the target species in the phantom samples included 0.033M, 0.067M, 0.133M, and 0.267M. At each level, a different mixture of monosodium phosphate and potassium formate served as the non-target species, including non-target species 1 (monosodium phosphate: 0.067M; potassium formate: 0.133M), non-target species 2 (monosodium phosphate: 0.100M; potassium formate: 0.100M), and non-target species 3 (monosodium phosphate: 0.133M; potassium formate: 0.067M).

The Raman spectra of the phantoms were measured using a micro-Raman system (InnoRam-785S) with 20× objectives over a range from 600 to 1500 cm^{-1} at room temperature. The excitation wavelength was 785 nm, and the laser power was 50 mW. The exposure time in Raman measurement was 10 sec, and the spectral resolution was 4 cm^{-1} . For each of the samples, 20 spectra were collected from the randomly selected positions of a droplet with a volume of 10 μL spread out on an aluminum foil.

Dataset 2: Synthetic Data of DOX and Leukemia Cells

Doxorubicin (DOX), as a widely used anticancer drug, has been well recognized to be effective in the therapy of many malignant diseases, including leukemia, breast and lung cancer, etc. [142-144]. In this study, the SERS spectrum of DOX (Sigma-Aldrich) was measured from the DOX-loaded Ag NP@N-GQD nano-complex. The Ag NP@N-GQD nanoparticles were synthesized using a published recipe [145]. The DOX solution was prepared by dispersing 0.25 mg DOX in 5 mL

PBS solvent mixed with 10 μL dimethyl sulfoxide (DMSO). DOX-loaded Ag NP@N-GQD were synthesized by incubating the DOX solution in 5 mL Ag NP@N-GQD nano-complex at pH 7.4 for 24 h at a concentration of 0.2 mg/mL, followed by centrifugation to remove the excess DOX in the PBS buffer (pH=7.4). The spectral measurement of DOX were obtained using the same micro-Raman system mentioned above over a range from 600 to 1800 cm^{-1} , with the excitation wavelength of 785 nm and the exposure time of 10s. For each of the samples, 20 spectra were collected from randomly selected positions of droplet with a volume of 10 μL spread out on an aluminum foil.

Culture of leukemia cells has been introduced in section 3.2.2. The cell density was maintained to be 10^6 cells per milliliter (CPM). The Raman spectra of leukemia cells were measured using a micro-Raman system coupled to a microscope (Alpha 300, WITec, Germany) with 50 $\times$ objectives, over a range from 600 to 1800 cm^{-1} , with a spectral resolution of 2 cm^{-1} . The excitation wavelength was 785 nm and the exposure time was 10 s.

The test data were synthesized by computationally adding up the spectra of DOX and leukemia cells, each spectrum with a particular weight, where DOX was treated as the target species, while leukemia cells served as the non-target species. The spectral intensity from DOX or leukemia cells corresponds to a unit concentration under the same measurement configuration. We always refer to the concentration relative to the unit concentration in the dataset 2 unless stated otherwise. In the synthetic data, the relative concentration of the target species was varied from 0.2 to 2.0 with an increment of 0.2. The relative

concentration of the non-target species was set to be 1. It should be noted here that the peak intensities in the actual spectrum of the non-target species is proportional to that in the estimated spectrum. Therefore, orthogonal projection on the estimated spectrum of the non-target species can eliminate the possible contribution of the non-target species regardless of its actual concentration. For each concentration level of the target species, 20 observations were generated, where 10% of Gaussian noise were added besides the contributions of the target species and non-target species.

Basic Biochemical Components

To fit to cell spectra in general, measurement was taken for seven basic biochemical components including DNA, RNA, actin, albumin, glycogen, phosphatidylcholine and triolein. These components were selected to represent the major biochemical groups, such as nucleic acids, proteins, lipids and polysaccharides, in cellular compositions. The experimental configuration of Raman measurement for these biochemical components was identical to that for leukemia cells.

4.2.4 Spectral Analysis

Preprocessing

To reduce the variations in measurement, preprocessing steps including spike removal, smoothing and baseline correction were first performed on every Raman spectrum. Then fluorescent background was removed using an algorithm based on continuous wavelet transform and background fitting [77].

In the phantom study, the spectral intensities of the three selected chemicals were proved to be comparable with each other at the same concentration, thus no further normalization was performed. In the synthetic data of DOX and leukemia cells, the Raman spectrum of each individual component was normalized by dividing all spectral values by the maximum intensity before synthesizing the spectra of mixtures to avoid bias on detection selectivity to any of the composition.

Fitting

In our strategy, we attempted to estimate the Raman contribution of the non-target species by taking advantage of the validated observation [45, 129] that the Raman contribution of the non-target species can be approximated by the linear summation of spectra of the composing biochemical components. Two assumptions are made in the approximation:

- 1). The Raman spectrum of a mixture is simply the linear superposition of all composing biochemical components' contributions;
- 2). The Raman contribution of each biochemical component is linearly proportional to its concentration [59].

In other words, the acquired spectra can be assumed to be the weighted summation of spectra of the basic biochemical components. This model has been utilized in the study of both cells [129] and tissues [45, 133], using spontaneous Raman spectroscopy and found practically useful.

In this study, the modelling of leukemia cells' spectrum on the seven basic

biochemical components was conducted using the function 'lsqcurvefit' in MATLAB with a non-negative constraint. Then, the R-square (R^2) value was utilized to evaluate the goodness of fitting. It should be emphasized that fitting is not a required step in quantification, but to provide support for our assumption that leukemia cells' spectrum can be approximated as weighted sum of pure spectra of the basic biochemical components.

4.3 Results and Discussion

4.3.1 Dataset 1: Phantom Measurements

Raman spectra from urea solutions with concentrations of 0.033M, 0.067M, 0.133M, and 0.267M are shown in Fig. 4-1 (A). There is one prominent peak at 1005 cm^{-1} . Moreover, the linear relationship between the peak intensity at 1005 cm^{-1} and urea concentration is observed.

Monosodium phosphate owns two prominent peaks at 877 cm^{-1} and 1078 cm^{-1} , and potassium formate owns one prominent peak at 1353 cm^{-1} . Raman spectra of the two basic components and the three types of non-target species are shown in Fig. 4-1 (B). The concentrations of monosodium phosphate and potassium formate in the non-target species are 0.067M and 0.133M (non-target species 1), 0.100M and 0.100M (non-target species 2), 0.133M and 0.067M (non-target species 3), respectively.

Figure 4-1. (A) Raman spectra from urea (i.e. target species) with concentrations of 0.033M, 0.067M, 0.133M, and 0.267M. The embedded plot displays the linear relationship between peak intensity and concentration. (B) Upper panel: Illustration of spectra of non-target species generated by mixing monosodium phosphate and potassium formate with different mixing ratios. Lower panel: Representative Raman spectra from monosodium phosphate and potassium formate solution.

The representative average Raman spectra of such phantoms for non-target species 3 are shown in Fig. 4-2 (A), in which the concentrations of the target

species cover 0.033M, 0.067M, 0.133M, and 0.267M. It can be seen from the spectra that, the peak intensity values of monosodium phosphate and potassium formate (at 877 cm^{-1} , 1078 cm^{-1} , 1353 cm^{-1}) are similar across all spectra; while the peak intensities of urea show significant differences, which are proportional to the corresponding concentrations.

Fig. 4-2 (B) demonstrates the results of quantification for target concentration. For each level of target concentration, the estimation results are presented in three groups respectively (circle, square, and triangle markers), corresponding to non-target species 1 through 3. In general, all markers are clustered around the perfect regression line, which indicates the consistency in the absolute estimation accuracy regardless of the difference in the non-target species. There appears to be a systematic trend in the estimation error, i.e. the markers representing estimated values appear to follow a straight line slightly different from the perfect regression line in slope. The overall MAPE calculated from all estimations is 23.2%, which is primarily attributed to the samples with the target species at small actual concentrations. If the data corresponding to these small real concentrations such as 0.033M are removed, the overall MAPE will drop to 11.0%.

Figure 4-2. (A) Average Raman spectra from mixtures that are made by mixing the target and non-target species, where the concentrations of target species are 0.033M, 0.067M, 0.133M, and 0.267M. (B) Estimated concentration vs. real concentration of the target species. The standard deviation is indicated by error bars. Note that circle, square, and triangle markers correspond to the estimation results from mixtures with the target species of the identical concentration but non-target species of different constitution. Details about the types of non-target species can be referred to the upper panel in Fig. 4-1 (B). The Dash-dotted line indicates the perfect regression result.

4.3.2 Dataset 2: Synthetic data of DOX and Leukemia Cells

The average Raman spectra of DOX and seven biochemical components including actin (Ac.), albumin (Al.), DNA, glycogen (Gl.), phosphatidylcholine (Ph.), RNA, triolein (Tr.) are displayed in Fig. 4-3 (A) and (B). Fig. 4-3 (C) shows the average Raman spectrum of leukemia cells (non-target species) and its optimal fitting using the spectra of basic biochemical components. The R^2 value for the optimal fitting is 0.949.

Fig. 4-4 (A) shows average synthetic Raman spectra. Note that the synthetic Raman spectra were generated by adding up the spectra of DOX and the non-target species, as well as Gaussian noise, in which the concentration of DOX, the target species, was varied from 0.2 to 2.0 with an increment of 0.2. It can be observed that the overall spectral shape changes with the increase in the target concentration. The quantification result for the target species concentration is presented in Fig. 4-4 (B), and the overall MAPE values is 7.4%. There appears to be a systematic trend in the estimation error, i.e. the markers representing estimated values appear to follow a straight line slightly different from the perfect regression line in slope. This is quite similar to that found in Fig. 4-2 (B). This interesting observation will be discussed later.

Figure 4-3. (A–B) Average Raman spectra from the target species (DOX) and basic biochemical components. (C) Average Raman spectrum of the non-target species (leukemia cells). The optimal fitting to the spectrum using those of biochemical components as shown in (B) is presented in the red dashed line.

Figure 4-4. (A) Illustration of the average Raman spectra from synthetic mixtures that are created by adding up the contributions from the target species (DOX) and non-target species (leukemia cells), as well as Gaussian noise. The concentration of the target species is varied from 0.2 to 2.0 with an increment of 0.2. (B) Estimated concentration vs. real concentration of the target species. The standard deviation is indicated by error bars. It should be noted that there is no unit for real and estimated concentrations as indicated by “-” since they refer to the ratio of the real or estimated concentration to the unit concentration. The dash-dotted line indicates the perfect regression result.

4.3.3 Discussion

The use of Raman spectroscopy in both qualitative and quantitative analyses has

been well documented. Currently, in combination with chemometric methods such as PCR and PLSR, Raman spectroscopic technology allows the quantification of a given compound. Quantification of a target species in a complex mixture with multiple unknown non-target species is a general problem frequently encountered in many applications such as the identification of pathogens in organ infection and the contaminants in water quality monitoring. Such a problem is often solved using regression-based methods, such as multivariate linear regression (MLR), partial least squares regression (PLSR) and principal component regression (PCR). These regression-based methods typically yield good accuracy; however, they in general require a special dataset that contains the spectra of samples with known concentrations of the target species for training. This is not always available. Moreover, these methods frequently need to perform iterations thus hard to be implemented optically to speed up quantification as demonstrated previously [146]. In contrast to these regression-based methods, our method, i.e. orthogonal projection based on biochemical component analysis does not require training dataset thus will be more cost effective and convenient to use. Moreover, the new method is simple and straightforward in principle without the need of iterative operations and could be implemented optically in the future to speed up quantification.

The new method is built on two assumptions:

1. The Raman spectrum of a mixture is the linear superposition of contributions from the components (including both the target species and non-target species) that make up the mixture, and there is a linear relationship between the spectral

intensity and the concentration of component. This assumption is true when there is no significant interaction between the two species, or the interaction does not significantly alter their Raman contributions. This model is widely accepted as the basis for quantitative analysis in spectroscopic techniques [59, 99] and has been shown valid in many previous studies [11, 12, 16, 99, 147]. If this assumption is not exactly true, the linearization of the non-linear response will be needed prior to the application of the proposed method.

2. The Raman contribution of all non-target species can be approximated as the linear summation of all composing basic biochemical components. This assumption has been explored in several previous cell and tissue studies. Shafer-Peltier and Haka *et al.* successfully constructed a Raman model of breast tissue using the Raman spectra measured from cell cytoplasm, cell nucleus, collagen, cholesterol-like and water, etc. [45, 133]. Ong *et al.* fit the Raman spectra of K562 leukemia cells to the linear summation of contributions from seven basic biochemical components [129]. In these past studies, such an assumption was utilized to find the relative concentrations of all known biochemical components and facilitate the understanding of the physiological mechanism behind changes in Raman contributions.

There are three potential error sources as listed below:

- 1) The goodness in the estimation of Raman contribution from the non-target species determines the smallest target concentration that can be accurately quantified. For example, there were a total of four different concentrations in Fig. 4-2, i.e. 0.033M, 0.067M, 0.133M, and 0.267M, for the target species,

urea, for which the MAPE were 60.1%, 19.2%, 5.2% and 8.7%, respectively. It can be seen clearly that the accuracy of target quantification degrades with the decreasing concentration of the target species. This is due to the fact that the Raman contribution of the target species has become so small that the error generated when estimating the contribution of the non-target species started to significantly affect the target species quantification.

The key factors for the accurate estimation of Raman contribution from the non-target species using the contributions of basic biochemical components are to cover all potential biochemical components and minimize the fitting residuals. Any missing component could induce error in the Raman contribution of the non-target species and this error would be in turn relayed to the quantification of the target species. Therefore, in practice, it is suggested to check first whether all basic components are covered by fitting the linear summation of these components' spectra to the measured spectra of control samples that contain no target species. Frequently, the background Raman contribution from the substrate or container may also need to be included in addition to the underlying biochemical components.

- 2) The spectral similarity between the target species and non-target species affects the accuracy of target quantification. A highly overlapping region between the spectra of DOX and leukemia cells can be observed over a range from 1420 cm^{-1} to 1530 cm^{-1} (see Fig. 4-5 A). To demonstrate the influence of the overlap in the spectra of target species and non-target

species on the quantification accuracy, we performed our method on the synthetic Raman spectra generated in the same way as that in dataset 2, except that the range from 1420 cm^{-1} to 1530 cm^{-1} was excluded, as shown in Fig. 4-5 (B). The new overall MAPE values is 4.4%, which is significantly smaller than the MAPE value of 7.4% obtained according to dataset 2 using the whole range.

- 3) Another potential error source is the purity of Raman spectra. The fluorescence residual in a measured Raman spectrum that is not totally removed in spectral preprocessing can increase the error in the quantification of the target species.

Figure 4-5. (A) Spectra of the target species (DOX) and non-target species (leukemia cells). (B) Illustration of the average Raman spectra from synthetic mixtures excluding the spectral range between 1420 cm⁻¹ and 1530 cm⁻¹. (C) Estimated concentration vs. real concentration of the target species. The standard deviation was indicated by error bars. It should be noted that there is no unit for real and estimated concentrations as indicated by “-” since they refer to the ratio of the real or estimated concentration to the unit concentration. The dash-dotted line indicates the perfect regression result.

4.4 Application: Non-invasive Diagnosis of Cornea Infection

In recent years, Raman spectroscopy has been gradually accepted as a hot and mature analytical tool for the rapid and non-invasive characterization and identification of different molecules, cells and tissues in biomedical applications [61]. Raman spectroscopy provides broad information not only on the structure and the chemical constitutions, but also the interactions of biomolecules within the substance, thus it reveals the chemical fingerprints of samples and enables the specification of the minute biochemical differences between samples. Raman spectroscopy has shown tremendous power for the characterization and identification of microorganisms including bacteria, fungi, etc. at species or even strain levels [148].

Bacterial investigation is an important area which relates to wide aspects of modern life. In medical and clinical fields, it is highly significant if the source of a bacterial infection can be identified timely and promptly, because it can save patients' life. Besides, the issue of bacteria-caused hospital hygiene has drawn enormous attention recently. Test procedures are required in industrial production to prevent contaminated food from being released into the market. Constant control of bacterial contamination for the supply system of drinking water is also of great importance to guarantee that the consumers will not be exposed to any healthy risk. In addition, detection of bacteria is a hot topic in environmental science, because the spread of plant pathogens is a huge risk to

the whole ecosystems and can cause severe economic damage [149]. Corneal infection due to microorganisms such as bacteria and fungi, is a common disease. In clinical practice, the gold standard of the microorganism identification is based on their morphological characteristics [149]. Although it is reliable and effective, this technique comes along with some disadvantages. The process of scratching the infected samples is invasive and labor intensive, and it usually takes a few days for the culturing depending on the bacterial species. Moreover, it is assumed in this approach that the bacteria can always be cultured; however, it is not applicable to all bacterial types. Therefore, Raman spectroscopy has been considered as a promising alternative to diagnose cornea infection and quantify the content of microorganism in the infection.

Rabbit cornea infected by *Fusarium solani* (F. s) fungus was investigated in this work. The culture of *fusarium solani* fungus is conducted following the published protocol [150]. Yong New Zealand White Rabbits (Weight varying from 1.5 kg to 2.5 kg) purchased from the National University of Singapore were used for this study and all animal experiments were conducted in compliance with the ARVO statement for the use of animals in ophthalmic and vision research. Rabbits were locally immunosuppressed by injecting Dexamethasone (Dexamethasone, Giemsa stain, Sigma, St. Louis, USA) 0.8mg/eye/day sub conjunctiva for 5 days pre-inoculation. Rabbits were anesthetized by injecting intramuscularly the mixtures containing Ketamine (50mg/kg) (Ketamine, Sigma, St. Louis, USA) and Xylazine (5mg/kg) (Xylazine hydrochloride, Sigma, St. Louis, USA). Then 1% xylocaine (Xylocaine Ointment, Guardian, Singapore) was used to anesthetize

the rabbit cornea before fungus inoculation. Pre-inoculation slit-lamp examination was done for checking the corneas free from infections and inflammation. Then 10 μL of the prepared fungal solutions ($\approx 1.2 \times 10^5$ cfu/mL) was injected into the corneal stroma by using insulin syringes (30G, Becton Dickinson, New Jersey, USA) with the aid of dissecting microscope. Rabbit cornea were infected by *fusarium solani* for 72 hrs. The infectious status was confirmed by slip lamp examination after the infection. Then the rabbit eyes were harvested and deposited in a container for one hour prior to Raman measurement.

The experimental configuration of Raman measurement for the rabbit cornea was identical to that for leukemia cells. All the spectra were acquired over the spectral range from 300 to 2000 cm^{-1} with an excitation power of 50 mW. The excitation power and exposure time values were selected to enhance the spectral intensity without inducing thermal damage to fungus [151].

As mentioned before, the Raman spectra measured from the infected cornea can be regarded as a mixture that contains information of multiple components, including corneal tissues, microorganism, and possibly other unknown components. To remove the associated spectral contributions from other components, the proposed orthogonal projection (OP) strategy was applied to a basis of interference, which contains all chemical information except for the one of interest, i.e. F. s. Fig. 4-6 compares the Raman spectra of F. s and infected cornea before and after the OP treatment. An increase in similarity can be clearly observed after the OP processing. Besides, the estimated concentrations of F.s

increased significantly, as shown in Fig. 4-7.

Figure 4-6. Comparison of spectra of F.s fungus and infected cornea, upper panel: original spectra; lower panel: OP-processed spectra.

Figure 4-7. Estimated content of F.s in the original and OP-processed spectra.

Results reveal the effectiveness of orthogonal projection in removal of relevance information of non-target species shared with the target species. It also implies the potential of Raman spectroscopy coupled with the proposed chemometrics in identifying the presence of F.s on the surface of cornea, thus can be used for

non-invasive diagnosis of cornea infection in the future study.

4.5 Conclusions

Target species quantification is highly demanded in many applications such as pharmaceutical research, biological sample analysis, and food quality control, in which Raman spectroscopy has been established as a label-free analytical tool. In this study, a method involving orthogonal projection and biochemical component analysis, i.e. OPBCA, is proposed, which aims to quantify the target content using Raman spectroscopy in a common context that the target species is mixed with multiple non-target species and the exact spectral features of the non-target species are unavailable. The proposed method is validated on both the spectra of Raman phantoms and spectra synthesized from the experimental measurement of doxorubicin (DOX) and leukemia cells. In both cases, the non-target species are accurately estimated with high accuracy, and our results demonstrate the effectiveness of the proposed method for the quantification of the target species. Moreover, the new method requires neither iterations nor training step and is advantageous when the data for non-target species are very limited thus would yield inaccurate modelling in regression-based methods.

Chapter 5 A Strategy to Improve the Performance of EEG-based Cross-Individual Pain Prediction

5.1 Introduction

Pain is a displeasing experience related to substantive or potential tissue damage [109, 110]. According to the kind of damage, pain can be categorized into nociceptive pain caused by tissue damage, neuropathic pain caused by nerve damage, and psychogenic pain affected by psychological factors. Nociceptive pain is under studied in this work. The gold standard to determine the presence/absence and the degree of nociceptive pain perception in clinic practice is based on self-report [111, 112], but it may fail in certain patient populations, e.g., patients who suffer from consciousness disorders or are in coma [113]. Lack of accurate pain assessment in these populations can result in insufficient or suboptimal treatment of pain. Therefore, it is of high importance to develop a physiology-based pain assessment method that is independent of participants' subjective rating [109, 118, 152].

As a sensory perception that involves a complex set of brain activities, nociceptive pain has been under intensive investigations using brain imaging techniques, such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). A variety of neural signatures of pain have been

identified from brain imaging data, and the possibility to utilize pain-related neural signatures for pain assessment has been explored in many studies [18, 19, 109, 121, 122, 153, 154]. For example, a SVM trained on fMRI data was verified to be possible for pain assessment [109]. Schulz *et al.* also applied a multivariate pattern analysis to predicted individual's pain sensitivity using single-trial pain-related EEG [122].

Particularly, EEG-based pain assessment has attracted a growing interest in recent years, not only because the EEG technique is cheap, easy-to-use, and non-invasive, but also because the relationship between EEG responses and nociceptive pain perception has been relatively well recognized. In basic research of pain, EEG activities elicited by nociceptive laser heat pulses are widely used to assess neural processing of nociceptive pain [18-20]. A positive relation between the intensity of perceived pain and a variety of components (such as N2 and P2) in nociceptive related laser-evoked EEG responses has been well documented [21-27]. Based on the existing knowledge on relationship between laser-evoked EEG and pain, a method to predict the level of subjective pain perception using single-trial laser-evoked EEG responses [25] has been proved to be effective with good predictive accuracy.

Pain prediction using evoked EEG can be realized at two levels: within-individual pain prediction (the classifier/prediction model is trained on and applied to the same group of individuals) and cross-individual prediction (the classifier/prediction model is trained on a group of individuals but applied to a different group of individuals). Since within-individual pain prediction requires

real pain ratings for new individuals and is not applicable to people who are unable to reliably express their pain perception [121, 122], cross-individual pain prediction is more desired in clinical uses. However, it showed in earlier work [118] that cross-individual pain prediction has a significantly lower performance than within-individual prediction, mainly due to the inherent inter-individual variability in pain perception and neural responses. Therefore, incorporating individual factors that are particularly related to inter-individual variability of pain perception or pain-evoked neural activities into the cross-individual pain prediction model is crucial in pain assessment and pain therapy [155].

In an EEG-based pain prediction model that links EEG signals and subjective pain ratings, substantial inter-individual variability is involved in both EEG and ratings. However, it is difficult to improve the performance of cross-individual pain prediction by means of reducing the inter-individual variability of subjective pain ratings, because pain ratings for a new individual, as the unknown variables to be predicted, are not accessible. Therefore, we focused exclusively on decreasing the inter-individual variability of pain-related EEG responses and aim to develop a normalization strategy on pain-related EEG responses for a more accurate EEG-based cross-individual pain prediction.

In the present study, we hypothesize that an individual's spontaneous EEG activity can be used to normalize his/her pain-evoked EEG responses so as to improve the accuracy of cross-individual pain prediction. This hypothesis is induced by strong and consistent evidence showing that the magnitudes of a variety of pain-evoked EEG responses are highly correlated with that of

spontaneous EEG of the same individual. Actually, the magnitudes of both spontaneous and pain-evoked EEG activities are altered by the difference in individual-specific factors, such as cortical anatomy (e.g., the thickness of the skin and skull) and experimental conditions (e.g., electrode position and scale-electrode impedances) [156, 157]. Therefore, the magnitude of spontaneous EEG has the potential to serve as an individual scale to normalize the magnitude of pain-evoked EEG responses for a reduced inter-individual variability. The normalized magnitudes of pain-evoked EEG are used as features in the subsequent pain prediction. Next, a two-stage cross-individual pain prediction method is developed: a binary classifier to discriminate low-pain (Numerical Rating Scale, $NRS \leq 4$) and high-pain ($NRS > 4$) followed by a linear prediction model to predict the pain ratings (4-10) for high-pain trials only. The results showed that the proposed spontaneous EEG based normalization strategy can effectively decrease the inter-individual variability in the classifiers and prediction models, and consequently, can increase the accuracy of pain prediction, as compared with the prediction based on raw pain-evoked EEG responses.

5.2 Methodology

5.2.1 Experimental Design

Thirty-four healthy volunteers (17 females and 17males), aged 18 – 25 years (Mean $\pm$ SD: 21.6 ± 1.7) without chronic pain history, participated in the experiment. All volunteers gave their written informed consent and were paid

for their participation. The experiment procedures were approved by the local ethics committee. Before the experiment, they were familiarized with the experimental setup and task.

The nociceptive-specific radiant-heat stimuli were generated by an infrared neodymium yttrium aluminiumperovskite (Nd:YAP) laser (Electronical Engineering, Italy) with a wavelength of 1.34 μm . The laser beam was transmitted by focusing lenses through an optic fiber of which the diameter was approximately 7 mm ($\approx 38 \text{ mm}^2$). The laser pulses were directed at the medial side of the dorsum of left hand, between the first and third metacarpus. The duration of the laser pulse was fixed at 4 ms. After each stimulus, the laser beam was shifted by about 1 cm in a random direction to avoid nociceptor fatigue.

Participants were required to report the levels of their subjective pain perception induced by the laser stimuli ranging from 0 (no pain) to 10 (pain as bad as it could be) based on a numerical rating scale (NRS). Prior to EEG data collection, the highest energy of the laser stimulation was individually determined using the method of limits (from 1 J in step of 0.25 J) until a rating of 8 was reached. No withdrawal reflexes and motor contractions were observed until the stimulation intensity increased to 3.75 – 4.5 J. During the EEG data collection, 12 – 15 different levels of laser stimulation energies (from 1 J to 3.75 – 4.5 J, in step of 0.25 J) were adopted, and 10 laser pulses at each energy level, for a total of 120 – 150 pulses, were delivered in two blocks. Before each block, the surface temperatures of hand dorsum for each participant were measured using an infrared thermometer. The order of stimulus energies was pseudo-

randomized. The inter-stimulus interval (ISI) varied randomly between 10 to 15 s (uniformly distributed). An auditory tone delivered between 3 to 6 s after the laser pulse (uniformly distributed) prompting the participants to rate the level of the pain perception. The dataset with 12 – 15 levels of stimulation energy enables a more comprehensive and detailed investigation of the relationship between pain-evoked EEG responses and spontaneous EEG activities and the relationship between pain-evoked EEG responses and subjective pain ratings.

5.2.2 EEG Recording

Participants were seated in a comfortable chair in a temperature-controlled room, focusing their attention on the stimuli. The EEG data were collected using a 64-channel EEG cap with Ag-AgCl scalp electrodes placed according to the international 10 – 20 system (Brain Products GmbH, Munich, Germany), with the nose serving as the reference electrode. The impedances of all electrodes were kept lower than 10 k Ω .

5.2.3 EEG Data Analysis

EEG trials recorded from Cz channel (nose referenced) were used for the prediction of pain perception. The EEG data were filtered from 1 Hz to 30 Hz via a band-pass filter. EEG epochs from 0.5s pre-stimulus to 0.5s post-stimulus were extracted using a 1s analysis window, and then baseline-corrected using the pre-stimulus interval (-0.5 to 0s). Artifacts caused by eye movements or eye blinks were removed using independent component analysis. The above EEG data preprocessing were realized using EEGLAB [158] running in MATLAB

environment.

Each EEG trial consists of two segments: the pre-stimulus trial (-0.5s to 0s) is spontaneous EEG (sEEG) and the post-stimulus trial (0s to 0.5s) is dominated by pain-evoked EEG (pEEG). The magnitude of sEEG or pEEG trial is quantified by root mean square (RMS), as shown in Eq. (5-1)

$$RMS = \sqrt{\frac{1}{N} \sum_{k=1}^N x_k^2} \quad (5-1)$$

where x_k is the k -th sample of the trial, and N is the number of data points. The RMS of sEEG and pEEG are denoted as RMS_S and RMS_P and will be used as features in subsequent investigation of the relationship between pain and EEG and in pain prediction.

5.2.4 Relationship Between sEEG and pEEG

To test whether sEEG can serve as a baseline to normalize pEEG for a smaller inter-individual variability, we examined the relationship between RMS_S and RMS_P . We assume that an individual's RMS_S and RMS_P (both of which are averaged across all trials at each pain intensity level) are normally distributed, and calculate the mean and standard deviation (SD) of RMS_S and RMS_P across all trials at each pain intensity level. Then, cross-individual correlation between these mean and SD values were estimated.

5.2.5 Relationship Between Pain Perception and pEEG

In our experiments, participants were asked to report the level of their pain perception with 4 as the pinprick pain threshold (i.e. NRS > 4 refers to feeling of

pinprick pain). Thus, NRS = 4 serves as a threshold to differentiate low-pain and high-pain (i.e., low-pain: $NRS \leq 4$, high-pain: $NRS > 4$). To investigate the relationship between the rating of perceived pain and evoked EEG responses, RMS_p , which was averaged across trials with each identical pain level for each individual, was fitted using two models: a global linear model and a two-piecewise linear model (two segments with $NRS = 4$ as the break point). The global linear model is based on the assumption that magnitude of pEEG is linearly increased with the pain rating, while the piecewise linear model assumes that the relationship between pain rating and pEEG is different for low-pain ($NRS \leq 4$) and high-pain ($NRS > 4$). The performance of the two models are quantified with the mean square error (MSE) and compared across individuals using a paired sample t-test.

5.2.6 Feature Normalization Based on sEEG

There are different normalization methods available, and here we normalized the magnitude of a pEEG trial as the z-score of the population defined by sEEG trials. For each individual, the magnitude of the i -th pEEG trial, $RMS_p(i)$, was normalized by RMS_s of all sEEG trials as

$$nRMS_p(i) = \frac{RMS_p(i) - \mu(RMS_s)}{\sigma(RMS_s)} \quad (5-2)$$

where $nRMS_p(i)$ is the normalized magnitude of the i -th pEEG trial, μ and σ are respectively the mean and the SD of RMS_s of all trials of this individual.

To examine whether the inter-individual variability of pEEG magnitudes was decreased by the sEEG-based normalization, we performed an ANOVA F-test on

RMS_p and $nRMS_p$. More precisely, at each pain level, an ANOVA F-test was performed on RMS_p or $nRMS_p$ of all trials with this pain level across all individuals to check whether the means of RMS_p or $nRMS_p$ of this group of individuals are the same, and the resultant F-statistics denote the inter-individual variability relative to the within-individual variability of the variable under test. Then, we compared the F-statistics between RMS_p and $nRMS_p$ at each pain level, and it is expected that $nRMS_p$ has a smaller F-statistic than RMS_p .

We next investigate whether the proposed sEEG-based normalization strategy can reduce the cross-individual variability in the relationship between intensity of pain perception and the pEEG magnitude. Firstly, for each individual, we calculated an optimal threshold of RMS_p or $nRMS_p$ that can best classify (with the highest accuracy) the individual's trials into low-pain ($NRS \leq 4$) and high-pain ($NRS > 4$). The inter-individual variability of the binary classification thresholds of RMS_p and $nRMS_p$ were measured by variance. If the variance of the threshold obtained from $nRMS_p$ is smaller than that from RMS_p , the effectiveness of sEEG-based normalization in reducing the individual difference in the binary classification can be validated. A two-sample F-test was also conducted to check whether the thresholds of RMS_p and $nRMS_p$ have the same variance. Secondly, we consider the relationship between pain ratings and RMS_p (or $nRMS_p$) of high-pain ($NRS > 4$) trials to be a linear model specific to each individual, then the inter-individual variability in this relationship is indicated by the cross-individual variance of slope and intercept of the linear

model. So, slopes and intercepts of all individuals were calculated using two sets of features (RMS_p and $nRMS_p$), and their inter-individual variability were compared. If the cross-individual variance of slopes or intercepts obtained using $nRMS_p$ is smaller than that obtained using RMS_p , the effectiveness of sEEG-based normalization in reducing the individual variability in the continuous prediction model can be validated. Similarly, a two-sample F-test was also conducted to check whether the slopes or intercepts of RMS_p and $nRMS_p$ have the same variance.

5.2.7 Binary Classification and Continuous Prediction

A linear discriminant analysis (LDA) classifier was adopted to classify low-pain and high-pain trials using leave-one-individual-out cross validation. The classifier was first trained with RMS_p or $nRMS_p$ of training trials, which divided into two groups (low-pain: $NRS \leq 4$, and high-pain: $NRS > 4$), and then applied to the test trials to predict labels (low-pain vs. high-pain) from the corresponding RMS_p or $nRMS_p$. The classification performance was evaluated by accuracy, and the accuracies obtained from classification using RMS_p and $nRMS_p$ were compared using paired sample t-test.

After binary classification, only high-pain trials are involved in continuous pain prediction, because there is no significant correlation between pain ratings and pEEG of low-pain trials. To prove that the sEEG-based normalization is effective for continuous pain prediction regardless of the results of the preceding binary classification, we performed continuous pain prediction for trials predicted as

high-pain from binary classification as well as for real high-pain trials (NRS > 4).

Relationship between single-trial RMS_p (or $nRMS_p$) and the intensity of pain perception was modeled by linear regression. For the i -th pEEG trial, the pain rating R_i can be estimated as

$$R_i = \alpha \cdot RMS_p(i) + c \text{ or } R_i = \alpha \cdot nRMS_p(i) + c \quad (5-3)$$

where α and c are slope and intercept of the linear regression model. The model of (5-3) was trained and tested using leave-one-out cross validation.

The prediction performance of the linear regression model was evaluated by mean absolute error (MAE), which is defined as Eq. (5-4)

$$MAE = \frac{1}{N} \sum_{n=1}^N |R_i - \hat{R}_i| \quad (5-4)$$

where N is the total number of test trials, R_i and $\hat{R}_i$ are the real and predicted rating value for the i -th trial, respectively. MAE provides a straightforward measure on how precisely the generated linear regression model can represent the relationship between pain ratings and pEEG magnitudes. The MAE values obtained from prediction using RMS_p and $nRMS_p$ were compared using the paired sample t-test.

5.3 Results and Discussion

5.3.1 Relationship Between sEEG and pEEG

For each participant, the mean and SD of RMS_s and RMS_p at each pain intensity level were calculated. Since NRS > 8 was not available for some participants, we use a combined level of “NRS > 8” to denote all trials with an NRS greater than

8. It can be clearly seen from Fig. 5-1 and Table 5-1 that, a significant correlation ($p \leq 0.007$) between the mean values of RMS_S and RMS_P was obtained at each intensity level of pain perception. In addition, a significant correlation between the SD values of RMS_S and RMS_P was also obtained ($p \leq 0.02$) for overall intensity level of low-pain ($NRS \leq 4$) and high-pain ($NRS > 4$), though some individual intensity level is not significant (such as intensity level at 2 – 3, 4 – 5, 6 – 7, and 8 – 10). To conclude, the distributions of RMS_S and RMS_P are highly correlated, which verifies our hypothesis that magnitudes of pEEG are highly correlated with the magnitude of sEEG. This observation supports the idea that RMS_S could serve as an individual scale to normalize his/her RMS_P to reduce inter-individual variability of pain-related features in pain classification and prediction models.

Figure 5-1. Correlation between the mean and SD of RMS_S and RMS_P at levels of (A) $NRS \leq 4$, and (B) $NRS > 4$. Red dots represent the mean or SD of RMS_S and RMS_P , which are

averaged across all trials at each pain intensity level for each participant. Gray lines represent the best linear fit (Adopted from Fig. 1 in [119] with permission from the publisher).

Table 5-1. Correlation between the mean and SD of RMS_S and RMS_P at each pain level

NRS		(0, 4]	(0, 1]	(1, 2]	(2, 3]	(3, 4]	
Mean	R	0.89	0.72	0.93	0.73	0.67	
	P-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
SD	R	0.66	0.61	0.47	0.17	0.54	
	P-value	< 0.001	< 0.001	0.013	0.335	0.001	
NRS		(4, 10]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 10]
Mean	R	0.55	0.45	0.53	0.57	0.63	0.56
	P-value	< 0.001	0.007	0.001	< 0.001	< 0.001	0.003
SD	R	0.52	0.30	0.34	0.07	0.40	0.19
	P-value	0.002	0.082	0.049	0.718	0.025	0.346

5.3.2 Relationship Between Pain Perception and pEEG

Fig. 5-2 (A) shows the relationship between pain rating and the magnitude of pEEG (mean $\pm$ SD) of one participant. Overall, pain rating and RMS_P are positively related, but RMS_P does not increase significantly when the subjective pain ratings is smaller than or equal to 4 (referred to as “low-pain”); when the subjective pain rating is greater than 4 (referred to as “high-pain”), RMS_P is linearly increased with pain ratings. MSE of the global linear model (red line) or the two-piecewise linear model (blue line) was adopted to measure the accuracy of fitting, as shown in Fig. 5-2 (B). It can be seen from the group-level results in Fig. 5-2 (B) that, the fitting error of the piecewise linear model is significantly smaller than that of the global linear model. Therefore, the piecewise linear model can better describe the relationship between pain perception and RMS_P . The nonlinear relationship motivates us to develop the two-stage pain prediction (i.e., to classify low- and high-pain first, then to predict the pain rating

for high-pain trials only).

Figure 5-2. (A) Relationship between pain ratings and RMS_P (from one participant). Colored dots represent mean $\pm$ SD of RMS_P averaged across trials at each given level of pain perception. The red line represents the fitted global linear model, while the blue lines represent the fitted two-piecewise linear model. (B) Comparison of MSE (mean $\pm$ SD) of all participants between two fitting models (Adopted from Fig. 2 in [119] with permission from the publisher).

5.3.3 Feature Normalization Based on sEEG

We first confirmed that the magnitudes of sEEG trials and pEEG trials of each individual approximately follow a normal distribution ($p < 0.0001$ for all individuals, one-sample Kolmogorov–Smirnov test). Therefore, an individual's sEEG trials could form a distribution for normalizing pEEG trials into z-scores. Table 5-2 shows that, at each pain level, the F-statistic obtained from the ANOVA F-test on $nRMS_P$ is lower than that obtained from RMS_P , which proves that the sEEG-based normalization method can effectively reduce the inter-individual variability of pEEG trials.

Fig. 5-3 shows the optimal binary classification thresholds and slopes/intercepts of linear regression models for high-pain trials, which were obtained from

available trials and ratings of all individuals. We can clearly see that, after sEEG-based normalization, the variance of all above three parameters were remarkably decreased, which illustrates that the proposed sEEG-based normalization method can effectively reduce the inter-individual variability in classification and prediction models.

Table 5-2. Comparison of F-statistics obtained from the ANOVA F-test on RMS_p and $nRMS_p$

NRS	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 10]
RMS_p	5.28	6.48	5.80	10.56	17.02	13.50	17.28	10.42	18.72
$nRMS_p$	1.18	1.48	1.58	4.80	11.62	11.33	13.79	8.17	9.31

Figure 5-3. Effects of sEEG normalization on inter-individual variability of (A) binary classification thresholds, (B) slopes of linear regression models for high-pain trials, (C) intercepts of linear regression models for high-pain trials. Noted that mean values were removed from these parameters for illustration (Adopted from Fig. 3 in [119] with permission from the publisher).

Table 5-3 further shows that, cross-individual variances of all three classifier/regression model parameters were decreased after sEEG-based normalization. The two-sample F-test also demonstrates that the variances of all three classifier/regression model parameters are significantly different between using RMS_p and using $nRMS_p$ as features ($p < 0.0001$ for all).

Table 5-3. Comparison of binary classifier thresholds and regression model parameters between using RMS_p and using $nRMS_p$

Model Parameters		RMS_p	$nRMS_p$
Binary classification thresholds	Cross-individual variance	4.28	1.80
	P-value (F-test for equal variance)	< 0.001	
Slopes of linear regression models	Cross-individual variance	1.34	0.52
	P-value (F-test for equal variance)	< 0.001	
Intercepts of linear regression models	Cross-individual variance	7.53	3.25
	P-value (F-test for equal variance)	< 0.001	

5.3.4 Pain Prediction

The mean and SD of accuracy for binary pain classification (low-pain vs. high-pain) using RMS_p (i.e., pEEG features) and using $nRMS_p$ (i.e., sEEG-normalized pEEG features) are summarized in Table 5-4. Results show that $nRMS_p$ can improve the classification accuracy, though the performance improvement is only close to significant ($p = 0.092$).

Table 5-4. Accuracy of binary classification and prediction error (MAE) of continuous prediction

	RMS_p	$nRMS_p$	P-value (paired t-test)
Accuracy of binary classification (%)	68.95 ± 12.91	70.36 ± 14.18	0.092
MAE of continuous prediction (on predicted high-pain trials)	1.84 ± 0.60	1.63 ± 0.45	0.002
MAE of continuous prediction (on real high-pain trials)	1.24 ± 0.28	1.17 ± 0.28	0.003

The mean and SD of MAE for continuous pain prediction using RMS_p (i.e., pEEG responses) and using $nRMS_p$ (i.e., sEEG-normalized pEEG responses) for trials predicted as high-pain from binary classification and for real high-pain trials ($NRS > 4$) are summarized in Table 5-4. Results indicate that the proposed sEEG-

based normalization method can significantly improve the prediction accuracy for both predicted high-pain trials ($p = 0.002$) and real high-pain trials ($p = 0.003$) in continuous pain prediction.

5.3.5 Discussion

In this study, we proposed a strategy to normalize pain-evoked EEG responses using spontaneous EEG to improve the performance of EEG-based cross-subject pain prediction. Pain-related EEG responses have been used to predict the level of subjective pain, but the large inter-individual variability seriously degrades the performance of cross-individual pain prediction. In this work, we began with performing a comprehensive and detailed investigation of the relationship between pEEG responses and sEEG activities as well as the relationship between subjective pain ratings and pEEG responses. Our results revealed a strong inter-individual correlation between the magnitude of pEEG and sEEG. Besides, our results also confirmed a nonlinear relationship between pEEG and subjective pain ratings. Based on above observations, we proposed a new two-stage approach for pain prediction: (1) a binary classification to differentiate low-pain and high-pain trials; (2) a continuous regression to predict pain ratings of high-pain trials. In both steps, the normalization strategy based on sEEG was used to reduce the inter-individual variability in the magnitude of pEEG, so that a higher classification accuracy and a lower prediction error were achieved. The new sEEG-based normalization strategy has the potential to contribute to an applicable and reliable tool for pain assessment.

Relationship Between sEEG and pEEG

An individual's spontaneous EEG has been shown to be related to his/her genetic code, implying its uniqueness [159-164]. A strong inter-individual correlation between the magnitude of sEEG and pEEG was also obtained from our database. A potential interpretation for this phenomenon may be due to the skull thickness, the orientation of the grey matter and so forth. These anatomical factors are specific to each person and will remain relatively stable for adults. Experimental conditions, such as electrode position and scale-electrode impedances, could also contribute to the phenomenon, because they may influence the magnitudes of both pEEG and sEEG.

Relationship Between Pain Rating and Pain-Evoked EEG Responses

Numerous previous studies have shown that the perceived pain intensity is significantly correlated with the amplitude of a number of evoked EEG responses [25, 26]. In most of these works, the level of pain perception was assumed or found to be linearly correlated with the evoked EEG responses, but such a linear relationship has been challenged by growing evidence showing the nonlinearity between pain level and the neural responses. The assumption and observation of the linear relationship may due to the limited range of painful stimulus intensities used in most of pain experiments, which further limited the range of perceived pain intensity. To solve this problem, our experiment was designed to deliver a large number of laser pulses (> 100) with a wide energy range (from 1 J to 3.75 – 4.5 J; 12 – 15 levels) to each participant. The result that the fitting error of a two-piecewise linear model (with a break point of NRS = 4) was

significantly smaller than that of a global linear model indicated a nonlinear relationship between pain level and the evoked EEG responses.

Feature Normalization Based on sEEG

To normalize the magnitudes of pEEG trials of one individual, we proposed to estimate their z-scores in the population defined by sEEG trials of this individual. Although other normalization methods exist, the proposed z-scores can achieve better prediction results than other normalization methods, such as dividing RMS_p by the mean of sEEG or subtracting the mean of sEEG from RMS_p (results are not shown here). Besides its good performance, the proposed z-score normalization also reflects certain physiological meanings. It has been revealed that the variability of spontaneous neural activity can reflect the “dynamic range” of possible neural responses to incoming stimuli and can provide a powerful and accessible measure for understanding various individual difference variables [165-173]. Therefore, the distribution defined by the magnitudes of sEEG is indicative of the possible range of magnitudes of evoked pEEG and it can be used as a baseline distribution to normalize pEEG magnitudes to z-scores.

Applicability of the sEEG-Normalization Based Pain Prediction

Predictive power of pEEG responses for decoding the intensity of subjective pain perception has been well documented in previous studies [23-27], which further led to several cross-individual pain prediction methods, which do not need any subjective pain rating for new individuals and thus more promising for clinical uses. However, the accuracy of cross-individual pain prediction is still not satisfactory because of the inherent inter-individual variability in either pain

evoked responses or pain ratings. A practical solution to this problem is to incorporate individual traits that are related to inter-individual variability into the pain prediction model [155]. In an earlier study [25], the features of single-trial evoked EEG were normalized with regard to the mean and SD of the individual's evoked EEG features, and single-trial ratings of pain perception were rescaled into the range from 0 to 10. Although above normalization on both evoked EEG features and pain ratings can significantly increase the prediction accuracy, its drawback was obvious. First, the normalization was based on the distribution of evoked EEG features, which can only be obtained from a large number of painful stimuli and may not be accepted by participants. Second, it still needs subjective pain rating from a new individual, which is not suitable for participants with communication impairments. As compared with the conventional normalization strategy [25], the proposed method has two main advantages: first, it will not introduce any pain experience to a new participant because the normalization is based on spontaneous EEG; second, it can well deal with the difficult situation that no reliable pain rating is available because no subjective rating is needed. Therefore, the proposed sEEG-based normalization method is more practical and feasible for clinical research and applications.

Sex Differences in Prediction

We also investigated the sex differences in pain prediction. Firstly, we examined the difference of pain ratings with same stimulus energy (14 levels from 1 – 4.25 J, in step of 0.25 J) between males and females, as shown in Fig. 5-4 (A). It can be seen from Table 5-5 that significant sex-dependent difference was observed

at most of the stimulus levels (except levels 3, 12, and 13). Next, a similar analysis was conducted to compare RMS_p values with same rating values (NRS from 0 – 8) between males and females, as shown in Fig. 5-4 (B). Statistical analysis in Table 5-6 illustrated significant sex-dependent difference at every rating value. Please note that all above differences were obtained using two sample t-test.

Figure 5-4. (A) Difference in pain ratings with the same stimulus energy between males and females. (B) Difference in RMS_p with same rating value between males and females (Modified from Fig. S1 in [119] with permission from the publisher).

Table 5-5. P-values of comparisons of pain ratings with the same stimulus energy between males and females

Level of Stimulus Energy	1	2	3	4	5	6	7
P-value	2.3e-5	0.01	0.15	0.05	1.2e-4	6.4e-4	2.7e-4
Level of Stimulus Energy	8	9	10	11	12	13	14
P-value	0.03	8.2e-5	6.0e-7	2.9e-5	0.11	0.84	0.04

Table 5-6. P-values of comparisons of RMS_p with the same rating value between males and females

NRS	0	1	2	3	4	5	6	7	8
P-value	3.0e-6	5.7e-6	5.7e-5	6.8e-8	1.4e-9	2.5e-7	6.9e-8	<1e-9	<1e-9

The results showed significant sex-dependent difference in pain perception as

well as in pain-evoked EEG responses, and the findings are consistent with the literatures [174-177].

5.4 Conclusions

An effective physiological pain assessment method that complements the gold standard of self-report is highly desired in pain clinical research and practice. Recent studies have shown that pain-evoked EEG responses could be used as a readout of perceived pain intensity. Existing EEG-based pain assessment is normally achieved by cross-individual prediction (i.e., to train a prediction model from a group of individuals and to apply the model on a new individual), so its performance is seriously hampered by the substantial inter-individual variability in pain-evoked EEG responses. In this study, to reduce the inter-individual variability in pain-evoked EEG and to improve the accuracy of cross-individual pain prediction, we examined the relationship between pain-evoked EEG, spontaneous EEG, and pain perception on a pain EEG dataset, where a large number of laser pulses (> 100) with a wide energy range were delivered. Motivated by our finding that an individual's pain-evoked EEG responses is significantly correlated with his/her spontaneous EEG in terms of magnitude, we proposed a normalization strategy for pain-evoked EEG responses using one's spontaneous EEG to reduce the inter-individual variability. In addition, a nonlinear relationship between the level of pain perception and pain-evoked EEG responses was obtained, which inspired us to further develop a new two-stage pain prediction strategy, a binary classification of low-pain and high-pain trials followed by a continuous prediction for high-pain trials only, both of which

used spontaneous-EEG-normalized magnitudes of evoked EEG responses as features. Results show that the proposed normalization strategy can effectively reduce the inter-individual variability in pain-evoked responses, and the two-stage pain prediction method can lead to a higher prediction accuracy.

Chapter 6 Conclusions and Future Directions

6.1 Conclusions

This dissertation develops several signal processing techniques for Raman spectroscopy and biomedical signal analysis.

First, we propose a denoising method for Raman spectroscopy based on Wiener estimation. Wiener estimation using a numerical calibration dataset (NCD-WE) is developed by considering the common characteristics of Raman spectra such that the method does not require any experimental measurement. This method is validated by the Raman measurement of three types of samples, including a phantom sample, human fingernail and leukemia cells, with various noise levels. The results indicate that the NCD-WE method is able to recover the spectral shape and the main peaks of the Raman spectrum accurately even when the peaks are overwhelmed by noise in the original spectrum. Comparison of denoising performance with the other two widely used methods, moving-average (MA) filtering and Savitzky-Golay (SG) filtering, suggests that NCD-WE method provides comparable or even better denoising performance especially in low SNR cases. Another advantage worth highlighting about the proposed method is that it demands fewer experiences from users to process spectra with unknown noise levels because it is less sensitive to the choices of parameters.

Second, we investigate the possibility of using Raman spectroscopy for the quantitative analysis in a context that the target species is mixed with multiple

non-target species and the exact spectral features of the non-target species are unavailable. A method involving orthogonal projection and biochemical component analysis, i.e. OPBCA, is proposed, which is inspired by the previous studies that the Raman spectrum of biological cells can be approximated by the linear summation of spectra contributed by the composing basic biochemical components. We extend this observation to the estimation of the unknown non-target species, which is then used in orthogonal projection to find the clean contribution of the target species for quantitative analysis. The proposed method is validated on both Raman phantoms and spectra synthesized from the experimental measurement of doxorubicin (DOX) and leukemia cells. Results demonstrate the effectiveness of the OPBCA method for the quantification of the target species accompanying the highly accurate estimation of the non-target species in both cases. Furthermore, the proposed method does not require a training step thus is advantageous when the data for non-target species are very limited.

Thirdly, we explore the possible reasons of performance degradation of classification/prediction methods at the cross-individual level in the application of EEG-based pain assessment. A comprehensive investigation of the relationship between pain-evoked EEG (pEEG) responses and spontaneous EEG (sEEG) activities as well as the relationship between subjective pain ratings and pEEG responses is carried out. Results reveal that an individual's pain-evoked EEG responses is significantly correlated with his/her spontaneous EEG in terms of magnitude. Making use of this observation, we propose a normalization

strategy to reduce the inter-individual variability in pain-evoked EEG and to improve the accuracy of cross-individual pain prediction. In addition, a nonlinear relation is found between the level of pain perception and pain-evoked EEG responses. Inspired by this observation, we further develop a new two-stage pain prediction strategy, a binary classification of low-pain and high-pain trials followed by continuous prediction for high-pain trials only, in both of which stages, the spontaneous-EEG based normalization is applied. Results demonstrate that the proposed normalization strategy can effectively reduce the inter-individual variability in pain-evoked responses, and the two-stage pain prediction method lead to a higher prediction accuracy of pain perception. Beyond, our approach may help establish a guideline to address the challenge of cross-Individual classification/prediction.

6.2 Future Work

In this dissertation, we mainly focus on developing signal processing techniques and the advanced algorithms on three topics. The future direction of each topic is described as follows:

Processing of Raman Spectroscopic Data

Due to the inherently weak signal, Raman scattering often leads to a low signal-to-noise ratio to the Raman spectrum, which limits its capabilities under certain conditions. Nevertheless, many techniques have been introduced to overcome this issues, such as various smoothing methods [2-4] and wavelet transformation [6-9]. These efforts have advanced Raman spectroscopy to be

more compatible with practical application and clinical use [178]. However, most of the existing algorithms require significant experiences from users to ensure good and consistent performance. In this study, we attempt to develop an optimal and user-friendly denoising method for dealing with Raman spectroscopic data of which the performance is less sensitive to parameter choices. Results confirm that decent performance is achieved as long as the parameter values fall in the designated ranges. In our strategy, a numerical calibration dataset is created, and the Wiener matrices are obtained from a series of calibration datasets each with a single peak in spectra after the peaks in the column vectors of the Wiener matrices are spectrally shifted along the spectral dimension properly. This design simplifies the procedure and expand the universality of the method to a great extent. However, performing Wiener estimation at every data point is time consuming.

In the future work, we will focus on further improving the speed and efficiency of the method. The computational efficiency of the NCD-WE method could be improved by performing the reconstruction using the Wiener matrix (after spectral shifting) with an adaptive shifting increment according to the bandwidth of Raman peaks. In addition, we believe it is possible to improve the proposed algorithm performance using a multi-scale approach, which is similar to wavelet reconstruction. In this approach, multiple calibration datasets will be created each with a different peak width in the initial single-peak spectrum. The reconstruction will be performed in multiple steps and the calibration datasets associated with progressively smaller peak widths will be sequentially used in

these steps to reconstruct the details of the target spectrum in a series of spectral scales. This approach is expected to yield more accurate details at the cost of longer computation. Another aspect of potential improvement is to incorporate model-driven machine learning techniques into signal processing to enhance the efficiency of signal processing. The recent advances in artificial intelligence (e.g. deep learning) have made it feasible for the computer to be trained to perform signal processing in an efficient and intelligent manner. It is desirable to develop automatic algorithms to effectively and efficiently process data on behalf of domain experts in the near future.

Quantification Using Raman Spectroscopy

Raman spectroscopy has been demonstrated with great potentials in quantitative analysis in many areas with the help of various multivariate statistical analysis techniques. Despite numerous previous studies, challenges remain when the target being studied is in a complex environment, where the contributions from non-target species are present.

Many advanced algorithms have been developed to solve the mixture analysis problem [179-181]. Among them, Multivariate Curve Resolution (MCR) [182] intends to recover the concentrations and response profiles of the components in unsolved mixtures. It has been shown as an effective tool to describe the contribution of each pure component in a multiple-components system. It is expected that the use of multivariate curve resolution alternating least squares (MCR-ALS) method applied to Raman measurement will allow the recovery of Raman spectra and pure spectral concentration profiles of different species

involved in the sample. In our future work, we are aiming to develop a tailored MCR-ALS method by taking into account the physically/chemically meaningful constraints, allowing to deal with different types of deviations from ideal situations, e.g. change of spectral features from samples to samples because of the interactions among fluorophores. We also want to explore more potential applications of multivariate analysis-assisted Raman spectroscopic technology in practice, for example, in the quantification of bacteria in environmental science research, and monitoring drug release in biomedicine and clinical diagnosis.

Cross-individual Pain Prediction

In practical applications, pain prediction is highly desired to be performed on new subjects, for example, a participant in a drug trial or a patient after hospital admission [183]. However, it is common that when single-subject decoding is accurate, decoding across subjects is inaccurate on trials from new subjects. The cause of this issue is the variability across subjects. Such variability reflects not only in the aspect of pain-related EEG responses, but also in the subjective pain perception. In this study, the proposed normalization strategy focused solely on the features of pain-evoked EEG responses, simply because real values of pain perception are generally considered to be unknown in clinical scenarios. However, not only EEG responses but also the pain ratings are characterized by tremendous inter-individual variability. Each individual perceives pain perception differently in response to the same painful stimulus. For example, we have found pronounced sex-dependent difference in pain perception as well as

in pain-evoked EEG responses. Taking into account gender difference (such as using sex as a predictor) may lead to more accurate pain prediction. Mechanisms contributing to inter-individual differences in pain sensitivity include genetic, environmental, psychological, and cognitive factors [32, 122, 184], and it may be caused at any stage in pain processing from the skin to the brain. Highly sensitive individuals may activate stronger neural responses and/or pain experience than insensitive individuals [184, 185]. Variations in pain sensitivity is an important issue worthy of further investigation, because understanding the contributing factors of pain sensitivity will help greatly in developing a more accurate and practical method for diagnosis of pain [32, 186]. Our future study will aim to address the above difficult problems, such as how to normalize pain ratings and pain sensitivity and how to incorporate personal traits and environmental factors in the prediction model, to develop a more accurate and practical EEG-based prediction assessment method.

In sum, it is our hope that our research will contribute to the translation of technical advances in biomedical optics and bio-photonics to solutions for practical problems in health care and medical diagnostics.

Author Publications

1. **Y.R. Bai**, Q. Liu, "Denoising Raman spectra by Wiener estimation with a numerical calibration dataset," Biomedical Optics Express 11(1): p. 200-214, 2020.
2. **Y.R. Bai**, G. Huang, Y.S. Huang, et al., "Normalization of pain-evoked neural responses using spontaneous EEG improves the performance of EEG-based cross-individual pain prediction", Frontiers in computational Neuroscience, p. 31, 2016.
3. **Y.R. Bai**, K.R. Chen, S. Chen, et al., "A Raman spectroscopic method for target species quantification based on orthogonal projection and biochemical component analysis", Biomedical Express (prepare the manuscript).
4. **Y.R. Bai**, K.R. Chen, A. Mishra, et al., "Feasibility of quantitatively diagnosing cornea infection using Raman spectroscopy" (Oral Presentation). SPIE BIOS. Vol. 10054. 2017: SPIE.
5. **Y.R. Bai**, Y. Hu, and Z.G. Zhang, "Spontaneous EEG-based normalization of pain-evoked neural responses: Effect on improving the accuracy of pain prediction", 2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2016.

References

1. Jit, M., *BIOMEDICAL SIGNAL ANALYSIS*. 2003, McGraw-Hill Education: New York.
2. Tarcea, N. and J. Popp, *Raman data analysis*, in *Raman spectroscopy applied to Earth sciences and cultural heritage*, J. Dubessy, M.C. Caumon, and F. Rull, Editors. 2012, Mineralogical Society of Great Britain and Ireland. p. 193-226.
3. Kwiatkowski, A., et al., *Algorithms of Chemicals Detection Using Raman Spectra*. 2010. **17**(4): p. 549.
4. Barton, S.J., T.E. Ward, and Bryan M. Hennelly, *Algorithm for optimal denoising of Raman spectra*. *Analytical Methods*, 2018. **10**(30): p. 3759-3769.
5. Lasch, P., *Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging*. *Chemometrics and Intelligent Laboratory Systems*, 2012. **117**: p. 100-114.
6. Chen, H., et al., *An adaptive denoising method for Raman spectroscopy based on lifting wavelet transform*. *Journal of Raman Spectroscopy*, 2018. **49**(9): p. 1529-1539.
7. Ehrentreich, F. and L. Sümmchen, *Spike Removal and Denoising of Raman Spectra by Wavelet Transform Methods*. *Analytical Chemistry*, 2001. **73**(17): p. 4364-4373.
8. Ramos, P.M. and I. Ruisánchez, *Noise and background removal in Raman spectra of ancient pigments using wavelet transform*. *Journal of Raman Spectroscopy*, 2018. **49**(9): p. 1529-1539.

- Spectroscopy, 2005. **36**(9): p. 848-856.
9. Villanueva-Luna, A.E., et al., *Fluorescence and noise subtraction from Raman spectra by using wavelets*. Opt. Mem. Neural Netw., 2010. **19**(4): p. 310-317.
 10. El-Abassy, R.M., et al., *Fast determination of milk fat content using Raman spectroscopy*. Vibrational Spectroscopy, 2011. **56**(1): p. 3-8.
 11. Nedeljković, A., et al., *Raman Spectroscopy as a Rapid Tool for Quantitative Analysis of Butter Adulterated with Margarine*. Food Analytical Methods, 2016. **9**(5): p. 1315-1320.
 12. Strachan, C.J., et al., *Raman spectroscopy for quantitative analysis of pharmaceutical solids*. J Pharm Pharmacol, 2007. **59**(2): p. 179-92.
 13. Johansson, J., S. Pettersson, and S. Folestad, *Characterization of different laser irradiation methods for quantitative Raman tablet assessment*. Journal of Pharmaceutical and Biomedical Analysis, 2005. **39**(3-4): p. 510-516.
 14. Chuchuen, O., et al., *Quantitative analysis of microbicide concentrations in fluids, gels and tissues using confocal Raman spectroscopy*. PLoS One, 2013. **8**(12): p. e85124.
 15. Huang, J., et al., *Tracking the intracellular drug release from graphene oxide using surface-enhanced Raman spectroscopy*. Nanoscale, 2013. **5**(21): p. 10591-10598.
 16. Zhu, Q., R.G. Quivey, Jr., and A.J. Berger, *Raman spectroscopic measurement of relative concentrations in mixtures of oral bacteria*. Appl Spectrosc, 2007. **61**(11): p. 1233-7.
 17. Zhu, Q., R.G. Quivey, and A.J. Berger, *Measurement of bacterial*

- concentration fractions in polymicrobial mixtures by Raman microspectroscopy*. Journal of Biomedical Optics, 2004. **9**(6): p. 1182-1186.
18. Bromm, B. and R. Treede, *Nerve fibre discharges, cerebral potentials and sensations induced by CO2 laser stimulation*. Human neurobiology, 1983. **3**(1): p. 33-40.
 19. Iannetti, G., et al., *Evidence of a specific spinal pathway for the sense of warmth in humans*. Journal of neurophysiology, 2003. **89**(1): p. 562-570.
 20. Treede, R.-D., J. Lorenz, and U. Baumgärtner, *Clinical usefulness of laser-evoked potentials*. Neurophysiologie Clinique/Clinical Neurophysiology, 2003. **33**(6): p. 303-314.
 21. Arendt-Nielsen, L., *Characteristics, detection, and modulation of laser-evoked vertex potentials*. 1994: Wiley Online Library.
 22. Beydoun, A., et al., *Variability of laser-evoked potentials: attention, arousal and lateralized differences*. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 1993. **88**(3): p. 173-181.
 23. Bromm, B. and R. Treede, *Laser-evoked cerebral potentials in the assessment of cutaneous pain sensitivity in normal subjects and patients*. Revue neurologique, 1990. **147**(10): p. 625-643.
 24. Garc-Larrea, L., et al., *Association and dissociation between laser-evoked potentials and pain perception*. Neuroreport, 1997. **8**(17): p. 3785-3789.
 25. Huang, G., et al. *Single-trial laser-evoked potentials feature extraction for prediction of pain perception*. in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. 2013. IEEE.

-
26. Iannetti, G., et al., *Simultaneous recording of laser-evoked brain potentials and continuous, high-field functional magnetic resonance imaging in humans*. Neuroimage, 2005. **28**(3): p. 708-719.
 27. Kakigi, R., H. Shibasaki, and A. Ikeda, *Pain-related somatosensory evoked potentials following CO₂ laser stimulation in man*. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 1989. **74**(2): p. 139-146.
 28. Schulz, E., et al., *Decoding an individual's sensitivity to pain from the multivariate analysis of EEG data*. Cerebral Cortex, 2011: p. bhr186.
 29. Legrain, V., et al., *Attentional modulation of the nociceptive processing into the human brain: selective spatial attention, probability of stimulus occurrence, and target detection effects on laser evoked potentials*. Pain, 2002. **99**(1): p. 21-39.
 30. Lorenz, J. and L. Garcia-Larrea, *Contribution of attentional and cognitive factors to laser evoked brain potentials*. Neurophysiologie Clinique/Clinical Neurophysiology, 2003. **33**(6): p. 293-301.
 31. Ohara, S., et al., *Cutaneous painful laser stimuli evoke responses recorded directly from primary somatosensory cortex in awake humans*. Journal of neurophysiology, 2004. **91**(6): p. 2734-2746.
 32. Nielsen, C.S., R. Staud, and D.D. Price, *Individual differences in pain sensitivity: measurement, causation, and consequences*. The journal of pain, 2009. **10**(3): p. 231-237.
 33. Gautam, R., et al., *Review of multidimensional data processing approaches*

- for Raman and infrared spectroscopy*. Epj Techniques and Instrumentation, 2015. **2**(1).
34. Roussel, S., et al., *Multivariate Data Analysis (Chemometrics)*, in *Process Analytical Technology for the Food Industry*, C.P. O'Donnell, C. Fagan, and P.J. Cullen, Editors. 2014, Springer New York: New York, NY. p. 7-59.
35. Ryder, A.G., G.M. O'Connor, and T.J. Glynn, *Quantitative analysis of cocaine in solid mixtures using Raman spectroscopy and chemometric methods*. Journal of Raman Spectroscopy, 2000. **31**(3): p. 221-227.
36. Goicoechea, H.C. and A.C. Olivieri, *Enhanced Synchronous Spectrofluorometric Determination of Tetracycline in Blood Serum by Chemometric Analysis. Comparison of Partial Least-Squares and Hybrid Linear Analysis Calibrations*. Analytical Chemistry, 1999. **71**(19): p. 4361-4368.
37. Chang, H.-H. and J.M. Moura, *Biomedical signal processing*. Biomedical engineering and design handbook, 2010. **2**: p. 559-579.
38. Smith, S.J.M., *EEG in the diagnosis, classification, and management of patients with epilepsy*. Journal of Neurology, Neurosurgery & Psychiatry, 2005. **76**(suppl 2): p. ii2-ii7.
39. Hazarika, N., T. Ah Chung, and A.A. Sergejew, *Nonlinear considerations in EEG signal classification*. IEEE Transactions on Signal Processing, 1997. **45**(4): p. 829-836.
40. Calvo, N.L., R.M. Maggio, and T.S. Kaufman, *Characterization of pharmaceutically relevant materials at the solid state employing*

- chemometrics methods*. J Pharm Biomed Anal, 2017.
41. Evans, C.H., *Applications of infrared, Raman and Resonance Raman spectroscopy in biochemistry*. Trends in Biochemical Sciences, 1984. **9**(6): p. 289.
42. Vandenabeele, P., *Raman spectroscopy*. Analytical and Bioanalytical Chemistry, 2010. **397**(7): p. 2629-2630.
43. Ellis, D.I. and R. Goodacre, *Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy*. Analyst, 2006. **131**(8): p. 875-85.
44. Krafft, C., et al., *Disease recognition by infrared and Raman spectroscopy*. J Biophotonics, 2009. **2**(1-2): p. 13-28.
45. Haka, A.S., et al., *Diagnosing breast cancer by using Raman spectroscopy*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(35): p. 12371-12376.
46. Mahadevan-Jansen, A. and R.R. Richards-Kortum, *Raman spectroscopy for the detection of cancers and precancers*. J Biomed Opt, 1996. **1**(1): p. 31-70.
47. Murugesan, M. and R. Sukanesh. *Automated Detection of Brain Tumor in EEG Signals Using Artificial Neural Networks*. in *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*. 2009.
48. Min, W. and G. Luo, *Medical Applications of EEG Wave Classification*. CHANCE, 2009. **22**(4): p. 14-20.
49. Lotte, F., et al., *A review of classification algorithms for EEG-based brain-*

- computer interfaces: a 10 year update*. Journal of Neural Engineering, 2018. **15**(3): p. 031005.
50. Harvey, T.J., et al., *Classification of fixed urological cells using Raman tweezers*. J Biophotonics, 2009. **2**(1-2): p. 47-69.
51. Mika, S., et al. *Fisher discriminant analysis with kernels*. in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*. 1999.
52. Ellis, D.I., et al., *Rapid identification of closely related muscle foods by vibrational spectroscopy and machine learning*. Analyst, 2005. **130**(12): p. 1648-54.
53. Widjaja, E., W. Zheng, and Z. Huang, *Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines*. Int J Oncol, 2008. **32**(3): p. 653-62.
54. Berger, A.J., I. Itzkan, and M.S. Feld, *Feasibility of measuring blood glucose concentration by near-infrared Raman spectroscopy*. Spectrochim Acta A Mol Biomol Spectrosc, 1997. **53a**(2): p. 287-92.
55. Brudzewski, K., et al., *Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach*. Fuel, 2006. **85**(4): p. 553-558.
56. Geladi, P., *Chemometrics in spectroscopy. Part 1. Classical chemometrics*. Spectrochimica Acta Part B-Atomic Spectroscopy, 2003. **58**(5): p. 767-782.
57. Geladi, P., et al., *Chemometrics in spectroscopy*. Spectrochimica Acta Part B: Atomic Spectroscopy, 2004. **59**(9): p. 1347-1357.

-
58. Müller, U., M. Diem: *Introduction to modern vibrational spectroscopy*, J. Wiley, New York, Chichester, 1993, ISBN 0-471-59584-5, 285 Seiten, Preis: £ 49.50. Berichte der Bunsengesellschaft für physikalische Chemie, 1994. **98**(10): p. 1347-1348.
59. Hanlon, E.B., et al., *Prospects for in vivo Raman spectroscopy*. Physics in Medicine and Biology, 2000. **45**(2): p. R1.
60. Raman, C.V. and K.S. Krishnan, *A new type of secondary radiation*. Nature, 1928. **121**: p. 501-502.
61. Lorenz, B., et al., *Cultivation-Free Raman Spectroscopic Investigations of Bacteria*. Trends in Microbiology, 2017. **25**(5): p. 413-424.
62. Neugebauer, U., P. Rosch, and J. Popp, *Raman spectroscopy towards clinical application: drug monitoring and pathogen identification*. Int J Antimicrob Agents, 2015. **46 Suppl 1**: p. S35-9.
63. Kirsch, M., et al., *Raman spectroscopic imaging for in vivo detection of cerebral brain metastases*. Anal Bioanal Chem, 2010. **398**(4): p. 1707-13.
64. Good, D.M., et al., *Body fluid proteomics for biomarker discovery: lessons from the past hold the key to success in the future*. J Proteome Res, 2007. **6**(12): p. 4549-55.
65. Lin, J., et al., *Erythrocyte membrane analysis for type II diabetes detection using Raman spectroscopy in high-wavenumber region*. Applied Physics Letters, 2014. **104**(10): p. 104102.
66. Chen, K., et al., *Towards ultrasensitive malaria diagnosis using surface enhanced Raman spectroscopy*. Scientific Reports, 2016. **6**: p. 20177.

-
67. Chen, K., et al., *Review of Surface Enhanced Raman Spectroscopy for Malaria Diagnosis and a New Approach for the Detection of Single Parasites in the Ring Stage*. IEEE Journal of Selected Topics in Quantum Electronics, 2016. **22**(4): p. 179-187.
68. Bocklitz, T.W., et al., *Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!?* Analytical Chemistry, 2016. **88**(1): p. 133-151.
69. Bocklitz, T., et al., *How to pre-process Raman spectra for reliable and stable models?* Anal Chim Acta, 2011. **704**(1-2): p. 47-56.
70. Burger, J. and P. Geladi, *Hyperspectral NIR image regression part I: calibration and correction*. Journal of Chemometrics, 2005. **19**(5-7): p. 355-363.
71. Mozharov, S., et al., *Automated Cosmic Spike Filter Optimized for Process Raman Spectroscopy*. Applied Spectroscopy, 2012. **66**(11): p. 1326-1333.
72. Li, S. and L. Dai, *An improved algorithm to remove cosmic spikes in Raman spectra for online monitoring*. Appl Spectrosc, 2011. **65**(11): p. 1300-6.
73. Schulze, H.G. and R.F.B. Turner, *A Two-Dimensionally Coincident Second Difference Cosmic Ray Spike Removal Method for the Fully Automated Processing of Raman Spectra*. Applied Spectroscopy, 2014. **68**(2): p. 185-191.
74. Cappel, U.B., I.M. Bell, and L.K. Pickard, *Removing Cosmic Ray Features from Raman Map Data by a Refined Nearest Neighbor Comparison Method as a Precursor for Chemometric Analysis*. Applied Spectroscopy, 2010. **64**(2): p. 195-200.

-
75. Vincent_Larat. *Data treatment and analysis*. 2014; Available from: https://raman-wurm2014.sciencesconf.org/conference/raman-wurm2014/pages/Vincent_Larat.pdf.
76. Guo, S., T. Bocklitz, and J. Popp, *Optimization of Raman-spectrum baseline correction in biological application*. Analyst, 2016. **141**(8): p. 2396-2404.
77. Zhang, Z.-M., et al., *An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy*. Journal of Raman Spectroscopy, 2010. **41**(6): p. 659-669.
78. Perez-Pueyo, R., M.J. Soneira, and S. Ruiz-Moreno, *Morphology-Based Automated Baseline Removal for Raman Spectra of Artistic Pigments*. Applied Spectroscopy, 2010. **64**(6): p. 595-600.
79. Liu, H., et al., *Joint Baseline-Correction and Denoising for Raman Spectra*. Applied Spectroscopy, 2015. **69**(9): p. 1013-1022.
80. Baek, S.-J., et al., *A simple background elimination method for Raman spectra*. Chemometrics and Intelligent Laboratory Systems, 2009. **98**(1): p. 24-30.
81. Lieber, C.A. and A. Mahadevan-Jansen, *Automated Method for Subtraction of Fluorescence from Biological Raman Spectra*. Applied Spectroscopy, 2003. **57**(11): p. 1363-1367.
82. Zhao, J., et al., *Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy*. Applied Spectroscopy, 2007. **61**(11): p. 1225-1232.
83. Mosier-Boss, P.A., S.H. Lieberman, and R. Newbery, *Fluorescence Rejection*

- in Raman Spectroscopy by Shifted-Spectra, Edge Detection, and FFT Filtering Techniques*. Applied Spectroscopy, 1995. **49**(5): p. 630-638.
84. Schulze, G., et al., *Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation*. Applied Spectroscopy, 2005. **59**(5): p. 545-574.
85. O'Grady, A., et al., *Quantitative Raman Spectroscopy of Highly Fluorescent Samples Using Pseudosecond Derivatives and Multivariate Analysis*. Analytical Chemistry, 2001. **73**(9): p. 2058-2065.
86. Zhang, D. and D. Ben-Amotz, *Enhanced Chemical Classification of Raman Images in the Presence of Strong Fluorescence Interference*. Applied Spectroscopy, 2000. **54**(9): p. 1379-1383.
87. Leger, M.N. and A.G. Ryder, *Comparison of Derivative Preprocessing and Automated Polynomial Baseline Correction Method for Classification and Quantification of Narcotics in Solid Mixtures*. Applied Spectroscopy, 2006. **60**(2): p. 182-193.
88. Camerlingo, C., et al., *Wavelet data processing of micro-Raman spectra of biological samples*. Measurement Science and Technology, 2006. **17**(2): p. 298.
89. Hu, Y., et al., *A background elimination method based on wavelet transform for Raman spectra*. Chemometrics and Intelligent Laboratory Systems, 2007. **85**(1): p. 94-101.
90. Jirasek, A., et al., *Accuracy and Precision of Manual Baseline Determination*. Applied Spectroscopy, 2004. **58**(12): p. 1488-1499.

91. Mazet, V., et al., *Background removal from spectra by designing and minimising a non-quadratic cost function*. Chemometrics and Intelligent Laboratory Systems, 2005. **76**(2): p. 121-133.
92. Byrne, H.J., et al., *Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells*. Chemical Society Reviews, 2016. **45**(7): p. 1865-1878.
93. Guiñón, J.L., et al., *Moving average and Savitzki-Golay smoothing filters using Mathcad*. Papers ICEE, 2007. **2007**.
94. Schafer, R.W., *What is a Savitzky-Golay filter*. IEEE Signal processing magazine, 2011. **28**(4): p. 111-117.
95. Zhao, A., et al. *The parameters optimization selection of Savitzky-Golay filter and its application in smoothing pretreatment for FTIR spectra*. in *2014 9th IEEE Conference on Industrial Electronics and Applications*. 2014.
96. Radzol, A.R.M., et al. *Optimization of Savitzky-Golay smoothing filter for salivary surface enhanced Raman spectra of non structural protein 1*. in *TENCON 2014 - 2014 IEEE Region 10 Conference*. 2014.
97. Awal, M.a., S. Mostafa, and M. Ahmad, *Performance Analysis of Savitzky-Golay Smoothing Filter Using ECG Signal*. International Journal of Computer and Information Technology, 2011. **1**(02): p. 24.
98. Randolph, T.W., *Scale-based normalization of spectral data*. Cancer Biomark, 2006. **2**(3-4): p. 135-44.
99. Pelletier, M.J., *Quantitative Analysis Using Raman Spectrometry*. Applied Spectroscopy, 2003. **57**(1): p. 20A-42A.

-
100. Hedoux, A., *Recent developments in the Raman and infrared investigations of amorphous pharmaceuticals and protein formulations: A review*. Adv Drug Deliv Rev, 2016. **100**: p. 133-46.
101. Franzen, L., et al., *Combining confocal Raman microscopy and freeze-drying for quantification of substance penetration into human skin*. Exp Dermatol, 2014. **23**(12): p. 942-4.
102. Franzen, L., et al., *Towards drug quantification in human skin with confocal Raman microscopy*. Eur J Pharm Biopharm, 2013. **84**(2): p. 437-44.
103. Tatum IV, W.O., *Handbook of EEG interpretation*. 2014: Demos Medical Publishing.
104. Chernecky, C.C. and B.J. Berger, *Laboratory Tests and Diagnostic Procedures-E-Book*. 2012: Elsevier Health Sciences.
105. Chiappa, K.H., *Evoked potentials in clinical medicine*. 1997: Lippincott Williams & Wilkins.
106. Bell, M.A. and K. Cuevas, *Using EEG to Study Cognitive Development: Issues and Practices*. Journal of cognition and development : official journal of the Cognitive Development Society, 2012. **13**(3): p. 281-294.
107. Janvale, G.B., et al., *Songs induced mood recognition system using EEG signals*. Annals of neurosciences, 2010. **17**(2): p. 80-84.
108. Li, X., et al., *The impact of mood on empathy for pain: Evidence from an EEG study*. Psychophysiology, 2017. **54**(9): p. 1311-1322.
109. Brown, J.E., et al., *Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal*

- stimulation*. PLoS One, 2011. **6**(9): p. e24124.
110. Loeser, J.D. and R.-D. Treede, *The Kyoto protocol of IASP basic pain terminology*. Pain, 2008. **137**(3): p. 473-477.
 111. Cruccu, G., et al., *EFNS guidelines on neuropathic pain assessment: revised 2009*. European Journal of Neurology, 2010. **17**(8): p. 1010-1018.
 112. Haanpää, M., et al., *NeuPSIG guidelines on neuropathic pain assessment*. PAIN®, 2011. **152**(1): p. 14-27.
 113. Schnakers, C. and N.D. Zasler, *Pain assessment and management in disorders of consciousness*. Current opinion in neurology, 2007. **20**(6): p. 620-626.
 114. Zaslansky, R., et al., *Pain-evoked potentials: what do they really measure?* Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 1996. **100**(5): p. 384-391.
 115. Treede, R., et al., *Evidence for two different heat transduction mechanisms in nociceptive primary afferents innervating monkey skin*. The Journal of Physiology, 1995. **483**(3): p. 747-758.
 116. Carmon, A., et al., *Single trial analysis of evoked potentials to noxious thermal stimulation in man*. Pain, 1980. **8**(1): p. 21-32.
 117. Iannetti, G.D., et al., *Determinants of laser-evoked EEG responses: pain perception or stimulus saliency?* Journal of neurophysiology, 2008. **100**(2): p. 815-828.
 118. Huang, G., et al., *A novel approach to predict subjective pain perception from single-trial laser-evoked potentials*. Neuroimage, 2013. **81**: p. 283-93.

-
119. Bai, Y., et al., *Normalization of Pain-evoked Neural Responses Using Spontaneous EEG Improves the Performance of EEG-based Cross-individual Pain Prediction*. *Frontiers in Computational Neuroscience*, 2016. **10**.
120. Yin, Z., et al., *Cross-Subject EEG Feature Selection for Emotion Recognition Using Transfer Recursive Feature Elimination*. *Frontiers in neurorobotics*, 2017. **11**: p. 19-19.
121. Brodersen, K.H., et al., *Decoding the perception of pain from fMRI using multivariate pattern analysis*. *Neuroimage*, 2012. **63**(3): p. 1162-1170.
122. Schulz, E., et al., *Decoding an Individual's Sensitivity to Pain from the Multivariate Analysis of EEG Data*. *Cerebral Cortex*, 2012. **22**(5): p. 1118 - 1123.
123. Villarroel, M., et al., *Time derivatives in air temperature and enthalpy as non-invasive welfare indicators during long distance animal transport*. *Biosystems Engineering*, 2011. **110**(3): p. 253-260.
124. Lieber, C.A. and A. Mahadevan-Jansen, *Automated method for subtraction of fluorescence from biological Raman spectra*. *Appl Spectrosc*, 2003. **57**(11): p. 1363-7.
125. Chen, S., et al., *Recovery of Raman spectra with low signal-to-noise ratio using Wiener estimation*. *Opt Express*, 2014. **22**(10): p. 12102-14.
126. Chen, S., Y.H. Ong, and Q. Liu, *A Method to Create a Universal Calibration Dataset for Raman Reconstruction Based on Wiener Estimation*. *IEEE Journal of Selected Topics in Quantum Electronics*, 2016. **22**(3): p. 164-170.
127. Sompel, D.V.d., et al. *Comparison of Gaussian and Poisson noise models*

- in a hybrid reference spectrum and principal component analysis algorithm for Raman spectroscopy*. in *SPIE BiOS*. 2013. SPIE.
128. Shen, H.-L. and J.H. Xin, *Spectral characterization of a color scanner based on optimized adaptive estimation*. Journal of the Optical Society of America A, 2006. **23**(7): p. 1566-1569.
129. Ong, Y.H., M. Lim, and Q. Liu, *Comparison of principal component analysis and biochemical component analysis in Raman spectroscopy for the discrimination of apoptosis and necrosis in K562 leukemia cells*. Optics Express, 2012. **20**(20): p. 22158-22171.
130. Bai, Y. and Q. Liu, *Denoising Raman spectra by Wiener estimation with a numerical calibration dataset*. Biomedical Optics Express, 2019. **11**(1): p. 200.
131. Puppels, G.J., et al., *Laser irradiation and Raman spectroscopy of single living cells and chromosomes: Sample degradation occurs with 514.5 nm but not with 660 nm laser light*. Experimental Cell Research, 1991. **195**(2): p. 361-367.
132. Tarcea, N., *Raman data analysis*. 2012.
133. Shafer-Peltier, K.E., et al., *Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis in vivo*. Journal of Raman Spectroscopy, 2002. **33**(7): p. 552-563.
134. Semrau, D., R.I. Killey, and P. Bayvel, *The Gaussian Noise Model in the Presence of Inter-Channel Stimulated Raman Scattering*. Journal of Lightwave Technology, 2018. **36**(14): p. 3046-3055.

-
135. Tian, D., et al. *Raman spectroscopy denoising based on smoothing filter combined with EEMD algorithm*. in *Tenth International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2017)*. 2018. SPIE.
136. Khoshroo, H., et al., *Quantitative analysis of essential oils of Thymus daenensis using laser-induced fluorescence and Raman spectroscopy*. Appl Opt, 2015. **54**(32): p. 9533-9.
137. Dörfer, T., et al., *Quantitative mineral analysis using Raman spectroscopy and chemometric techniques*. Journal of Raman Spectroscopy, 2010. **41**(6): p. 684-689.
138. Higgins, K., J.H. Kalivas, and E. Andries, *Evaluation of target factor analysis and net analyte signal as processes for classification purposes with application to benchmark data sets and extra virgin olive oil adulterant identification*. Journal of Chemometrics, 2012. **26**(3-4): p. 66-75.
139. Rezatofighi, S.H., et al. *A New Approach to White Blood Cell Nucleus Segmentation Based on Gram-Schmidt Orthogonalization*. in *2009 International Conference on Digital Image Processing*. 2009. IEEE.
140. Boulet, J.-C. and J.-M. Roger, *Pretreatments by means of orthogonal projections*. Chemometrics and Intelligent Laboratory Systems, 2012. **117**: p. 61-69.
141. Boiret, M., et al., *An iterative approach for compound detection in an unknown pharmaceutical drug product: Application on Raman microscopy*. J Pharm Biomed Anal, 2016. **120**: p. 342-51.

-
142. Lee, C.-J., et al., *The study of doxorubicin and its complex with DNA by SERS and UV-resonance Raman spectroscopy*. Bulletin of the Korean Chemical Society, 2004. **25**(8): p. 1211-1216.
143. Olszówka, M., et al., *A computational approach to the resonance Raman spectrum of doxorubicin in aqueous solution*. Theoretical Chemistry Accounts, 2016. **135**(2): p. 27.
144. Farhane, Z., F. Bonnier, and H.J. Byrne, *Monitoring doxorubicin cellular uptake and trafficking using in vitro Raman microspectroscopy: short and long time exposure effects on lung cancer cell lines*. Analytical and Bioanalytical Chemistry, 2017. **409**(5): p. 1333-1346.
145. Ju, J., et al., *Sustained and Cost Effective Silver Substrate for Surface Enhanced Raman Spectroscopy Based Biosensing*. Sci Rep, 2017. **7**(1): p. 6917.
146. Wilcox, D.S., et al., *Digital compressive chemical quantitation and hyperspectral imaging*. Analyst, 2013. **138**(17): p. 4982-4990.
147. Cordeiro Dantas, W.F., J.C. Laurentino Alves, and R.J. Poppi, *MCR-ALS with correlation constraint and Raman spectroscopy for identification and quantification of biofuels and adulterants in petroleum diesel*. Chemometrics and Intelligent Laboratory Systems, 2017. **169**: p. 116-121.
148. Stockel, S., et al., *The application of Raman spectroscopy for the detection and identification of microorganisms*. Journal of Raman Spectroscopy, 2016. **47**(1): p. 89-109.
149. Pahlow, S., et al., *Isolation and identification of bacteria by means of*

- Raman spectroscopy*. Advanced Drug Delivery Reviews, 2015. **89**: p. 105-120.
150. CLSI, C., *Reference Method for Broth Dilution Antifungal Susceptibility Testing of Filamentous fungi; Approved Standard—Second Edition—Document M38-A2*. 2008: Wayne, Pa, USA.
151. Vural, E., et al., *The effects of laser irradiation on Trichophyton rubrum growth*. Lasers in Medical Science, 2008. **23**(4): p. 349-353.
152. Terhaar, J., et al., *Differential processing of laser stimuli by A δ and C fibres in major depression*. PAIN®, 2011. **152**(8): p. 1796-1802.
153. Marquand, A., et al., *Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes*. Neuroimage, 2010. **49**(3): p. 2178-2189.
154. Zhang, Z., et al., *Gamma-band oscillations in the primary somatosensory cortex—a direct and obligatory correlate of subjective pain intensity*. The Journal of Neuroscience, 2012. **32**(22): p. 7429-7438.
155. Davis, K.D., *Neuroimaging of pain: what does it tell us? Current opinion in supportive and palliative care*, 2011. **5**(2): p. 116-121.
156. Klistorner, A.I. and S.L. Graham, *Electroencephalogram-based scaling of multifocal visual evoked potentials: effect on intersubject amplitude variability*. Investigative ophthalmology & visual science, 2001. **42**(9): p. 2145-2152.
157. You, Y., et al., *Normalization of visual evoked potentials using underlying electroencephalogram levels improves amplitude reproducibility in rats*. Invest Ophthalmol Vis Sci, 2012. **53**(3): p. 1473-1478.

-
158. Delorme, A. and S. Makeig, *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis*. Journal of neuroscience methods, 2004. **134**(1): p. 9-21.
159. Anokhin, A.P., et al., *Genetic influences on dynamic complexity of brain oscillations*. Neuroscience letters, 2006. **397**(1): p. 93-98.
160. Doležal, J., J. Štastný, and P. Sovka. *Recognition of direction of finger movement from EEG signal using markov models*. in *IFMBE Proc.* 2005.
161. Marcel, S. and J.R. Del Millan, *Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007. **29**(4): p. 743-752.
162. Näpflin, M., M. Wildi, and J. Sarnthein, *Test–retest reliability of resting EEG spectra validates a statistical signature of persons*. Clinical Neurophysiology, 2007. **118**(11): p. 2519-2524.
163. Tran, Y., A. Craig, and P. Mclsaac, *Extraversion–introversion and 8–13 Hz waves in frontal cortical regions*. Personality and Individual Differences, 2001. **30**(2): p. 205-215.
164. Zietsch, B.P., et al., *Common and specific genetic influences on EEG power bands delta, theta, alpha, and beta*. Biological psychology, 2007. **75**(2): p. 154-164.
165. Garrett, D.D., et al., *Moment-to-moment brain signal variability: A next frontier in human brain mapping?* Neuroscience & Biobehavioral Reviews, 2013. **37**(4): p. 610-624.
166. Barlow, J.S., *Rhythmic activity induced by photic stimulation in relation to*

- intrinsic alpha activity of the brain in man*. Electroencephalography and clinical neurophysiology, 1960. **12**(2): p. 317-326.
167. Lee, T.-W., et al., *Do resting brain dynamics predict oddball evoked-potential?* BMC neuroscience, 2011. **12**(1): p. 121.
168. Nash, K., M. Inzlicht, and I. McGregor, *Approach-related left prefrontal EEG asymmetry predicts muted error-related negativity*. Biological psychology, 2012. **91**(1): p. 96-102.
169. Polich, J., *On the relationship between EEG and P300: individual differences, aging, and ultradian rhythms*. International journal of psychophysiology, 1997. **26**(1-3): p. 299-317.
170. Ramos-Loyo, J., et al., *Relationship between resting alpha activity and the ERPs obtained during a highly demanding selective attention task*. International Journal of Psychophysiology, 2004. **54**(3): p. 251-262.
171. Rodin, E., et al., *Relationship of EEG background rhythms to photic evoked responses*. Electroencephalography and clinical neurophysiology, 1965. **19**(3): p. 301-304.
172. Rogers, L., *Ipsilateral and contralateral correlations between EEG and EP principal components*. Electroencephalography and clinical neurophysiology, 1980. **50**(5): p. 441-448.
173. Schiller, B., et al., *Individual differences in inhibitory control—Relationship between baseline activation in lateral PFC and an electrophysiological index of response inhibition*. Cerebral cortex, 2013: p. bht095.

-
174. Eli, I., et al., *Effect of gender on acute pain prediction and memory in periodontal surgery*. European Journal of Oral Sciences, 2000. **108**(2): p. 99-103.
175. Fillingim, R.B., *Sex, gender, and pain: Women and men really are different*. Current Review of Pain, 2000. **4**(1): p. 24-30.
176. Riley, J.L., 3rd, et al., *Sex differences in the perception of noxious experimental stimuli: a meta-analysis*. Pain, 1998. **74**(2-3): p. 181-7.
177. Wiesenfeld-Hallin, Z., *Sex differences in pain perception*. Gender Medicine, 2005. **2**(3): p. 137-145.
178. Kong, K., et al., *Raman spectroscopy for medical diagnostics--From in-vitro biofluid assays to in-vivo cancer detection*. Adv Drug Deliv Rev, 2015. **89**: p. 121-34.
179. de Juan, A. and R. Tauler, *Chemometrics applied to unravel multicomponent processes and mixtures*. Analytica Chimica Acta, 2003. **500**(1-2): p. 195-210.
180. Windig, W. and D.A. Stephenson, *Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach*. Analytical Chemistry, 1992. **64**(22): p. 2735-2742.
181. Zimonja, M., et al., *Multivariate curve resolution of mixed bacterial DNA sequence spectra: identification and quantification of bacteria in undefined mixture samples*. Journal of Chemometrics, 2008. **22**(5-6): p. 309-322.
182. de Juan, A., J. Jaumot, and R. Tauler, *Multivariate Curve Resolution (MCR). Solving the mixture analysis problem*. Anal. Methods, 2014. **6**(14): p. 4964-

- 4976.
183. Hu, L. and G.D. Iannetti, *Painful Issues in Pain Prediction*. Trends Neurosci, 2016. **39**(4): p. 212-220.
184. Coghill, R.C., *Individual differences in the subjective experience of pain: new insights into mechanisms and models*. Headache: The Journal of Head and Face Pain, 2010. **50**(9): p. 1531-1535.
185. Coghill, R.C., J.G. McHaffie, and Y.-F. Yen, *Neural correlates of interindividual differences in the subjective experience of pain*. Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8538-8542.
186. Edwards, R.R., *Individual differences in endogenous pain modulation as a risk factor for chronic pain*. Neurology, 2005. **65**(3): p. 437-443.