Signal Processing for Raman Spectroscopy

by

Aly Khedr

A thesis submitted to the Faculty of Graduate and Postdoctoral
Affairs in partial fulfillment of the requirements
for the degree of

Master of Applied Science

in

Biomedical Engineering

Carleton University
Ottawa, Ontario

© 2018

Aly Khedr

Abstract

Raman spectroscopy has been promoted as a non-invasive diagnostic technique capable of distinguishing molecular changes in biological samples. The quantification of spectra is limited by high frequency cosmic spikes, and broader fluorescence background artifacts.

Current preprocessing techniques eliminate artifacts but also distort Raman spectral bands in the process. The preprocessing techniques depend on many parameters that are difficult to optimize and rely on signal characteristics such as background shape and spectral and cosmic spike widths. This thesis presents an algorithm to remove cosmic spikes based on the principles of empirical mode decomposition and energy detection, and two algorithms for background fluorescence noise rejection by applying the principles of asymmetric penalized least squares minimization using non-quadratic cost functions. Performance analysis using simulated datasets show that the proposed algorithms outperform the state-of-the-art in both artifact suppression and Raman band preservation. This thesis also shows the importance of preprocessing Raman Spectra on classification.

Acknowledgments

I would like to thank Dr. Sreeraman Rajan for his support and guidance during the course of my studies. I will always be grateful for your inspiration and guidance as I continue my journey. I am very thankful for having the opportunity of having you as my mentor.

I would also like to thank Dr. Sangeeta Murugkar (Department of Physics), Dr. Balazs Nyiri (The Ottawa Hospital, Cancer Centre) and Dr. Vinita Chauhan (Health Canada) for offering their time to explain the physics behind Raman Spectroscopy, the data collection process, and the database used in this work. Thanks to their expertise in Raman Spectroscopy and numerous discussions, I was able to appreciate the complexities and dependencies in the dataset. I also want to acknowledge the help of Ms. Pavitra Ramachandran (University of Ottawa) for demonstrating how data was collected.

Most especially, I would like to thank my mom for her moral support and blessings.

Table of Contents

Abstract	ii
Acknowledgments	iii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
List of Abbreviations	x
1 Introduction	1
1.1 Motivation and Thesis scope	2
1.2 Contributions	3
1.3 Thesis organization	3
1.4 Data acquisition procedure for datasets	4
1.5 Instrumentation	5
1.6 Spectrum preprocessing	6
2 Literature survey	7
2.1 Smoothing	7
2.2 Cosmic ray spikes removal	9
2.3 Fluorescence background removal methods	13
2.3.1 Differentiation-based methods	13
2.3.2 Transform domain-based methods	14
2.3.3 Polynomial fitting and spline smoothing-based methods	17
2.4 Data normalization	18
2.5 Conclusion	20
3 Cosmic spike detection and removal algorithm	21
3.1 Introduction	21
3.2 The proposed algorithm	29
3.3 Experiments on datasets simulated	33
3.3.1 Simulated datasets	33

3.3.2	Results and Discussion	41
3.4	Conclusion	49
4	Background estimation and removal	50
4.1	Introduction	50
4.2	Weighted penalized least squares approximation with a non-quadratic cost function	54
4.3	Parameter optimization of the proposed algorithms	55
4.3.1	Effect of spectrum peak width on algorithm parameters	55
4.3.2	Effect of SNR on algorithm parameters	63
4.4	Simulated datasets	66
4.4.1	Effect of number of peaks on background estimation	68
4.5	Conclusion	75
5	Classification	76
5.1	Classification procedure	76
5.2	Principal Component Analysis (PCA)	76
5.3	Linear Discriminant Analysis (LDA)	77
5.4	Feature selection	78
5.5	Performance metrics	78
5.6	Smoothing	79
5.7	Test 1: Effect of data normalization on classification	82
5.7.1	Experimental procedure	83
5.7.2	Parameters settings	83
5.7.3	Results	83
5.8	Test 2: Effect of background subtraction on classification	84
5.8.1	Experimental procedure	84
5.8.2	Results	85
5.9	Test 3: Effect of cosmic spikes removal on classification	86
5.9.1	Results	86
5.10	Conclusion	87
6	Conclusion and Future work	88
6.1	Conclusion	88
6.2	Limitations:	89
6.3	Future Work	90

List of Tables

2.1	Common wavelets and their abbreviations	15
3.1	Pure Raman spectrum simulation parameters for dataset cD1	34
4.1	Parameters used for simulated spectra of datasets fD1, fD2 and fD3	67
5.1	A summary of classification results for vector, AUC, SNV, and non-normalized spectra for a) Experiment 1 b) Experiment 2	84
5.2	A summary of classification results for a) Experiment 3 showing classification results using optimally subtracted spectra . b) Experiment 4 showing classification results using spectra contaminated with fluorescence	86
5.3	A summary of classification results for a) Experiment 5 showing classification results using spike-free spectra. b) Experiment 6 showing classification results using spike contaminated spectra.	87

List of Figures

1.1	Jablonsky energy diagram: Quantum energy transitions for Rayleigh and Raman scattering [1]	2
1.2	Data acquisition process	5
1.3	Order of preprocessing steps after data acquisition in order to prepare data for classification	6
2.1	Sample irradiated spectrum with 5 Gy radiation from Donor 52 contaminated with cosmic spikes	9
2.2	Wavelet decomposition process. Adopted from [2]	16
3.1	a) Mono-component linear chirp signal, $S(t)$. b) Spectrogram of linear mono-component chirp signal, $S(t)$	23
3.2	Extracted instantaneous frequency of a mono-component linear chirp signal	24
3.3	a) Multi-component $S(t)$ (sum of two linear chirps). b) Spectrogram of multi-component $S(t)$	25
3.4	Extracted instantaneous frequency of sum of two linear chirps	26
3.5	a) Simulated Raman spectrum contaminated with spikes. b) IMF_1 extracted using EMD. c) IMF_1 extracted using EEMD	28
3.6	a) $c_m[n]$ of Raman spectrum from Donor 52 b) $c_{mlocalized}[n]$ of Raman spectrum from Donor 52	31
3.7	a) Event detection and localized event detection windows for $L=150$. b) $S_{filtered}[n]$ after first iteration of algorithm. c) $S_{filtered}[n]$ after second iteration of algorithm.	32
3.8	a) 50 simulated spectra of dataset cD1 before adding spikes. b) A simulated spectrum from dataset cD1 before adding spikes	35
3.9	Subset six spectra from dataset cD1, and the localized detection vector extracted on each of the 6 spectra before majority vote correction	36
3.10	Subset six spectra from dataset cD1, and the corrected event detection vector after majority vote correction	37
3.11	$IMF_1[n]$ of one filtered spectrum after spike removal and the corresponding detection vectors	38
3.12	A simulated spectrum from dataset cD2 a) before adding spikes b) after adding spikes	39

3.13	a) Spike from a dataset cD2 spectrum covering 2 pixels. b). Spike from a dataset cD3 spectrum covering 4 pixels	40
3.14	Dot plot showing location of spikes, if any, in dataset cD4	41
3.15	Difference between a simulated pure spectrum in dataset cD1 and spectrum after spikes are removed using different algorithms, a) Katsumoto-Ozaki (Kat-Oz) algorithm b) Zhang-Henson (ZH) algorithm c) The proposed algorithm	43
3.16	Difference between a simulated pure spectrum in dataset cD3 and spectrum after spikes are removed using a median filter with different window sizes. a) $L_{med}) = 3$. b) $L_{med}) = 9$	44
3.17	Difference between a simulated pure spectrum ($R[n]$) in dataset cD2 and spectrum after spikes are removed using the Katsumoto-Ozaki algorithm with median filter of a) $L_{KAT-Oz} = 3, 5, 23$, and 25 for $\eta_{KAT-Oz} = 1$. b) $L_{KAT-Oz} = 3, 5, 23$, and 25 for $\eta_{KAT-Oz} = 0.5$	45
3.18	NMSE box plot for dataset cD1	46
3.19	NMSE box plot for dataset cD2	46
3.20	NMSE box plot for dataset cD3	47
3.21	Difference spectrum for proposed algorithm and Tian algorithm. a) Raw Raman spectrum without any spikes. b) Tian algorithm using dB3 wavelet. c) Tian algorithm using sym2 wavelet. d) Proposed algorithm	48
3.22	Difference spectrum for a) Katsumoto-Ozaki algorithm b) Median filtering. c) Zhang-Henson	48
3.23	Spectra in cD4 before and after spike removal using proposed algorithm	49
4.1	Asymmetric cost functions adapted from [3] at $s=2$	53
4.2	Effect of peak width on NMSE for both algorithms using randomly chosen values	55
4.3	Effect of peak width on NMSE as a function of all the three parameters for spectrum with smallest simulated width ($w_x=10$) and spectrum with largest simulated width ($w_x=100$) for s and p varying from 10^{-4} to 0.9 in steps of 0.9 and ε varying from 10^2 to 10^3 in steps of 45	56
4.4	Effect of width on NMSE as a function of p for both proposed algorithms	58
4.5	Background estimation of spectrum for $p = 0.006$, $p=7.5e-4$, and $p=0$. for a) $w_x = 10$. b) $w_x=100$	59
4.6	Effect of width on NMSE as a function of s for both algorithms	60
4.7	Effect of width on NMSE as a function of ε for both algorithms	62
4.8	Background estimation for spectrum with peak width of a) width = 10. b) width=100 using $\varepsilon = 10^9$	63
4.9	Effect of estimating the background of a spectrum contaminated with noise with a small $p=0.001$ for ε set to 10^6 , $s=1$ and $\sigma_n=1$	64
4.10	Effect of estimating the background of a spectrum contaminated with noise with different values of p	65
4.11	Effect of estimating the background of a spectrum contaminated with noise with different values of ε for WTQPLS and WHPLS	66

4.12	One simulated spectrum from a) Dataset fD1. b) Dataset fD2. c) Dataset fD3 with background estimated using WTQPLS with parameters optimized for dataset fD1.	68
4.13	Box plot showing MSE of dataset fD1 and fD3 respectively using optimum parameters set for fD1 dataset	69
4.14	MSE of dataset fD1, fD2, and fD3 respectively using its own optimum parameters	71
4.15	MSE box plot for test ensemble from Data set fD1 with $s=0$	72
4.16	a) Spectrum from D52 and background estimated using WTQPLS ($p=0.0001$, $s=0.05$, $\varepsilon=100$). b) Background-free spectrum.	73
4.17	a) Spectrum from D52 and background estimated using WHPLS ($p=0.0001$, $s=0.05$, $\varepsilon=100$). b) Background-free spectrum	74
4.18	a) Spectrum from D52 and background estimated using WHPLS ($p=0.0001$, $s=0.013$, $\varepsilon=100$). b) Background-free spectrum	75
5.1	Region used for classification: $1504 - 1649 \text{ cm}^{-1}$ region	78
5.2	Effect of spectrum smoothing with SG filter with different window lengths	80
5.3	Effect of smoothing on Raman band at $727\text{-}780 \text{ cm}^{-1}$ (pixels 528-581)	81
5.4	Effect of smoothing on Raman band at $1498\text{-}1698 \text{ cm}^{-1}$ (pixels 1299-1499)	82
5.5	a) Optimal background subtraction (Experiment 4). b) Under-estimating the background (Experiment 3)	85
6.1	Twenty spectra background subtracted with parameters selected to produce a visually pleasing result using one spectrum	91

List of Abbreviations

APLS	Probability-based adaptive weight penalised least squares algorithm
APOLY	Probability-based adaptively weighted polynomial fit algorithm
AsLs	Asymmetric least squares smoothing
AUC	Area under the curve
CCD	Charged-coupled device
CWT	Continuous wavelet Transform
DWT	Discrete wavelet Transform
EEMD	Ensemble empirical mode decomposition
EMD	Empirical mode decomposition
HB	Asymmetric Huber cost function
HT	Hilbert transform
I-MODPOLY	Improved modified polyfit algorithm
IF	Instantaneous frequency
IMF	Intrinsic mode function
IPSA	Iterative polynomial smoothing algorithm
KAT-Oz	Katsumoto-Ozaki algorithm
LDA	Linear discriminant analysis
MODPOLY	Modified polyfit algorithm
MSC	Multiple scatter correction
MSE	Mean squared error
NMSE	Normalized mean squared error
PCA	Principal component analysis
PLS	Penalized least squares

SNR	Signal-to-noise ratio
TQ	Asymmetric truncated quadratic cost function
WHPLS	Proposed weighted Huber PLS algorithm
WTQPLS	Proposed weighted truncated quadratic PLS algorithm
ZH	Zhang-Henson algorithm

Chapter 1: Introduction

Raman spectroscopy is a non-contact, non-invasive, and non-destructive optical technique used to relay information about molecular vibrations of a sample. Characteristics of a sample can be inferred by studying the scattering of light when it is exposed to a beam of light. There are several types of scattering of light. Rayleigh Scattering, an elastic scattering, happens when light enters a medium where the molecules that form the medium are much smaller than the wavelength of the incident light. These molecules get excited and the excited molecule relaxes back down to the ground state and emit photons with equal energy to that of the incident photons. In elastic scattering, the kinetic energy is conserved. There is another type of scattering that occurs, the inelastic scattering, also known as Raman scattering. In this scattering, two different scenarios are possible. In the first scenario, the excited molecule may emit a photon with energy lower than the incident photon, this is referred to as Stokes Raman scattering. In the second scenario, the molecule is already at a high phonon level, and hence is excited above the virtual state. Once it relaxes back to the ground state, it emits a photon with greater energy than the incident photon. This type of scattering is known as Anti-Stokes Raman scattering. Figure 1.1 shows a Jablonsky diagram explaining the quantum energy transitions for Raman and Rayleigh scattering. Raman spectroscopy exploits the small fraction of light which undergoes Raman scattering. The Raman scattering light is separated into its component wavelengths using the diffraction grating, with each component being projected onto a different element (pixel) of the CCD detector. This results in a Raman spectrum which gives a measure of the intensity of Raman scattered energy as a function of the frequency difference from the incident energy.

Figure 1.1. Jablonsky energy diagram: Quantum energy transitions for Rayleigh and Raman scattering [1]

Since the discovery of the Raman scattering effect by Sir C.V Raman in 1922, the utility of Raman spectroscopy as a biomedical diagnostic tool has been demonstrated for biological tissues. It has also been widely used in pharmacology, and for tissue and biofluid analysis [4]. It has also shown high accuracy in differentiating between organic materials using portable spectrometers outside a lab setting. One drawback of Raman spectroscopy is that the resulting Raman spectrum is heavily contaminated with unwanted artifacts. These artifacts manifest themselves as sharp peaks, or broader peaks due to convolution with the sensing system's transfer function, or as slowly varying profiles superimposed on the artifact free Raman spectrum. The presence of such artifacts hinder the Raman spectra from being a reliable diagnostic technique. If Raman spectra should be made reliable and usable, these artifacts need to be removed. This thesis mainly focuses on signal processing techniques that enable the removal of artifacts from Raman spectra.

1.1 Motivation and Thesis scope

Studies have shown the negative consequences of ionizing radiation on blood and blood components such as loss of blood cell count, impairment of function and death. A more thorough review on the effects of radiation on whole blood cells can be found in [5]. In recent times, concern over mass radiation exposition has lead to research over reliable fast techniques to screen for radiation exposure. Existing technologies use cytogenetic analysis which depends on sophisticated laboratories and expensive equipment, and is not practical

as a response to a real world disaster. To address the limitations of existing technologies, Raman spectroscopy has been proposed as a rapid identification method.

The dependence of the Raman spectrum's intensity on many factors such as sample concentration, laser power, instrumental effects, and noise demands further processing before the spectrum can be interpreted in a meaningful manner. As Raman spectra are information-rich, and have the potential to serve as a diagnostic tool, the underlying useful information from the spectra need to be extracted. In order to extract the useful information contained in the Raman spectra, appropriate signal processing needs to be done. Two dominant processes which affect the Raman spectra are the spikes introduced in the spectrum by cosmic rays and the fluorescence that adds a background to the spectra. The overall purpose of this thesis is to remove the cosmic spikes and the background noise in Raman spectra through a signal processing solution without adversely affecting the useful information contained in the spectra. The overall aim of this thesis comprises the following:

- To provide novel artifact elimination algorithms that out perform the state-of-the art.
- To understand the role play of data normalization, spikes, and the fluorescence in classification of Raman spectra.

1.2 Contributions

To address cosmic spike and background removal from the Raman spectra, three novel algorithms are proposed. They are:

- A novel cosmic spike detection algorithm based on the principles of energy detection and ensemble empirical mode decomposition (EEMD). The detection algorithm presented in this thesis can be combined with any spike removal method.
- Two novel background estimation and subtraction methods based on the principles of minimizing a non-quadratic cost function proposed by [3] and asymmetric penalized least squares (AsLs) proposed by [6]. The proposed algorithms minimize the same objective function of AsLs but with a non-quadratic cost function instead of the normal least squares criterion.

1.3 Thesis organization

The thesis is divided into the following chapters:

1. Introduction
2. Literature survey
3. Cosmic spike detection and removal
4. Background estimation and removal

5. Classification

6. Conclusion and Future work

The first chapter (the current one) gives a brief summary of the Raman spectroscopic technique and provides relevant information on the thesis structure.

The second chapter gives a literature survey on approaches used to preprocess Raman spectra and their limitations. These include smoothing, cosmic ray spikes removal methods, fluorescence background removal methods, and data normalization.

The third chapter outlines a novel spike detection and removal algorithm. The proposed algorithm uses energy detection and ensemble empirical mode decomposition to extract and remove spikes from the Raman spectrum. After describing the algorithm, a comparative study is provided to highlight the performance of the proposed algorithm.

The fourth chapter proposes two novel background estimation methods by minimizing a weighted penalized least squares using non-quadratic cost functions. A detailed study of the algorithms' parameters is conducted to demonstrate their influence on the performance of the algorithms. A comparative study is also provided that compares the proposed algorithms to the state-of-the-art.

The fifth chapter combines preprocessing methods and multivariate techniques in order to assess the importance of spectra processing on differentiating between non-irradiated and irradiated (exposed to 1,2, and 5 Gy radiation) whole blood cells. In the experiments conducted, the effects of three preprocessing techniques on classification are tested: data normalization, cosmic spikes removal, and background elimination.

The final chapter summarizes the presented work, the limitation in the presented work and gives new potential solutions for preprocessing Raman spectra.

1.4 Data acquisition procedure for datasets

In this thesis, tests conducted were done on spectra collected from Health Canada on two volunteers (Donor 52 and Donor 125).

Both donors were male with ages unknown, ranging somewhere between 20-60 years. Peripheral blood from these volunteers were drawn via peripheral venipuncture with informed consent into 2 x 10 ml EDTA vacutainer tubes (Becton Dickinson and Company, Franklin Lakes, NJ). A 100 μ l whole blood sample was used for a complete blood count (CBC) via automatic haemocytometer (Beckman Coulter, Mississauga, ON).

The blood (3 ml) collected from the donors was aliquoted into flow tubes (Falcon) and irradiated at doses of 0 (control) 1,2 and 5 Gy using the X-RAD 320 X-ray irradiation system (Precision X-ray, Inc., North Branford, CT, USA) at a dose rate of 0.98 ± 0.05 Gy/min. The blood was rocked in a 37°C incubator for three hours and then immediately frozen at -80°C . Samples were then thawed and 50 μ l of blood was placed in an aluminum

slide with a well. The sample was then spread evenly across the well. Twenty spectra were collected for each dose using a new drop of blood each time for a total of 18 times (18 blood samples) for all donors. The acquisitions were conducted in a darkened room, containing minimal ambient light. For Donor 125, the collection process was done fully on two separate occasions, once on 7/24/2017 (day 1) and again on 8/18/2017 (day 2). The collection process for Donor 52 was done on 7/18/2017. For each donor, a total of 360 spectra were collected for each dose of radiation, with a total of 1440 Spectra.

The block diagram below in Figure 1.2 shows different steps in the acquisition process:

Figure 1.2. Data acquisition process

1.5 Instrumentation

The spectra were acquired using a Hyperflux Raman spectroscopy system (Tornado). The settings included a 785 nm diode laser, 20 second acquisition time, 5 averages, with 50 mW power. The spectral range (Raman shifts, cm^{-1} , from 785 nm excitation) was from 200 cm^{-1} to 3300 cm^{-1} . The system's CCD detector comprised 3101 detector elements (pixels) with each element corresponding to one wavenumber. The Hyperflux Raman spectroscopy

system has a proprietary software, SpectralSoft, which provides a real-time graphical spectrum display, an option to export recorded sessions and control the laser's power settings.

1.6 Spectrum preprocessing

After data (spectrum) acquisition, the raw Raman spectrum undergoes a series of preprocessing steps in order to remove anomalies in the spectrum and prepare the spectrum for classification. These steps include: smoothing, cosmic ray spikes removal, fluorescence background removal, and data normalization. Figure 1.3 shows the order of preprocessing steps taken after data acquisition.

Figure 1.3. Order of preprocessing steps after data acquisition in order to prepare data for classification

Chapter 2: Literature survey

This chapter gives a literature survey on approaches currently used to preprocess Raman spectra, and their limitations.

2.1 Smoothing

Raman spectrum strength depends on laser power. Samples such as blood are generally excited with a laser operating at low power to ensure that they do not get destroyed. As a consequence, the Raman spectra of these samples are contaminated with high levels of noise. To mitigate the noise some degree of smoothing is required before any further analysis is done. All smoothing involve convolving the raw spectrum with some weighted smoothing function.

The most basic form of smoothing is an equally weighted moving average, where each point in a smoothed spectrum, $\mathbf{z}$, is the average of N points from the raw spectrum, $\mathbf{S}$. This procedure can be expressed by the following equation (2.1)

$$z[n] = \frac{1}{N} \sum_{i=0}^{N-1} S[n+i], n = 1, 2, \dots N \quad (2.1)$$

where N is the number of points used in the moving average filter and n is the running index of the ordinate data. In this case, the moving average filter is a convolution using a rectangular pulse kernel of length N . Families of the moving average filter include different convolution smoothing kernels. For example, two passes of the moving average is equivalent to convolving the original spectrum with a triangular kernel, while a third pass is the same as convolving the original spectrum with a Gaussian one. Applying multiple smoothing with a small window improves SNR without heavily distorting the spectrum, as opposed to smoothing with a larger window, which increases the width of peaks, and in effect reduces spectral resolution.

Instead of using a moving average filter as the basis of smoothing, Savitsky Golay (SG) filter was proposed [7] for smoothing Raman spectra. The SG filter is based on a least square fit convolution. The main idea of the algorithm is that for every unsmoothed data point, i , a polynomial is fit by a least squares criterion, centered around i inside a sliding window

of fixed number of data points. Consequently, the central value of the polynomial at that point replaces the corresponding point in the unsmoothed spectrum. The window is then shifted by 1 data point, and the process is repeated [8]. The advantage of the SG method over the sliding average is that it provides a better trade-off between spectrum distortion and improvement of signal-to-noise ratio (SNR). In addition, the SG filter has no delays, hence does not cause shifts in the spectral peaks. The SG filter uses two parameters: window length, (L_{SG}), and polynomial order. If the window length is too long, it will lead to loss of potentially important information while if it is too short no noise reduction will occur. Setting the order too low leads to over-smoothing and spectrum distortion, while setting it too high leads to false artificial peaks. In general, optimum smoothing provides a balance between reducing noise and preserving the spectrum [9].

Another approach widely used is based on penalized least squares, proposed by Whittaker in 1923. The Whittaker smoother is fast and can adapt automatically to boundaries. It is more robust than SG in handling missing data. This smoothing algorithm was revived in chemistry by Eilers [6, 10, 11, 12]. The main idea of the algorithm is that while smoothing data a balance between two conflicting goals must be realized [6]:

- Fidelity to the data (J)
- Roughness of the estimated smoothing (E)

Fitting to the data can be expressed using the sum of square of differences, as in ordinary least squares, given by (2.2):

$$J = | \mathbf{S} - \mathbf{z} |^2 \quad (2.2)$$

where $\mathbf{S}$ and $\mathbf{z}$ are the unsmoothed and smoothed spectrums, respectively.

Roughness, E , is expressed by the sum of the squared second order differences of the smoothed vector $\mathbf{z}$ given by (2.3):

$$E = \sum_i (\Delta^2 z_i)^2 \quad (2.3)$$

where $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2})$ where i represents the elements of the smoothed vector $\mathbf{z}$.

The Whittaker smoother balances the two conflicting goals as given in the vector form by the following equation (2.4):

$$H = J + \varepsilon E \quad (2.4)$$

The ε parameter ranges from 0 to infinity. It provides a balance between smoothing and fidelity. The larger ε is, the smoother the spectrum will be at the cost of a worse fit to

the data. After minimizing H , the smoothed spectrum is given by(2.5):

$$\mathbf{z} = (\mathbf{I} + \varepsilon \mathbf{D}'\mathbf{D})^{-1}\mathbf{S} \quad (2.5)$$

with $\mathbf{D} = \Delta^2 \mathbf{z}$. and where $'$ denotes the transposition of the vector.

2.2 Cosmic ray spikes removal

Cosmic spikes are anomalies that occur due to the sensitivity of charge-coupled devices (CCD) to cosmic rays (muons, protons, and neutrons). Particles such as muons, protons, and neutrons cause electrons to charge up consecutive elements of the CCD detector. This leads to sharp peaks in the output of the detector, termed as cosmic spikes [13]. The cosmic spikes have very sharp peaks with widths ranging from 1 cm^{-1} to 3 cm^{-1} , and with variable amplitudes [14]. Due to the stochastic nature of the ionizing particles collision process, the spikes are randomly distributed in both time and space [15]. Raman spectra often contain one or more cosmic spike artifacts as can be seen in Figure 2.1 or may not contain any spikes at all.

Figure 2.1. Sample irradiated spectrum with 5 Gy radiation from Donor 52 contaminated with cosmic spikes

Cosmic spikes destroy meaningful chemical information from Raman spectra, rendering both nonparametric and parametric features selection for classification and classification

itself ineffective. For example, cosmic spikes lead to false principal component directions while conducting principal component analysis (PCA)[16]. There are many signal processing methods available in literature to both detect and remove cosmic spikes. These methods can be mainly divided into two categories:

1. Single-scan methods
2. Multiple-scan methods

Single-scan methods detect and remove cosmic spikes using individual Raman spectrum. They are based on the assumption that cosmic spikes have a smaller bandwidth than that of the Raman peaks. Single-scan methods include smoothing, moving window filtering [17], and wavelet-based methods [13, 18]. Smoothing and moving window filtering methods remove the cosmic spikes but heavily distort the Raman spectrum in non-spike regions as they apply a filtering operation to the whole spectrum. To mitigate filtering the whole spectrum, Katsumoto and Ozaki (KAT-Oz) [19], employed a moving average method for spike detection. The logic behind the algorithm is that if a running window contained a spike, the value of the spike should show up as an anomalous deviation in the ordinate with respect to the window average. After spikes are detected as anomalous deviations, above a specified threshold, η_{KAT-Oz} , they are replaced by values obtained by the moving-average of the window. After this step, the moving-average process is repeated and spikes are detected by comparing the difference of the averaged spectrum with the original one.

Wavelet-based methods decompose the spectrum and localize the cosmic ray spikes. This targeted removal is preferred as it theoretically does not alter the Raman spectrum in non-spike regions. However, wavelet-based removal is prone to false detections due to their reliance on the choice of appropriate wavelet, threshold, and the level of decomposition. A general limitation of single-spectrum methods is the assumption that spikes are always narrower than the Raman peaks. Despite this limitation, single scan methods are relevant in cases where only individual spectra are available.

Multiple-scan methods detect and remove cosmic spikes using more than one Raman spectra. They are based on the assumption that the probability of a cosmic spike occurring at the same wavenumber for two different spectra is very low [14]. The simplest multiple-scan methods involve taking the difference between consecutive spectra in order to identify the spikes. Advanced algorithms such as the Upper-bound spectrum method (UBS) go beyond using the difference approach. UBS was one of the first multiple-scan methods and is used in most Raman software [16]. UBS assumes that any two consecutive spectra, S_i and S_j , in a dataset are related by (2.6):

$$S_i = C_{UBS,i,j}S_j + \Delta \quad (2.6)$$

where S_i, S_j are two consecutive spectra, $C_{UBS,i,j}$ is a constant scalar factor which represents the difference in intensities between the two spectra, and Δ is a noise vector representing white noise and cosmic spikes. The $C_{UBS,i,j}$ constant factor is determined by setting it to a

value that causes 50% of a residual vector, x (2.7), to be negative.

$$x = S_i - C_{UBS,i,j}S_j \quad (2.7)$$

This ensures reliable scaling of spectra. After $C_{UBS,i,j}$ is set, it is used to construct an upper-bound spectrum. For two collected spectra, the resulting upper-bound spectra is given by (2.8):

$$\begin{aligned} U_1 &= C_{UBS,1,2}S_2 + 3.5\sqrt{\min[C_{UBS,1,2}S_2, S_1] + \sigma^2v_r} \\ U_2 &= C_{UBS,2,1}S_1 + 3.5\sqrt{\min[C_{UBS,2,1}S_1, S_2] + \sigma^2v_r} \end{aligned} \quad (2.8)$$

where σ^2v_r is the read noise factor. The read noise factor depends on the A/D converter. Its value ranges from 25 to 10^5 . The upper bound spectra are compared to the original spectra. If the original spectrum, S_i , contains an intensity greater than its corresponding U_i , the intensity is marked as a potential spike. After all potential spikes are identified, the constant factor, $C_{UBS,i,j}$, is re-approximated through a least squares criterion such that $\hat{S}_i$ and $\hat{S}_j$ are equal (where $\hat{S}_i$ and $\hat{S}_j$ are S_i and S_j after removing the suspected spikes), that is, $\hat{S}_i$ and $\hat{S}_j$ are equated as follows by equation (2.9):

$$\hat{S}_i = C_{UBS,i,j}\hat{S}_j \quad (2.9)$$

The suspected spikes at each point for one spectrum are then replaced with the suspected spikes of the other spectrum at the corresponding point multiplied by the re-approximated $C_{UBS,i,j}$ parameter. The main limitation of the UBS algorithm is the difficulty in setting the noise factor, σ^2v_r , which does not depend on the Raman spectrum but on the A/D converter of the Raman laser. Wrong selection may lead to both miss and false detections. In addition, as with the simple multiple-scan methods, the UBS method performs poorly when consecutive spectra have different fluorescence background profiles or when Raman peaks of consecutive spectra do not appear exactly at the same location [16]. Non-alignment of peaks is especially a problem with spectra of dynamic samples where sequentially acquired spectra may not be identical [14]. For instance, if several spectra of blood that contain different count of red blood cells for a given volume are acquired, then the Raman spectrum would differ from one sample to another.

More recent algorithms [14, 15, 20] resolve these issues by relaxing the requirement of multiple spectra being fully identical. An algorithm proposed by Tian *et.al* [15], combines wavelet transforms and data clustering methods to automatically detect and remove spikes from Raman spectra. The first step of the algorithm applies the wavelet transform up to the k^{th} level to decompose each spectra. K-means clustering, for each decomposition level, is then employed on the coefficients of all spectra. Clustering search is then performed to locate the erroneous coefficients. The spectrum is then reconstructed without the erroneous coefficients using the inverse wavelet transform. The algorithm works under the assumption

that measured spectra are relatively similar with the exception of spikes. Under this assumption, the wavelet coefficients of the unfiltered Raman spectrum should be similar enough in order to aggregate into one cluster while the coefficients of the spikes should aggregate in a separate cluster. To account for variations between spectra, the cluster radius of higher decomposition levels is set to a large value. The main drawback of the algorithm is that an appropriate choice of wavelet is needed or else distortion of the pure Raman spectrum during the reconstruction phase of the algorithm will occur. In addition, the clustering radius is empirically selected to account for potential variation between different spectra. If not correctly set, it will lead to either false or misdetections.

Another promising and simple algorithm is that proposed by Zhang *et al.* The Zhang-Henson algorithm (ZH) [21] can be summed up as follows:

- All spectra are median filtered.
- The correlation coefficients of all spectra combinations is calculated. For each spectra, its neighboring spectrum is defined as the one which produced the highest correlation coefficient.
- Each spectra is then approximated with a linear combination of its most correlated neighbor and reference spectra (if available) using the following linear model expressed in vector form: (2.10).

$$\mathbf{z} = a\mathbf{p} + b + \sum_{n=0}^N C_{ZH,n}S[n] + \mathbf{e} \quad (2.10)$$

where $\mathbf{z}$ is the approximated spectrum, $\mathbf{p}$ is $\mathbf{z}$'s nearest neighbor, $\mathbf{S}$ is the reference spectrum with a length of N , a is a scaling factor, b is an offset used to account for spectral intensity variations between the spectrum and its nearest neighbor, $C_{ZH,n}$ is a scaling factor corresponding to the pure unfiltered reference spectrum, and $\mathbf{e}$ is the intrinsic model error.

- If no reference spectra are present, the last term in (2.10) is omitted. Spike outliers are then detected and removed by replacing them with the corresponding points of the approximated spectra. Detection occurs by applying a dual threshold to a residual vector, $\mathbf{x}$, calculated as follows: (2.11):

$$\mathbf{x} = \frac{\mathbf{p} - \mathbf{z}}{\sigma_z} \quad (2.11)$$

where σ_z is the standard deviation of $\mathbf{z}$. For each residual vector, all values greater than a low threshold $\eta_{ZH,1}$ are identified. A second higher threshold, $\eta_{ZH,2}$, is then defined. Any values above $\eta_{ZH,2}$ for the values identified by $\eta_{ZH,1}$ are considered as spikes.

The main shortcoming of the algorithm is its dependence on reference spectra, which is impractical in real-life scenarios where the spectra may not be known *a priori*. The second shortcoming is that there is no mechanism to choose either threshold.

2.3 Fluorescence background removal methods

Background noise manifests itself as a broad, slowly varying signal superimposed on the Raman spectrum. It arises due to the absorption of fluorescence from organic molecules, and from stray light entering the detector [3, 22]. The intensities of fluorescence noise are orders of magnitude higher than that of the Raman scattering, and therefore background estimation and removal are important tasks in order to extract Raman peaks masked by the background noise.

Literature contains plenty of methods to estimate and remove the background noise. The most common methods include:

1. Differentiation-based methods.
2. Transform domain-based methods.
3. Polynomial fitting and spline smoothing-based methods.

A brief summary of these methods along with their advantages and disadvantages are given below.

2.3.1 Differentiation-based methods

Differentiation-based methods are one of the earliest techniques used in analytical spectroscopy for background estimation and removal [23]. Differentiation applies a "band-sharpening effect" on Raman spectra. It suppresses low frequency bands while enhancing sharper high frequency ones. This phenomena increases with increasing derivative order [24] due to - for Lorentzian and Gaussian bands - the k^{th} order derivative's amplitude being inversely proportional to the spectrum's original bandwidth raised to the k^{th} power.

The main disadvantage in using differentiation methods to extract the relatively sharp Raman spectrum from the broader background is that signal-to-noise ratio (SNR) decreases with increasing derivative orders [25]. To show how differentiation affects noise we consider the following experiment. Consider a signal consisting of only white Gaussian noise. The noise level of the *zeroth* order derivative for this signal is given by the standard deviation of its elements, σ_0 [25]. For the k^{th} order derivative, the noise is given by a weighted sum of σ_0^2 as follows:

$$\sigma_k = \sigma_0 \sqrt{\sum_{m=0}^k \left(\frac{k!}{(k-m)!m!} \right)^2} \quad (2.12)$$

To mitigate this, derivative techniques require that the spectrum be first smoothed. Higher order derivatives require more smoothing than lower order derivatives. This extra smoothing may itself distort the spectrum. After smoothing, signal differentiation is applied

to remove the baseline [26]. In general, derivative algorithms are not recommended as they change original peak shapes and therefore render data analysis meaningless.

2.3.2 Transform domain-based methods

Fourier and wavelet transforms are powerful tools that are widely used in conventional signal processing [27]. The Fourier transformation, given by (2.13), use both cosines and sines as a basis function for extracting frequency (“k”-space) information from the raw Raman spectrum $S(\lambda)$.

$$S(f) = \int_{-\infty}^{\infty} S(\lambda) e^{-i2\pi f\lambda} d\lambda \quad (2.13)$$

It decomposes the Raman spectrum into linear combinations of sines and cosines, with each combination representing a unique frequency. (Note: For ease of understanding, we interchangeably use frequency for k-space). The amplitude of the frequency, in turn, represents the combinations contribution to the Raman spectrum. Baseline removal using the Fourier transform is done by filtering low frequencies which correspond to the broad background with a high pass filter [28]. The Raman spectrum is then reconstructed by applying the inverse Fourier transform given by (2.14):

$$S(\lambda) = \int_{-\infty}^{\infty} S(f) e^{i2\pi f\lambda} df \quad (2.14)$$

In order to handle non-stationarity in the Raman spectrum, the wavelet transform is often used. It uses the mother wavelet, ψ , as a basis function [29]. The mother wavelet is a prototype from which other basis functions can be derived by both translating and dilating it, as given by (2.15), to provide a whole family of wavelets.

$$\psi_{\alpha,\kappa}(\lambda) = \frac{1}{\sqrt{|\alpha|}} \psi\left(\frac{\lambda - \kappa}{\alpha}\right) \quad (2.15)$$

where α is the dilation constant, and κ is the translation constant.

Common mother wavelet families [30], and their abbreviations, are provided in Table 2.1.

The continuous wavelet transform is given by (2.16):

$$CWT(\alpha, \kappa) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} S(\lambda) \psi\left(\frac{\lambda - \kappa}{\alpha}\right) d\lambda \quad (2.16)$$

It can be thought of as the inner product of the original spectrum with a scaled, shifted version of the mother wavelet. Calculating wavelet coefficients at every scale is inefficient because of over representation (or redundant decompositions). Instead, wavelet analysis is

Table 2.1. Common wavelets and their abbreviations

Name of Wavelet families	Abbreviation
Haar Wavelet	haar
Daubechies	Db
Symlets	sym
Coiflets	coif
Biorthogonal wavelets	bior
Meyer wavelet	meyr
Discrete approximation of the Meyer wavelet	dmey

done on dyadic (based on power of 2) scales (dilations) and positions (translations). This is obtained using the discrete wavelet transform (DWT) given by:

$$DWT(j, k) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} S(\lambda) \psi\left(\frac{\lambda - 2^j k}{2^j}\right) d\lambda \quad (2.17)$$

with the scale and position, α and κ , being replaced with 2^j and $k2^j$, respectively [31]. DWT can be interpreted as a multi-stage filter bank with high-pass (HP) and lowpass (LP) filters performing a series of decompositions on the spectrum. Coefficients obtained after the HP filters are called *detail* coefficients while those after the LP filter are called the *approximation* coefficients. In the first decomposition level, for a Raman spectrum of length N , the spectrum is decomposed into $N/2$ low frequencies, represented by the first *approximation* coefficient, and $N/2$ high frequencies, represented by the first detail coefficients. Subsequent decompositions at each level further decompose the first *approximation* coefficients in order to improve frequency resolution at low frequencies [30]. Figure 2.2 shows the decomposition process explained above.

Figure 2.2. Wavelet decomposition process. Adopted from [2]

Following this scheme, background subtraction using DWT is done by first applying an n -level decomposition on the Raman spectrum, and then nullifying the n th approximation coefficient, $A_{DWT,n}$. The Raman spectrum is then reconstructed using the inverse discrete wavelet transform [2, 31, 32]. Wavelets have also been used in literature for Raman peak identification [27], after which other subtraction methods were used to remove the background [27, 33].

The main disadvantage of using wavelets is that it assumes that the background and the spectrum are well separated in the transform domain. In addition, the method requires a lot of parameters to be optimized such as the mother wavelet, the decomposition level, and threshold employed, all of which affect accurate estimation.

2.3.3 Polynomial fitting and spline smoothing-based methods

Polynomial fitting and spline smoothing are considered to be the most promising type of background estimation methods as they approximate the fluorescence background, B , while retaining the shape of the Raman peaks [34]. In this class of methods, the background is modeled as a polynomial of order k with $\beta = [\beta_0, \beta_1, \dots, \beta_k]$ coefficients such that the approximation function (2.18) best represents the fluorescence background fit. Different approaches have been attempted to estimate the coefficients using ordinary least squares (OLS), non-quadratic cost functions, and asymmetric least squares estimation techniques.

Consider the following $(k - 1)^{th}$ degree polynomial.

$$z(\lambda) = \beta_0 + \beta_1\lambda + \beta_2\lambda^2 + \dots + \beta_{k-1}\lambda^{k-1} \quad (2.18)$$

where λ represents the spectral range of the spectrum. Ordinary least squares minimizes the objective function, $H(\beta)$, given by (2.20), using a quadratic cost function (2.19).

$$\varphi(\mathbf{x}) = \mathbf{x}^2 \quad (2.19)$$

A quadratic cost function minimizes the sum of squares of the differences between the estimated background vector, $\mathbf{z}$, and the associated measured Raman spectrum given by $\mathbf{S}$. where $\mathbf{x}$ is $\mathbf{z} - \mathbf{S}$.

$$\begin{aligned} H(\beta) &= \varphi(\mathbf{z} - \mathbf{S}) \\ &= (\mathbf{z} - \mathbf{S})^2 \\ &= (\mathbf{z} - \mathbf{S})'(\mathbf{z} - \mathbf{S}) \\ &= (\beta X - \mathbf{S})'(\beta X - \mathbf{S}) \\ &= \beta' X' X \beta - \beta' X' \mathbf{S} - \mathbf{S}' X \beta + \mathbf{S}' \mathbf{S} \end{aligned} \quad (2.20)$$

(Note that $'$ represents transposition of a vector). To minimize H with respect to β , in order to find the optimal coefficients, the gradient is set to 0.

$$\begin{aligned} \frac{\partial H}{\partial \beta} &= 0 = \beta' X' X + (X' X \beta)' - (\mathbf{S}' X) - (X' \mathbf{S})' \leftrightarrow X' X \beta = X' \mathbf{S} \\ \beta &= (X' X)^{-1} X' \mathbf{S} \end{aligned} \quad (2.21)$$

where X is the Vandermonde matrix. Through this process, polynomial fits using different orders [35, 36, 37, 38, 39] were performed. To ensure that Raman peaks did not contribute to the coefficient calculation, estimation was done on a predetermined spectral range that did not contain Raman peaks.

To further improve background estimation, [37] proposed first dividing the Raman spectrum into multiple segments since the background is usually too complex to be fitted

properly using a single polynomial. After fitting each segment, the segments are attached to produce a continuous estimated background. The problem with this method is that adjacent segments produce high frequency artifacts due to abrupt changes in slope of the spectrum’s boundary regions. To mitigate this, careful manual selection of the segments was needed to ensure collinearity between segment boundaries. Manual fitting, however, is neither effective nor fast, and requires *a priori* studying of the data. Moreover, like previous methods, the fit depends on a manually selected polynomial order, and a chosen spectral fitting range. Selecting a high order may lead to overshoots in the estimated background, while a lower one may not be enough to model it.

To overcome the shortcomings of simply fitting a polynomial to estimate the background [40] proposed an iterative weighted least squares algorithm, ModPoly, to reduce dependence on the polynomial order and fitting the data on a chosen spectral range. The method smoothed the spectrum such that Raman peaks were clipped, leaving the background intact. This method works by fitting a polynomial to the original spectrum. For each wavenumber, the minimum of the fitted and original spectrum is calculated to form a new spectrum which is in turn refitted. This process is repeated until some criteria is met. ModPoly was further improved by [41], who introduced I-ModPoly, by taking into account signal-to-noise ratio during fitting to ensure noise greater than the estimated background is not falsely treated as a Raman peak. However, both these methods still had limitations. The limitations were that the polynomial order was fixed during the whole iteration process, and the result still depended on a chosen spectral range. To overcome this limitation, [42] proposed a spectral range independent iterative polynomial smoothing algorithm (IPSA). It took into consideration the iteration method used in I-ModPoly and the use of a second order Savitzky-Golay polynomial filter to smooth the spectrum.

All the previous algorithms discussed used OLS based minimization in order to estimate the background. More recently, however, [3, 43] paved the way for a new class of background estimation methods based on minimizing non-quadratic cost functions, and asymmetric least squares smoothing, respectively. Chapter 4 provides further information on these methods, and two novel algorithms which make use of their principles.

2.4 Data normalization

Another important step to prepare Raman spectra for data analysis is normalization. Data normalization is especially important for Raman spectra acquired sequentially as successive scans may exhibit variations in intensity due to subtle changes in conditions during data collection. The goal of normalization is to ensure that different spectra, when no biological difference is present, are equivalent. Apart from time variation due to successive scans, variation on spectra acquired on different days where experimental conditions may not be the same must also be accounted for. For a calibration set, V , of m spectra, $V = [S_1, S_2, \dots, S_m]$, collected from the same biological sample, but under different conditions, optimal data normalization is one which ensures that the variations due to subtle changes in conditions during data collection is removed such that all spectra are identical [44]. There

are many different normalization techniques in literature [45, 44, 27]. These include: vector, standard normal variate (SNV), and area (AUC) normalization.

Vector normalization is a scaling transformation that is applied to normalize successive spectra containing different amounts of energy by transforming them to unit length vectors. Mathematically, this is done as follows (2.22):

$$\begin{aligned} fn_i &= \frac{S_{ij}}{\|S_j\|} \\ &= \frac{S_{ij}}{\sqrt{S_{i1}^2 + S_{i2}^2 + \dots + S_{iN}^2}}, \quad j = 1, 2, \dots, N; i = 1, 2, \dots, m \end{aligned} \quad (2.22)$$

with the unnormalized i -th spectrum $S_i = [S_{i1}, S_{i2}, \dots, S_{iN}]$, N being the number of pixels (spectral data points) in each spectrum, the normalized i -th spectrum is given by $Fn_i = [fn_1, fn_2, \dots, fn_N]$, f_{ij} and fn_j being the intensity of the spectrum at the corresponding Raman shift j for the initial and normalized spectrum respectively,

The SNV transformation is part of a family of normalization procedures that minimize the variation of Fn_i about a mean, μ_{S_i} , [44]. It is given by (2.23):

$$fn_i = \frac{S_{ij} - \mu_{S_i}}{\sqrt{\text{var}(S_i)}} \quad j = 1, 2, \dots, N \quad (2.23)$$

with $\text{var}(S_i)$ being the variance of the unnormalized i -th spectrum. Unlike vector normalization, SNV normalization is a weighted normalization. It normalizes the Raman spectrum by the total energy of S_i centered about its mean, leaving the Raman spectrum with unit standard deviation. A closely related and more common normalization is the multiplicative scatter correction (MSC) method [46]. MSC assumes each spectrum is a linear combination of an ideal spectrum, $\bar{S}$, given by the average of the calibration set V , and some offset as seen in (2.24)

$$S_i = a + b\bar{S}_i + \text{offset} \quad (2.24)$$

The coefficients a and b are estimated by ordinary least-squares regression between S_i and $\bar{S}$. The offset can be either assumed to be known or can also be part of the regression analysis. Under this assumption, MSC finds each intensity of the corrected spectrum by (2.25):

$$fn_i = \frac{S_{ij} - a}{b} \quad \text{for } j = 1, 2, \dots, N \quad (2.25)$$

MSC requires more than one spectra, and therefore is not suitable for classification where it is unknown whether spectra in a dataset are of the same chemical compound or not.

Another normalization typically used is area under the curve (AUC) normalization. In this type of normalization, each spectrum is divided by its area such that their AUC is unity.

2.5 Conclusion

This chapter summarized current preprocessing methods being used in the field of Raman spectroscopy.

The first type of preprocessing method discussed was cosmic spike removal. Current techniques are divided into two categories: single-scan methods, and multiple scan methods. The main limitation in both categories is their reliance on many parameters that are difficult to optimize, and their failure in preserving the pure Raman spectrum.

The second type of preprocessing method discussed was smoothing. Smoothing is employed in order to improve SNR for Raman spectra acquired using a lower power laser. The most common type of smoothing used on Raman spectra is Savitzky-Golay (SG) smoothing. SG smoothing requires two parameters: the polynomial order and window length. Generally, the parameters are heuristically selected such that an acceptable balance between spectrum preservation and increase in SNR is achieved. Another algorithm that is widely used, especially when there is missing data, is the Whittaker smoother whose objective function provides a balance between data fidelity and smoothing.

The third type of preprocessing method discussed was background removal. Current techniques are generally divided into the following three categories: differentiation-based techniques, transform-based techniques, and polynomial fitting and spline smoothing-based techniques. Differentiation-based techniques heavily distort Raman band shapes, while transform-based techniques assume that the background and spectrum are well separated in the transform domain. Polynomial techniques are the most promising of the three categories. Most of the polynomial techniques in literature make use of minimization of a least-squares criterion. However, recent research is leaning towards using cost functions that give a lower cost to positive residuals (the positive Raman peaks), effectively not including them in the polynomial estimation.

The last type of preprocessing method discussed was data normalization. Data normalization attempts to remove all non-chemical variations between spectra in order to make data analysis meaningful. Normalization methods include, and are not limited to: vector, SNV, MSC, and AUC normalization.

Chapter 3: Cosmic spike detection and removal algorithm

3.1 Introduction

Cosmic spikes present themselves as high frequency artifacts superimposed on the Raman spectrum. They come in variable intensities, and occur randomly. If not treated, they reduce the effectiveness of using a Raman spectra for classification. In this chapter a novel algorithm based on ensemble empirical mode decomposition (EEMD) and energy detection is investigated as a tool to detect cosmic spikes. Once found, the spikes are mapped to the real spectrum, and then replaced.

Empirical Mode Decomposition (EMD) [47] is an adaptive method for analyzing non-linear and non-stationary data. The decomposition extracts energy associated with fast and slow oscillations in the signal - each represented by a finite number of intrinsic mode functions (IMFs). Each IMF can be viewed as a sub-band of the signal, and must satisfy the following conditions:

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

The motivation behind developing EMD method arises from ambiguity regarding the definition of instantaneous frequency (IF) in non-stationary data [48]. One of the major advancements in understanding IF was proposed by Gabor *et.al* [48]. Gabor generated a complex analytic signal, $z(t)$, from a real one, $S(t)$, by finding the Fourier transform of $S(t)$, suppressing its negative frequencies, and then multiplying the amplitudes of the positive ones by two. In the time domain, the procedure is as follows:

$$z(t) = S(t) + jHT[S(t)] \quad (3.1)$$

$$= am(t)e^{j\theta(t)} \quad (3.2)$$

where $HT[S(t)]$ is the Hilbert transform of $S(t)$, $am(t)$ is the instantaneous amplitude, and $\theta(t)$ is the instantaneous phase as given in (3.3).

$$\begin{aligned} HT[S(t)] &= \frac{1}{\pi} p.v. \int_{-\infty}^{+\infty} \frac{S(t')}{t-t'} dt' \\ am(t) &= \sqrt{S^2(t) + HT[S(t)]^2} \\ \theta(t) &= \arctan\left(\frac{HT[S(t)]}{S(t)}\right) \end{aligned} \quad (3.3)$$

where p.v. is the principle value. Note that $HT[S(t)]$ is out of phase with $S(t)$ by $(\pi/2)$ in essence canceling the negative frequency component of the analytic signal. Consequently, the instantaneous frequency is defined as follows (3.4):

$$\omega = \frac{d\theta(t)}{dt} \quad (3.4)$$

According to this definition, the instantaneous frequency is a single value function of time. It can only represent one component at any given time. This does not make sense for multi-component signals (signals with more than one frequency for a given instance of time). To show this, a linear chirp signal is simulated by the following equation (3.5).

$$\begin{aligned} S_{chirp}(t) &= A \cos(2\pi f(t)t + \theta) \\ f(t) &= \frac{k}{2}t + f_0 \\ k &= \frac{f_1 - f_0}{T} \end{aligned} \quad (3.5)$$

where the frequency is initially set at $f_0=100$ Hz, and increases to $f_1=200$ Hz after $T=1$ s. The simulated chirp contained one frequency for every instance of time, as can be visualized in Figure 3.1. It is said to be mono-component.

Figure 3.1. a) Mono-component linear chirp signal, $S(t)$. b) Spectrogram of linear mono-component chirp signal, $S(t)$

In this case, the instantaneous frequency correctly extracts the varying frequencies in time as can be seen in Figure 3.2.

Figure 3.2. Extracted instantaneous frequency of a mono-component linear chirp signal

Now consider a signal defined as the sum of two linear chirps simulated by the following equation (3.6):

$$\begin{aligned}
 S_{chirp}(t) &= A \cos(2\pi f_1(t)t + \theta) + A \cos(2\pi f_2(t)t + \theta) \\
 f_1(t) &= \frac{f_1 - f_0}{2T}t + f_0 \\
 f_2(t) &= \frac{f_4 - f_3}{2T}t + f_3
 \end{aligned} \tag{3.6}$$

with $f_0=100$ Hz, $f_1=200$ Hz, $f_3=300$ Hz, and $f_4=400$ Hz with $T=1$ s.

Note that there are two components at each time-point in the spectrogram of the simulated signal, as seen in Figure 3.3.b. In such a scenario the signal is referred to as a multi-component signal.

Figure 3.3. **a)** Multi-component $S(t)$ (sum of two linear chirps). **b)** Spectrogram of multi-component $S(t)$

The instantaneous frequency estimate does not give any physical meaning for such a signal as can be seen in Figure 3.4 below.

Figure 3.4. Extracted instantaneous frequency of sum of two linear chirps

To resolve this, Huang proposed decomposing multi-component data into a group of functions known as intrinsic mode functions (IMF) which satisfy the following conditions [47]:

- “The number of extrema and the number of zero crossings must either be equal or differ at most by one;”
- “At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.”

The IMF in each cycle, defined by the zero crossings, involves only one mode of oscillation; that is, an IMF satisfies the mono-component condition. The IMF is not restricted to a narrow band range but can represent amplitude and frequency modulated signals.

The EMD algorithm extracts intrinsic mode functions through a sifting process which works as follows:

- Identify all the local extrema in the data.
- Connect all local maxima by a cubic spline as the upper envelope.
- Connect all local minima by a cubic spline as the lower envelope.

The mean of both envelopes is designated as μ_1 and is subtracted by the original discrete signal, $S[n]$, to extract the first component $h_1[n]$ (3.7).

$$h_1[n] = S[n] - \mu_1 \quad (3.7)$$

Ideally $h_1[n]$ is the first IMF; however, overshoots and undershoots in the cubic spline fit require a sifting process to eliminate riding waves. In the second sifting step, $h_1[n]$ is treated as the new data whose envelopes and mean are extracted (3.8).

$$h_{11}[n] = h_1[n] - \mu_{11} \quad (3.8)$$

The sifting process is repeated g times (3.9) until $h_{1g}[n]$ becomes an IMF after a stopping criteria is met (3.10).

$$h_{1g}[n] = h_{1(g-1)}[n] - \mu_{1g} \quad (3.9)$$

In the seminal paper [47], the authors proposed to stop the sifting process when the ratio as defined below in (3.10) was less than a chosen threshold, η .

$$SD = \sum_{n=0}^N \frac{|(h_{1(g-1)}[n] - h_{1g}[n])|^2}{\|h_{1(g-1)}[n]\|^2} \quad (3.10)$$

The procedure is stopped when $SD \leq \eta$.

Unlike Fourier and wavelet-based methods, this sifting process is completely data-driven and does not require a predefined set of basis functions to represent the signal [49]. The most significant drawback in EMD, however, is mode mixing where one IMF may contain signals of different scales or where one scale is present in more than one IMF [50]. To resolve this issue, [51] proposed ensemble EMD (EEMD), a noise-assisted EMD algorithm. Steps of EEMD are as follows [50]:

1. Add a white noise series to the original signal.
2. Decompose the signal with added white noise into IMFs using EMD.
3. Repeat steps (1) and (2) for N trials with different white noise realizations.
4. Obtain the corresponding IMF components of the decompositions.
5. Average the IMFs obtained from the ensemble of trials to obtain the true IMF.

This method reassigns the frequency bands onto proper scales referenced by the white noise [50]. Repeating the decomposition N times with different noise realizations ensures that the ensemble mean cancels the noise in the final IMF estimation. Figure 3.5 shows the first IMF of a sample simulated spectrum contaminated with spikes extracted using EMD, and EEMD.

Figure 3.5. a) Simulated Raman spectrum contaminated with spikes. b) IMF_1 extracted using EMD. c) IMF_1 extracted using EEMD

While both decompositions contain the simulated spikes, the IMF signal extracted using EMD contains broader Raman peak profiles superimposed to the signal. On the other hand, the signal extracted using EEMD completely extracts the high frequency spikes, making them a good candidate for spike detection and removal.

Cosmic spikes are high frequency components with narrow bandwidth relative to Raman peaks. Harmonics relating to the spikes are therefore isolated in the first IMF (IMF_1) after EEMD decomposition. The proposed algorithm uses the principles of energy detection and EEMD to detect and remove cosmic spikes. The next section outlines the steps of the proposed algorithm.

3.2 The proposed algorithm

The first IMF of a Raman spectrum contaminated with cosmic ray spikes will comprise both spikes and white noise. If the spectrum does not contain any spikes, however, then its first IMF will only contain white noise. Under these two scenarios, in the discrete domain, the $IMF_{e1}[n]$ of each individual spectrum, $S_e[n]$, can be given as (3.11):

$$IMF_{e1}[n] = \begin{cases} W_e[n] & : H_0 \\ W_e[n] + Q_e[n] & : H_1 \end{cases} \quad (3.11)$$

where e is the spectrum number ranging from 1, 2, $\dots, k$ with k being the total number of available spectra, n representing the pixel number (a point in the spectrum where each pixel represents one wavenumber), N representing the number of pixels (number of points in the spectrum), $W_e[n]$ representing a signal with only white noise, and $Q_e[n]$ representing a signal with only cosmic spikes. H_0 is the null hypothesis that $IMF_{e1}[n]$ contains only white noise as it represents the decomposition of a spike-free spectrum, while H_1 is the alternative hypothesis that $IMF_{e1}[n]$ contains both white noise and spikes as it represents the decomposition of a spectrum containing spikes.

The proposed algorithm follows the following steps based on this hypothesis formulation:

1. **Step 1: Decomposition of a subset of the total available spectra using EEMD:**

A subset of the total available spectra are decomposed using EEMD in order to extract their corresponding first IMF, $IMF_{m1}[n]$, for $m=1, 2, \dots, q$ where m is the spectrum number, and q is the number of spectra in the selected subset.

2. **Step 2: Application of energy detection algorithm on $IMF_{m1}[n]$:**

Energy detection on each $IMF_{m1}[n]$ is applied. The energy detector measures the energy, given by (3.12), associated with $IMF_{m1}[n]$ over a specified window length (L). This measurement is then compared with an appropriately selected threshold, η , to decide the presence or absence of cosmic spikes.

$$\Lambda = \sum_{n=i \times L}^{(i+1) \times L - 1} |IMF_{m1}[n]|^2 \quad (3.12)$$

for $i=0, 1, 2, \dots, M-1$, where Λ is the energy detection test statistic to be compared with the threshold. It is assumed that there are M windows of length L in the total N pixels.

After the test statistic is calculated, a detection vector, $c_m[n]$, is defined by to determine whether a cosmic spike event occurred or not. $c_m[n]$ is populated by comparing the test statistic to a selected threshold as follows (3.13):

$$c_m[n] = \begin{cases} 1 & \text{if } \Lambda \geq \eta \\ 0 & \text{if } \Lambda < \eta \end{cases} \quad (3.13)$$

for $n = i \times L$ to $(i + 1) \times L - 1$. For $i=0$, Λ is calculated by summing the absolute squared IMF signal from 0 to L . If $\Lambda \geq \eta$, then $c_m[n]$ for $n=0, 1, \dots, L - 1$ will be set equal to one, else, it will be set equal to zero. Λ is then calculated over the next portion of the IMF signal, at $i=1$, from L to $2L - 1$, and the process is repeated until the entire IMF signal has been processed. After, cosmic spike pixel locations are identified as the location of the maximum peak inside pixels with $c_m[n] = 1$ as can be seen in Figure 3.6.

The primary aim of an energy detector threshold is to give an estimate of the noise floor when the noise variance of the signal is unknown. Median based thresholds are typically used due to their resilience to outliers, when the majority of the signal is noise as in the first IMF [52]. The threshold η is therefore selected by median absolute deviation (MAD) estimator given by (3.14).

$$\eta = \frac{\text{median}|IMF_{m1}[n]|}{\sqrt{2}erf^{-1}(0.5)} \quad (3.14)$$

where

$$\sqrt{2}erf^{-1}(0.5) = 0.6745$$

and N is the number of points in the spectrum. The median absolute deviation is a measure of dispersion that is resilient to outliers in data, and therefore is preferred over standard deviation for estimating the noise floor.

Depending on the L set, $c_m[n]$ may include data corresponding to the pure Raman spectrum (spike-free regions). $c_m[n]$ is therefore further localized to the local minimum points around the peak of the spikes. The localized event detection vector will be referred to as $c_{localized}[n]$. Figure 3.6 shows the detection vector before and after localization. For visualization purposes the detection vectors were multiplied by 2000.

Figure 3.6. a) $c_m[n]$ of Raman spectrum from Donor 52 b) $c_{mlocalized}[n]$ of Raman spectrum from Donor 52

3. Step 3: Suppression of sharp Raman peaks:

After application of the energy detection algorithm on all q spectra, the localized detection vector $c_{mlocalized}[n]$ is automatically corrected to remove potential false detections. The corrected detection vector will be referred to as $c_{mcorrected}[n]$. False detections occur due to some Raman peaks having sharp peaks with similar characteristics to the cosmic ray spikes. They are suppressed before spike removal through the following majority voting scheme described below:

- If $c_{mlocalized}[n]$ equals 1 at pixels $n = i \times L, \dots, (i + 1) \times L - 1$, for all spectra in the subset, then $c_{ecorrected}[n]$ at these pixels are set to zero, where $e=1, \dots, k$ total spectra available. The logic behind this is that having spikes at the same pixels for different spectra is a rare event given the random nature of spike generation. Therefore, if all spectra in the subset have a spike at the same location, then it is most likely a sharp Raman peak that will be falsely detected as a spike for all spectra in the dataset. At all other pixels, $c_{ecorrected}[n]$ is set equal to $c_{elocalized}[n]$.

4. Step 4: Cosmic spike removal:

After the falsely detected sharp Raman peaks are suppressed, data values in $S_e[n]$ corresponding to $c_{ecorrected}[n]=1$ are replaced by interpolated approximate values

of $S_e[n]$ using a piecewise cubic hermite interpolating polynomial for all k spectra. The filtered spectrum, $S_{e\text{filtered}}[n]$, is then checked to see whether another iteration of the algorithm is needed. If no false detections exist, $c_{e\text{localized}}[n]$ is equated to $c_{e\text{corrected}}[n]$.

5. Step 5: Iteration check:

Figure 3.7 shows the effect of setting the energy window length, L , to a large value. In this case an event detection window may span two spikes. In that case, the localized detection vector will only cover one spike while disregarding other spanned spikes. To compensate for this missed detection, the algorithm requires to go through another iteration to detect and filter the missed spikes. The number of iterations is determined as follows:

- After spike removal, reapply steps 1 with all filtered spectra, $S_{e\text{filtered}}[n]$ as input.
- Reapply step 2 for all spectra. If $c_{e\text{corrected}}[n]=0$ for all n , then accept the null hypothesis that $IMF_1[n]$ of $S_{e\text{filtered}}[n]$ does not contain any more spikes. In that case all spikes of spectrum e have been filtered and there is no need for another iteration.
- If $c_{e\text{corrected}}[n]$ is not zero for all n , reapply step 4 on $S_{e\text{filtered}}[n]$.

Figure 3.7. a) Event detection and localized event detection windows for $L=150$. b) $S_{\text{filtered}}[n]$ after first iteration of algorithm. c) $S_{\text{filtered}}[n]$ after second iteration of algorithm.

3.3 Experiments on datasets simulated

In order to test the proposed algorithm four different datasets were created. The first three contained 50 simulated spectra, with 4 cosmic spikes. Each spectrum, $S_e[n]$, consisted of the following parts (3.15):

$$S_e[n] = R_e[n] + B_e[n] + Q_e[n] \quad (3.15)$$

where n are the pixel number ranging from 0 to 3101 in steps of 1, e is the spectrum number, $R_e[n]$ represented the pure Raman spectrum modeled by the sum of symmetrical, positively valued, Gaussian peaks given by (3.16)

$$R_e[n] = A_n \times \exp\left[\left(\frac{n - r_n}{wx_n}\right)^2\right] \quad (3.16)$$

where r_n is the position of the peak, wx_n is the width of the peak at the full width half-maximum (FWHM), and A_n is the total area under the curve above the background. $B_e[n]$ represented the background modeled by a Gaussian function given by (3.17), where C_1 and C_2 govern the curve's peak height, μ_1 , and μ_2 govern the center of the peak, σ_1 , and σ_2 control the width of the peak.

$$B_e[n] = C_1 \exp\left[-\frac{(n - \mu_1)^2}{2\sigma_1^2}\right] + C_2 \exp\left[-\frac{(n - \mu_2)^2}{2\sigma_2^2}\right] \quad (3.17)$$

For all datasets C_1 and C_2 were set to 1. μ_1 and μ_2 varied from 246 to 834 in steps of 12, while σ_1 varied from 386 to 974 in steps of 12, and σ_2 varied from 586 to 1174 in steps of 12. $Q_e[n]$ represented four randomly simulated cosmic ray spikes modeled using (3.16).

3.3.1 Simulated datasets

The first dataset cD1 ('c' stands for cosmic ray) contained Raman Spectra with sharp Raman spectral features whose bandwidth is comparable to that of the cosmic spikes. The purpose of this dataset is to check the ability of the proposed algorithm in distinguishing between sharp Raman peaks and sharp cosmic ray spikes in the detection phase of the algorithm.

The pure Raman spectrum, $R_e[n]$, for each spectrum in cD1 was simulated based on the values in Table 3.1.

Table 3.1. Pure Raman spectrum simulation parameters for dataset cD1

r_n	wx_n	A_n
500	44	23
555	50	34
1130	62	63
700	74	83
1002	44	20
1027	59	9
1071	55	42
1259	42	56
1265	33	44
1302	42	32
1341	45	37
1444	0.0001	20
1652	32	23
2000	43	23
2500	31	43
2210	54	34
1214	40	45
1700	43	20
1814	54	40
1953	93	59
2230	42	35
1656	43	52
1553	44	23
943	44	41

As mentioned, for each simulated pure spectrum, four spikes were randomly generated. Each spectrum also had a shifted version of the background superimposed, as described by the parameters of (3.17), in order to allow for variation in the spectra. Figure 3.8 shows the spectra in cD1 before spikes were randomly generated and added.

Figure 3.8. a) 50 simulated spectra of dataset cD1 before adding spikes. b) A simulated spectrum from dataset cD1 before adding spikes

As can be seen, spectra in dataset cD1 had one sharp Raman peak at pixel 1444. It is much narrower than the other Raman peaks and is similar to that of cosmic spikes. The proposed algorithm initially detects this sharp peak as a spike. This is automatically corrected by the algorithm's majority voting scheme employed in Step 4 of the algorithm on a subset of the simulated spectra. Figure 3.9 shows a subset of six different spectra with randomly generated spikes. The figure also shows the localized detection vector $c_{m\text{localized}}[n]$ of each spectrum.

Figure 3.9. Subset six spectra from dataset cD1, and the localized detection vector extracted on each of the 6 spectra before majority vote correction

As can be seen, all six spectra falsely detected the Raman peak at pixel 1444 as a spike. Each localized event detection vector is corrected, given by $c_{m,corrected}[n]$ as not to include this common peak as shown in Figure 3.10 below.

Figure 3.10. Subset six spectra from dataset cD1, and the corrected event detection vector after majority vote correction

After correction of the event detection vector, the spikes are removed as discussed. EEMD is then reapplied on each filtered spectrum to check whether all spikes have been removed for all spectra available (as described in Step 5 of the algorithm). If all the spikes have been filtered then the first IMF of the filtered spectrum will contain only noise, except at the pixels corresponding to the sharp Raman peak. This peak will be suppressed by the corrected detection vector as seen in Figure 3.11 below.

Figure 3.11. $IMF_1[n]$ of one filtered spectrum after spike removal and the corresponding detection vectors

The second dataset (cD2) contained Raman peaks whose spectral features were broad relative to the cosmic spikes. The same values in Table 3.1 were used to simulate the pure Raman spectrum with the exception of the peak at pixel 1444, whose width was changed to 40. Figure 3.12 shows an example spectrum from dataset cD2 before and after spikes are added.

Figure 3.12. A simulated spectrum from dataset cD2 **a)** before adding spikes **b)** after adding spikes

Typical spikes range between 1 to 3 wavenumbers. The first two datasets (cD1 and cD2) had spikes varying from 1 to 3 pixels to simulate spectra with each wavenumber sampled with one pixel (one pixel represents a wavenumber). The third dataset (cD3) contained broader spikes varying from 4 to 7 pixels to simulate spectra with each wavenumber spanning multiple pixels as may occur when the spectrum is highly resolved. Figure 3.13 shows an example spike from dataset cD2 and cD3 with spikes spanning 2 and 4 pixels respectively.

Figure 3.13. a) Spike from a dataset cD2 spectrum covering 2 pixels. b). Spike from a dataset cD3 spectrum covering 4 pixels

Dataset cD4 contained 20 Raman Spectra of whole blood cells not exposed to radiation from Donor 125. The purpose of this real (not simulated) dataset is to first check the algorithm's ability to detect whether the spectrum contains spikes or not. Its second purpose is to check its ability to detect real spikes on spectra produced under real experimental conditions. Figure 3.14 shows the spike locations in each of the 20 spectra.

Figure 3.14. Dot plot showing location of spikes, if any, in dataset cD4

3.3.2 Results and Discussion

The performance of the proposed algorithm was evaluated using the four datasets discussed above. It was compared to median filtering, the algorithm proposed by Katsumoto and Ozaki [19], the algorithm proposed by Tian [15], and the algorithm proposed by Zhang-Henson [21] discussed in the literature review. Median-filtering is included in the comparisons as it is simple, and used in most Raman software packages for spike removal.

Parameter selection

Preliminary analysis was done to optimize the parameters that produced the lowest NMSE, given by (3.18), for all algorithms.

$$NMSE((R + B), (S_{filtered})) = \frac{\sum_{n=1}^N ((R_e[n] + B_e[n]) - S_{e_{filtered}}[n])^2}{N \sum_{n=1}^N (R_e[n] + B_e[n])^2} \quad (3.18)$$

where $R_e[n] + B_e[n]$ is the simulated spectrum before spikes are added (pure Raman spectrum and background), N is the length of the spectrum, and $S_{e_{filtered}}[n]$ is the filtered spectrum

with spikes removed. For the proposed algorithm, the only parameter that requires adjusting is L . Setting it to three was sufficient to detect all spikes. Setting it higher, depending on how close the spikes were simulated to each other, required additional iterations to fully remove all spikes. In general, to avoid missed detections due to a large L , the number of iterations should equal the number of spikes in the spectrum.

Performance comparison of algorithms

For dataset cD1, the median filter, Katsumoto-Ozaki, and Zhang-Henson algorithms were not able to differentiate between cosmic spikes and sharp Raman peaks. This is expected as the algorithms do not take advantage of multiple spectra in the removal process. The proposed algorithm, on the other, is capable of doing so through the majority voting scheme described in the proposed methodology. As well, it only produced negligible bias at pixels corresponding to the spikes. The proposed algorithm also fully preserved the pure Raman spectrum as can be seen in Figure 3.15 which shows a difference spectrum for the three algorithms applied to one spectrum in dataset cD1. The difference spectrum shows the difference between the spike-free simulated spectra, $R[n]+B[n]$, and the corresponding filtered spectrum, $S_{e\text{filtered}}[n]$, using the Katsumoto-Ozaki, Zhang-Henson, and proposed algorithms. The proposed algorithm was far better than Zhang-Henson and Katsumoto-Ozaki algorithms. Amongst the three algorithms, Katsumoto-Ozaki algorithm performed the worst for this dataset. Zhang-Henson also appeared to removed part of a true Raman spike as well.

Figure 3.15. Difference between a simulated pure spectrum in dataset cD1 and spectrum after spikes are removed using different algorithms, **a)** Katsumoto-Ozaki (Kat-Oz) algorithm **b)** Zhang-Henson (ZH) algorithm **c)** The proposed algorithm

A window size of 3, $L_{med} = 3$, was sufficient to remove all sharp spikes (dataset cD1 and cD2) using the median filter. This also removed the sharp Raman peak in dataset cD1 causing median filter to produce the highest median NMSE as seen in Figure 3.18. For dataset cD2, where there were no sharp peaks, the median filter produced the second best average NMSE as can be seen in Figure 3.19. For the broader spikes in cD3, a larger window size for the median filter was needed to filter all the spikes, leading to a larger spectral bias as can be seen in Figure 3.16 which shows the effect of increasing the median filter's window size on the difference spectrum. This caused the median filter to produce the worst median NMSE in this dataset.

Figure 3.16. Difference between a simulated pure spectrum in dataset cD3 and spectrum after spikes are removed using a median filter with different window sizes. **a)** $L_{med} = 3$. **b)** $L_{med} = 9$

For the selected parameters, one iteration of the Katsumoto-Ozaki algorithm did not fully remove either the narrow (cD1 and cD2) nor broad (cD3) spikes, leading to the worst median NMSE in dataset cD2, and second worst in cD1 and cD3. Even though spikes were not fully removed, the algorithm was able to preserve the pure Raman spectrum at the selected optimum values done in the preliminary analysis with $\eta_{KAT-Oz} = 1$, and $L_{KAT-Oz} = 23$ where the moving average window span is $2L_{KAT-Oz} + 1$. Figure 3.17 shows the effect of selecting a window size and threshold different than the optimal values for dataset cD2. As seen, decreasing the threshold had the effect of filtering non spike bands. Decreasing the window size below the optimum value over filtered the spikes, while increasing it started filtering the broader Raman peaks.

Figure 3.17. Difference between a simulated pure spectrum ($R[n]$) in dataset cD2 and spectrum after spikes are removed using the Katsumoto-Ozaki algorithm with median filter of **a)** $L_{KAT-Oz} = 3, 5, 23$, and 25 for $\eta_{KAT-Oz} = 1$. **b)** $L_{KAT-Oz} = 3, 5, 23$, and 25 for $\eta_{KAT-Oz} = 0.5$

The Zhang-Henson algorithm did not preserve the pure Raman spectrum as well as the Katsumoto-Ozaki algorithm, except at the sharp Raman peak where the bias was much lower as can be visualized in Figure 3.15. This contributed to the lower average NMSE values in comparison to the Katsumoto-Ozaki algorithm in dataset cD1 as seen in Figure 3.18. The proposed algorithm was the only one able to fully preserve the pure Raman spectrum for all datasets without any bias at the Raman bands. This is clear in the visual comparison of the difference spectra in Figure 3.15 where the difference between the filtered and simulated spectrum at the Raman bands was zero. In addition, the proposed algorithm best removed the spikes. This can be seen from the boxplots shown in Figures 3.18, 3.19, and 3.20. Not only did the proposed algorithm produce the lowest average NMSE for all datasets, it also produced the lowest interquartile range, suggesting it is more adept to variation in spectra. This is because the proposed algorithm depends on one parameter, L , which only affects the number of iterations needed for the algorithm to remove all spikes.

Figure 3.18. NMSE box plot for dataset cD1

Figure 3.19. NMSE box plot for dataset cD2

Figure 3.20. NMSE box plot for dataset cD3

In dataset cD4, all algorithms except that proposed by Katsumoto-Ozaki fully removed the cosmic spikes. Since the real data in dataset cD4 did not contain a gold standard, visual assessment on the spike-free spectra was employed to check each algorithm's ability to detect whether spikes are present or not. If the spectra did not contain spikes no filtering should be done, and hence the difference spectrum should be zero. This gives an idea to whether the algorithms are able to automatically preserve the pure Raman spectrum or not. In terms of spectrum preservation, the proposed algorithm best preserved the pure Raman spectrum. Compared with the other algorithms, the proposed algorithm was the only one that did no filtering on the spike-free spectrum. This can be visually seen in Figure 3.21 and Figure 3.22, which shows an example difference spectrum using the proposed algorithm, median filtering, the algorithm proposed by Tian, Katsumoto-Ozaki, and Zhang-Henson. For the algorithm proposed by Tian, it was noted that the choice of wavelet, unlike suggested by [15], played a major role in the amount of random noise detected as spikes. As seen in the figure, spike filtering heavily depended on the wavelet chosen which may vary from spectra to spectra. In general, the proposed algorithm outperformed the other methods in term of preserving non-spike bands.

Figure 3.21. Difference spectrum for proposed algorithm and Tian algorithm. **a)** Raw Raman spectrum without any spikes. **b)** Tian algorithm using dB3 wavelet. **c)** Tian algorithm using sym2 wavelet. **d)** Proposed algorithm

Figure 3.22. Difference spectrum for **a)** Katsumoto-Ozaki algorithm **b)** Median filtering. **c)** Zhang-Henson

The proposed algorithm provides a spike detection methodology capable of efficiently detecting spikes such that removing them will maintain the integrity of the Raman spectrum.

After spike detection and localization, any suitable preprocessing method could be used to replace values inside the event detection window containing spikes. Figure 3.23 shows Raman spectra in dataset cD4 before and after spike removal using the proposed algorithm. As can be visually assessed, all spikes have been successfully removed.

Figure 3.23. Spectra in cD4 before and after spike removal using proposed algorithm

3.4 Conclusion

This chapter proposed a novel cosmic spike technique that was able to detect and successfully remove spikes. It was able to remove spikes of small and large bandwidths, and was capable of automatically suppressing common Raman peaks falsely detected as spikes. The proposed technique uses the EEMD process which does not have limitations that wavelet decomposition comes with. The main novelty in the proposed cosmic spike preprocessing technique is its ability to detect spike locations in order to preserve the Raman bands. Once detected, the spikes can be removed using any other method. This ensures that the Raman bands are not distorted as in current methods.

In all the comparisons done, the proposed algorithm produced the lowest NMSE due to its ability to preserve the Raman bands on both simulated and real spectra.

Chapter 4: Background estimation and removal

4.1 Introduction

Unlike cosmic ray spikes, fluorescence background noise manifests as broad low frequency noise. The intensities associated with this type of noise are orders of magnitude higher than that of the Raman scattering, masking Raman intensities. In this chapter two novel algorithms are proposed based on the principles of penalized least squares, and the principles of minimizing a non quadratic cost function. Penalized least squares was first proposed for smoothing by Whittaker. Later, Eilers [6] modified its use for baseline detection and removal. The penalized least squares algorithm provides a least squares roughness penalty smooth which balances between roughness (E) and fidelity (J) to the data. Given a vector $\mathbf{S}$ representing a Raman spectrum, and another vector $\mathbf{z}$ representing the estimate of the spectrum, fidelity of the estimate can be estimated by the sum square differences between $\mathbf{S}$ and $\mathbf{z}$ given by (4.1):

$$J = \sum_{i=1}^N (S_i - z_i)^2 \quad (4.1)$$

where N is the length of the spectrum and S_i and z_i represents the elements of the vectors $\mathbf{S}$ and $\mathbf{z}$, respectively.

The roughness of the fit E is described by the k^{th} order differences is given by (4.2):

$$E = \sum_{i=1}^{N-k} (\Delta^k z_i)^2 \quad (4.2)$$

where for $k=2$,

$$(\Delta^2 z_i) = \Delta(\Delta z_i) = z_i - 2z_{i-1} + z_{i-2} \quad (4.3)$$

The Whittaker smoother, also known as penalized least squares algorithm, is the summation of these two terms given by (4.4):

$$H = J + \varepsilon E$$

$$H = \sum_{i=1}^N (S_i - z_i)^2 + \varepsilon \sum_{i=1}^{N-k} (\Delta^k z_i)^2 \quad (4.4)$$

Note that ε determines the strength of contribution of E to H . It provides a balance between the fidelity and roughness. A large ε provides more smoothing, while a smaller one gives more importance to fitting the data.

For fluorescence background estimation, Eilers proposed the Asymmetric least squares algorithm (AsLs) by introducing a weight vector, $\mathbf{W}$, on the cost function of the penalized least squares, in order to give more weight to negative residuals given that the positive ones give more influence to the Raman peaks. The weight vector is governed by a scalar asymmetry parameter p which iteratively computes the weights by the following condition:

$$\begin{cases} w_i = p, S_i > z_i \\ w_i = 1 - p, \text{otherwise} \end{cases} \quad (4.5)$$

where $0 < p < 1$. The purpose of p is to give a small weight where a Raman peak is present (represented by positive residuals), and a large weight of $1-p$ for negative residuals (where peaks are not present). For background estimation, the cost function in (4.4) is changed to (4.6).

$$H = \sum_{i=1}^N w_i (S_i - z_i)^2 + \varepsilon \sum_{i=1}^{N-k} (\Delta^k z_i)^2 \quad (4.6)$$

In vector form it is given by (4.7):

$$H = (\mathbf{S} - \mathbf{z})' \mathbf{W} (\mathbf{S} - \mathbf{z}) + \varepsilon \mathbf{z}' \mathbf{D}' \mathbf{D} \mathbf{z} \quad (4.7)$$

with $\mathbf{W} = \text{diag}(w)$, and $\mathbf{D}$ is a matrix such that $\mathbf{D}\mathbf{z} = \Delta^2 \mathbf{z}$ for $k=2$. Note that $'$ denotes the transpose operation.

Minimizing the equation with respect to $\mathbf{z}$ we get (4.8):

$$\hat{\mathbf{z}} = (\mathbf{W} + \varepsilon \mathbf{D}' \mathbf{D})^{-1} \mathbf{W} \mathbf{S} \quad (4.8)$$

The final estimated background is found by iteratively re-evaluating the penalized least squares function, each iteration updating the weights until they converge.

Following the principles of asymmetric least squares, Mazet addressed the background estimation problem as a polynomial fitting where the coefficients were estimated using two different non-quadratic costs, the asymmetric Huber (HB), and asymmetric truncated quadratic (TQ) cost functions [3]. In classical least squares, the regular quadratic cost function (2.19) assigns a high cost to large residuals, x , far away from the true background, thus biasing the coefficient estimation. Instead, Mazet considered the undermentioned two cost functions that provide a lower cost to large values:

- Asymmetric Huber function (HB) (4.9)

$$\forall x \in \mathbb{R}, \varphi(x) = \begin{cases} x^2 & \text{if } x \geq s, \\ 2sx - s^2 & \text{otherwise} \end{cases} \quad (4.9)$$

- Asymmetric truncated quadratic function (TQ) (4.10)

$$\forall x \in \mathbb{R}, \varphi(x) = \begin{cases} x^2 & \text{if } x \geq s, \\ s^2 & \text{otherwise} \end{cases} \quad (4.10)$$

where x are the residuals, and s is an introduced threshold that governs the slope of the cost functions at positive residuals such that the Raman peaks are not influenced in the background estimation. Figure 4.1 shows the asymmetric cost functions discussed with $s=2$.

Figure 4.1. Asymmetric cost functions adapted from [3] at $s=2$

These two cost functions assign a low cost to positive residuals (positive residuals represent the positive Raman peaks), but are quadratic for negative residuals. The estimated coefficients for both these functions are found iteratively using half-quadratic minimization. The objective function for this minimization is given by (4.11):

$$H(\beta, \mathbf{d}) = \sum_{k=1}^N \frac{1}{2} ((S_k - z_k - d_k)^2 + \psi(d_k)) \quad (4.11)$$

where $\mathbf{S}$ is the spectrum, $\mathbf{z}$ is the least-squares background estimate, and ψ is defined as (4.12):

$$\psi(d) = \sup_x (\varphi(x) - (x - d)^2/2) \quad (4.12)$$

The result of the minimization can be interpreted as the least square estimator of a signal, $\mathbf{S} + \mathbf{d}$, where $\mathbf{d}$ is an auxiliary variable introduced by the half quadratic minimization. For each iteration, $\mathbf{d}$ is found by the following formula (4.13).

$$d_k = -(S_k - z_k) + \varphi'(S_k - z_k) \quad (4.13)$$

for $k = 1, 2, \dots, N$ with N being the length of the spectrum. with

$$\varphi'(x) = \begin{cases} 2x & \text{if } x < s, \\ 2s & \text{otherwise,} \end{cases} \quad (4.14)$$

for the asymmetric Huber cost function, and

$$\varphi'(x) = \begin{cases} 2x & \text{if } x < s, \\ 0 & \text{otherwise,} \end{cases} \quad (4.15)$$

for the asymmetric truncated quadratic function. $\mathbf{d}$ is used to recalculate the coefficients which are in turn used to re-estimate the background. This process is repeated until a certain decision criteria is met.

In this chapter, two novel background removal techniques are proposed based on the principles of weighted asymmetric least squares and minimizing a non-quadratic cost function.

4.2 Weighted penalized least squares approximation with a non-quadratic cost function

In this work, the following equation is minimized (4.16):

$$H = \sum_{i=1}^N w_i \varphi(\cdot) + \varepsilon \sum_{i=1}^{N-2} (\Delta^2 z_i)^2 \quad (4.16)$$

with $\varphi(\cdot)$ representing two non quadratic cost functions. The two algorithms proposed are the weighted Huber penalized least squares (WHPLS) algorithm, where $\varphi(\cdot)$ is an asymmetric Huber cost function, and the weighted truncated quadratic penalized least squares (WTQPLS) algorithm, where $\varphi(\cdot)$ is an asymmetric truncated quadratic cost function. For both algorithms, the estimated background, $\hat{z}$, is given by (4.17):

$$\hat{z} = (W + \varepsilon D' D)^{-1} W (\mathbf{S} + \mathbf{d}) \quad (4.17)$$

where $\mathbf{d}$ is found by (4.13), as previously discussed. In the proposed algorithm, the cost function is governed by 3 parameters, p , s , and ε . The motivation behind this objective function is to make use of the non-quadratic cost function's suitability for background estimation, as discussed in [3], and the smoothing penalty of the AsLs algorithm as discussed in [6]. At $s=0$, both the WHPLS and WTQPLS algorithms converge to the same asymmetric quadratic cost function with the addition of the roughness penalty.

4.3 Parameter optimization of the proposed algorithms

4.3.1 Effect of spectrum peak width on algorithm parameters

Ten spectra were simulated based on (3.16) in order to study how spectrum peak width affected background estimation for a given set of parameters. The peak width, w_x , of each spectrum varied from 10 to 100 in steps of 10. The Raman spectrum was simulated with one Raman peak simulated at $r = 1500$, amplitude of $A = 20$, and a spectral range, r , covering 3101 pixels. The background added was a Gaussian signal given by (3.17) with $C_1 = 20$, $C_2 = 0$, $\mu_1 = 1100$, and $\sigma_1 = 1300$. Figure 4.2 shows a plot of the NMSE between the estimated and simulated background for a single Raman peak with varying peak widths. For this plot, the parameters were randomly selected to values $\varepsilon = 10^4$, $p = 0.01$, and $s = 0.01$ for both algorithms. The trend observed was that as the peak width of a simulated spectrum increased, so did the NMSE.

Figure 4.2. Effect of peak width on NMSE for both algorithms using randomly chosen values

This can be also seen with the help of the plot in Figure 4.3, which shows the NMSE as a function of all the three parameters for the WTQPLS algorithm. In this example, the NMSE was calculated for background estimation carried out with all combinations of s and p varying from 10^{-4} to 0.9 in steps of 0.045, and ε varying from 10^2 to 10^3 in steps of 45.

Figure 4.3. Effect of peak width on NMSE as a function of all the three parameters for spectrum with smallest simulated width ($wx=10$) and spectrum with largest simulated width ($wx=100$) for s and p varying from 10^{-4} to 0.9 in steps of 0.9 and ε varying from 10^2 to 10^3 in steps of 45

In this plot, each slice represented a different value of s . The trend observed was that for all parameter combinations, the spectrum with a wx of 10 produced a much lower NMSE than the spectrum with a higher width for both algorithms. As well, as seen in Figure 4.2, it was observed that as the width of a simulated peak increased, the WHPLS algorithm generally started to produce a higher NMSE than the WTQPLS algorithm. This is expected as the cost function of WHPLS gives a higher cost to positive deviations, and hence Raman peaks will have more influence in the background estimation with increasing peak width.

To see how each individual parameter affected background estimation, two parameters were randomly selected, with the NMSE calculated as a function of the third one. Figures 4.4a and 4.4b show a log plot with NMSE as a function of p , for $p = 0$ to 1, increasing in increments of 0.0020, $s = 0.001$, and $\varepsilon = 10^6$ for the simulated spectrum with wx set to 10, 50 and 100 respectively. Initially, for values smaller than the optimal p , NMSE for the spectrum with the largest width produced the least error. Here, the optimal p is the one which produced the lowest NMSE for the parameter values evaluated. At the optimal p , the NMSE was the lowest for the spectrum with the smallest peak width. In general, the trend was that as the width of the peak increased, a smaller p was needed to produce an optimal background estimation. This is because a spectrum with a larger peak width consequently increases the effect of positive deviations, and as a result the background estimation gets influenced more by the Raman peaks. A smaller weight must be given to these peaks for a more accurate background estimation. This can be seen in Figure 4.5 which shows the estimated background for a spectrum with a width of 10 and 100 respectively for variable p values. A p of 0.006 was able to estimate the background of the spectrum with a width of 10, while over-estimating the one with a width of 100. This overestimation was compensated by reducing p from 0.006 to 7.5e-4 in this example.

(a) Effect of width on NMSE as a function of p for WTQPLS

(b) Effect of width on NMSE as a function of p for WHPLS

Figure 4.4. Effect of width on NMSE as a function of p for both proposed algorithms

Figure 4.5. Background estimation of spectrum for $p = 0.006$, $p=7.5e-4$, and $p=0$. for **a)** $wx = 10$. **b)** $wx=100$.

The same analysis was repeated to see how the parameter s influenced the background estimation. The NMSE was calculated for $s=0$ to 1, increasing in increments of 0.0020 with ε set to 10^4 , and p set to 0.01. The general trend observed was that as the peak width of a spectrum increased, a smaller value of s was required to provide optimal background estimation. This can be seen in Figure 4.6 which shows a log plot of the NMSE as a function of s for wx set to 10 and 30 respectively.

(a) Effect of width on NMSE as a function of s for WTQPLS

(b) Effect of width on NMSE as a function of s for WHPLS

Figure 4.6. Effect of width on NMSE as a function of s for both algorithms

As can be seen, the optimal s decreased by a factor of 10 as the width increased from 10 to 30 for both algorithms. It was also noted that for the same width, the optimal s of the WTQPLS algorithm was higher than for the WHPLS algorithm. This makes sense as for the same s term, the Huber cost function will give a larger cost to positive deviations than the TQ cost function (refer to Figure 4.1). It is therefore recommended to set s to a lower value for the WHPLS to reduce peak influence in the background estimation.

To see the influence of ε on background estimation, the NMSE was calculated for $\varepsilon = 10^4$ to 10^{10} , with a total of 71 values, and 10 values for each decade of ε . Both p and s were set to 0.01. Figure 4.7 shows a log plot with the NMSE calculated as a function of ε for the WTQPLS and WHPLS algorithms respectively. The optimal ε increased with increasing peak width. Generally, for the same width, the optimal value was higher for the WHPLS algorithm than the WTQPLS algorithm. In both cases, at optimal ε values, the spectrum with the smallest width produced the least NMSE. As ε increased for both algorithms, the NMSE converged to the same NMSE for all widths. At this point, precedence in H is given to smoothing, limiting the role of both non-quadratic cost functions. Here, $(W + \varepsilon D'D)z = WS$ reduces to $D'Dz = WS/\varepsilon = 0$, leading to an estimation that approaches a straight line. This can be visualized in the Figure 4.8 where a high ε leads to a similar background estimation for the spectrum simulated with a width of 10 and 100 using the WTQPLS, WHPLS, and AsLs algorithms.

(a) Effect of width on NMSE as a function of ε for WTQPLS

(b) Effect of width on NMSE as a function of ε for WHPLS

Figure 4.7. Effect of width on NMSE as a function of ε for both algorithms

Figure 4.8. Background estimation for spectrum with peak width of **a)** width = 10. **b)** width=100 using $\epsilon = 10^9$

4.3.2 Effect of SNR on algorithm parameters

White random noise ($\sigma_n=1$) was added to the spectrum with a width of 10 in order to study the influence of SNR on parameter selection. For a constant ϵ and s , a low p parameter value tended to fit the estimated background to the negative extremities of the noise peaks, as shown in Figure 4.9.

Figure 4.9. Effect of estimating the background of a spectrum contaminated with noise with a small $p=0.001$ for ε set to 10^6 , $s=1$ and $\sigma_n=1$

As p was increased, for the same s and ε , the background estimation moved from the negative extremities of noise towards the true background as can be seen in Figure 4.10 that shows the estimated background for different values of p . If p is set too high, however, it starts to over-estimate the background. This behavior of p was similar with other algorithms that made use of variations of the AsLs algorithm [53].

Figure 4.10. Effect of estimating the background of a spectrum contaminated with noise with different values of p

Figure 4.11 shows a contaminated spectrum estimated with different ε values with $s = 1$ and p set to 0.001. For both algorithms a very high ε fitted the data to the most negative noise extremities. As ε decreased, it began to give precedence to more negative noise instances other than the extremities as can be seen in Figure 4.11. This occurs because a small ε gives more precedence to fitting to the noisy data, and not to smoothing. In general for a spectrum with low SNR, it is preferred to increase ε such that when p and s are set to push the estimate towards the real background, the estimation does not follow the non-smooth shape of the noise.

Figure 4.11. Effect of estimating the background of a spectrum contaminated with noise with different values of ε for WTQPLS and WHPLS

4.4 Simulated datasets

In order to compare both proposed algorithms' ability to estimate the background with the state-of-the-art methods, three datasets, each containing 100 simulated spectra, were created. Each simulated spectrum, $S_{eback}[n]$, consisted of the following parts (4.18):

$$S_{eback}[n] = R_{eback}[n] + B_{eback}[n] \quad (4.18)$$

where e is the spectrum number ranging from 1 to 100, $R_{eback}[n]$ is the pure background-free spectrum, and $B_{eback}[n]$ is the simulated background.

The noise-free Raman spectrum, $R_{eback}[n]$, is modeled by the sum of symmetrical, positively valued, Lorentzian peaks given by (4.19). This model is adapted from [42] where

$$R_{eback}[n] = \frac{2A_n}{\Pi} \frac{wx_n}{4(n - r_n)^2 + wx_n^2} \quad (4.19)$$

n is the pixel number ranging from 0 to 3101 in steps of 1, r_n is the position of the peak, w_{x_n} is the width of the peak at full width half-maximum (FWHM), and A_n is the peak amplitude. In each dataset the position, widths, and amplitudes were randomly selected over ranges defined in Table 4.1. All spectra spanned 3101 pixels, and had Raman peaks simulated between 600-2456 pixels.

Table 4.1. Parameters used for simulated spectra of datasets fD1, fD2 and fD3

Dataset	Num of peaks	wx range (pixels)	A range (peak intensity)	r range
fD1	5			
fD2	10	10– 40	30– 60	600 –2456
fD3	30			

Dataset fD1 comprised 100 spectra with 5 Lorentzian peaks, while datasets fD2 and fD3 comprised 100 spectra with 10 and 30 Lorentzian peaks respectively. Each simulated pure Raman spectrum was contaminated with a background modeled by the sum of a Gaussian and fifth order polynomial, given by (4.20).

$$B_{eback}[n] = P_e[n] + C_1 \exp\left[-\frac{(n - \mu)^2}{2\sigma^2}\right] \quad (4.20)$$

where $P_e[n]$, adapted from [42], is given by (4.21):

$$P_e[n] = n^5 - 0.148n^4 + 1.894 \times 10^{-6}n^3 - 1.189 \times 10^{-9}n^2 + 3.12 \times 10^{-13}n - 2.63 \times 10^{-17} \quad (4.21)$$

with $C_1 = 1$, $\mu = 1500$, $\sigma = 200$, and n evaluated for 0 to 3101 pixels in increments of 1.

The proposed algorithms are compared with state-of-the-art background techniques closely related to them. These include:

- Minimizing an asymmetric Huber cost function for background estimation (HB) proposed by [3].
- Minimizing an asymmetric truncated quadratic cost function for background estimation (TQ) proposed by [3].
- The Asymmetric penalized least squares algorithm (AsLs) proposed by [6].

The metric used to compare the algorithms is the mean squared error (MSE) between $R_{eback}[n]$ and the corresponding background removed spectrum, $S_{ebackfilt}[n]$. MSE is given by (4.22):

$$MSE(R_{eback}, S_{ebackfilt}) = \frac{\sum_{n=1}^N (R_{eback}[n] - S_{ebackfilt}[n])^2}{N} \quad (4.22)$$

4.4.1 Effect of number of peaks on background estimation

Figure 4.12 shows one simulated spectrum with 5, 10 and 30 peaks, randomly picked out from fD1, fD2, and fD3 respectively with the background estimated using the WTQPLS algorithm. The parameters used in this figure were optimized for dataset fD1. The values of the optimized parameters for this algorithm are as follows: $p = 0.0022$, $s = 0.005$, and $\varepsilon = 1258$.

Figure 4.12. One simulated spectrum from a) Dataset fD1. b) Dataset fD2. c) Dataset fD3 with background estimated using WTQPLS with parameters optimized for dataset fD1.

From the simulated spectra in the figure above, it can be seen that increasing the number of Raman peaks in the simulation also increased the potential of having overlapping peaks. The superposition of the overlapped peaks caused the spectrum's shape at these pixels to have a broadened overall width. As the number of overlapping peaks increased, so did the positive deviations, and so did the need to change the parameters accordingly. Without *a priori* knowledge of the spectrum shape, as in real spectroscopic measurements, it is difficult to choose a set of parameters that work for all spectra. This is a limitation in all background estimation methods [22]. To compensate for this, each of the parameters must be tuned accordingly for each dataset given the characteristics of the spectrum. This is confirmed by the tests done with fD1, fD2 and fD3, which show that the median MSE increased with increasing number of simulated peaks for parameters optimized for spectra in fD1. This can

be seen in the box plots in Figure 4.13. This is supported by work done in [22] which shows that increasing the number of peaks lead to more overlapping peaks, and that there was a need to accordingly adjust the parameters of 5 different background estimation algorithms (ModPoly, I-ModPoly, APoly, WPLS, and APLS).

Figure 4.13. Box plot showing MSE of dataset fd1 and fd3 respectively using optimum parameters set for fd1 dataset

After tuning the parameters for each dataset, the MSE of datasets fd2 and fd3 drastically reduced. Figure 4.14 shows the MSE results obtained using optimized parameter values for all algorithms. Parameter optimization was done on one spectrum from each dataset.

In all cases, the proposed WTQPLS algorithm produced the lowest median MSE for all 3 datasets, and had the lowest interquartile range. This indicates that for a given parameter combination, the WTQPLS algorithm is well suited to handle similar spectra with minor variations.

Figure 4.14. MSE of dataset fd1, fd2, and fd3 respectively using its own optimum parameters

To highlight the power of both proposed algorithms, s is set to 0 and the MSE of dataset fD1 is recalculated with the remaining parameters set to the previously selected optimal parameters. At this configuration, the WTQPLS and WHPLS converge to the same cost function. Similarly, the TQ and HB converge to the same asymmetric least squares cost function. Hence, the proposed algorithms and the non-quadratic cost functions produced the same results respectively, as can be seen in Figure 4.15.

Figure 4.15. MSE box plot for test ensemble from Data set fD1 with $s=0$

The addition of the roughness penalty of the penalized least squares algorithm with the asymmetric Huber and truncated quadratic cost functions lead to a superior background estimation in comparison to estimation with the TQ and HB algorithms. This suggests that the combination of the roughness penalty and asymmetric non-quadratic cost functions are better suited for background estimation.

Quantitative analysis of the proposed algorithms is not possible on real Raman data since the true fluorescence background is unknown. Both algorithms are therefore assessed qualitatively. Figure 4.16 and Figure 4.17 show the estimated background and background-free spectrum using one spectrum of a whole blood cell unexposed to radiation belonging to Donor 52 for both algorithms respectively.

Figure 4.16. a) Spectrum from D52 and background estimated using WTQPLS ($p=0.0001$, $s=0.05$, $\varepsilon=100$). b) Background-free spectrum.

Figure 4.17. a) Spectrum from D52 and background estimated using WHPLS ($p=0.0001$, $s=0.05$, $\varepsilon=100$). b) Background-free spectrum

For the same parameters, the WHPLS over-estimated the background at the wide Raman band (pixels 1300-1500). This is because the WHPLS algorithm is more influenced by positive deviations than the WTQPLS algorithm. As previously discussed, setting s lower for the WHPLS algorithm attenuates this issue as can be seen in Figure 4.18

Figure 4.18. **a)** Spectrum from D52 and background estimated using WHPLS ($p=0.0001$, $s=0.013$, $\varepsilon=100$). **b)** Background-free spectrum

4.5 Conclusion

This chapter provided two novel background estimation methods based on the principles of penalized least squares and minimizing a non-quadratic cost function. Both algorithms consistently outperformed the other methods on all simulated datasets when optimized parameters were used. Both algorithms produced the lowest median MSE in all datasets (fD1, fD2, and fD3) amongst all the algorithms considered in this thesis. In all cases, the WTQPLS outperforming the WHPLS algorithm.

Chapter 5: Classification

The main objective of this chapter is to demonstrate that signal processing (data normalization, fluorescence background removal, and cosmic spike removal) on the raw Raman spectrum is required before any further data analysis. Otherwise, the results will not be meaningful.

5.1 Classification procedure

The preprocessing procedures used to prepare spectra for data analysis are cosmic ray removal, smoothing, background subtraction, and data normalization. The preprocessed Raman spectra are then mean centered to reposition the data's centroid to the origin in preparation for dimensionality reduction using the principal components analysis (PCA). For an $N \times \chi$ data matrix where N is the number of spectra and χ is the number of variables (features), data centering is done by subtracting the mean of each feature. At the end of this procedure, all features will have zero mean.

After all spectra are preprocessed and mean centered PCA is performed to map the selected features to a lower dimension that captures uncorrelated variables important for analysis. In all tests conducted, discussed in the following sections, the number of principal components (PC) were selected by locating the elbow of an explained-variance plot. After dimensionality reduction, an LDA prediction model was trained and validated using 10-fold cross validation.

5.2 Principal Component Analysis (PCA)

PCA is a feature extraction and dimensionality reduction technique used to represent d-dimensional data in a lower-dimensional space that best explains it. This is done by finding a new set of variables (features), smaller than the original set, that retain most of the data's information. The new set of variables are uncorrelated with each other and referred to as principal components (PCs).

In the covariance matrix method, data is reduced into the PCA space as follows:

1. Calculate the covariance matrix of the data matrix (DM).
2. Calculate the eigenvalues and eigenvectors of the covariance matrix.

Each eigenvalue represents a principal component with the eigenvector with corresponding to the highest eigenvalue representing the first principal component. A more detailed description on PCA can be found in [54].

5.3 Linear Discriminant Analysis (LDA)

Classification methods are divided into two main categories: unsupervised methods and supervised methods [27]. Unsupervised methods are used when no class label is available. Supervised methods, on the other hand, are provided with data that contain class labels. In this thesis, only supervised classification method is considered. Supervised classification consists of two phases:

- **Training phase:** A labeled training dataset to fit a model (classifier).
- **Testing phase :** Data that is not part of the training set is used to test the model obtained in the training phase. This data is unseen by the classifier as it has not been used to train the classifier.

LDA is a supervised learning classifier that separates the classes by linearly combining the features. In a two class problem, the features belong to one of two classes (irradiated, or non-irradiated, in the context of this work). LDA classifies the features assuming that these two classes are linearly separable. Linearly separable classes are those which can be separated by a single linear decision surface. The LDA procedure classifies a feature, χ_i , as belonging to the class which minimizes the within-group variance, i.e.,

$$k = \underset{k}{\operatorname{argmin}} (\chi_i - \mu_k)' \sum^{-1} (\chi_i - \mu_k) \quad (5.1)$$

where k is the class label which minimizes the expression, χ_i is one observation in the feature vector, μ_k is the mean vector of the feature vector belonging to class k , and $\sum$ is the combined covariance matrix. Assuming that the different classes have a multivariate normal distribution with equal class covariances, this is the same as saying that the observation belongs to the group which maximizes the observation posterior probability. For a binary case, the discriminant function is given by (5.2):

$$d(\chi) = \chi' \sum^{-1} (\mu_1 - \mu_2) - \frac{1}{2} (\mu_1 + \mu_2)' \sum^{-1} (\mu_1 - \mu_2) + \log \frac{\pi_1}{\pi_2} \quad (5.2)$$

where π_k is the proportion of class k . When the proportion of classes are equal, the log term vanishes.

If $d(\chi) > 0$ then the feature will belong to class 1, else it will belong to class 2.

A more detailed description on LDA can be found in [54].

5.4 Feature selection

In this thesis, a specific band within the fingerprint region is selected as features. The fingerprint region is a region that contains most chemical information in the Raman spectrum. The region explored is pixels 1305 - 1450 which corresponds to $1504 - 1649 \text{ cm}^{-1}$. Figure 5.1 below shows a raw Raman spectrum and the intensities explored in classification, highlighted in red.

Figure 5.1. Region used for classification: $1504 - 1649 \text{ cm}^{-1}$ region

5.5 Performance metrics

The following performance metrics were used to assess the effects of preprocessing on distinguishing between non-radiated (0 Gy), and irradiated spectra (1, 2 and 5 Gy): sensitivity, specificity, precision (PPV), and negative predictive value (NPV). Sensitivity is

the ability of a test to correctly classify irradiated spectra. It is given by (5.3):

$$Sensitivity = \frac{TP}{TP + FN} \quad (5.3)$$

where TP is the number of spectra correctly identified as irradiated, and FN is irradiated spectra falsely identified as non-irradiated. Specificity is the ability of a test to correctly classify non-irradiated spectra. It is given by (5.4):

$$Specificity = \frac{TN}{TN + FP} \quad (5.4)$$

where FP is the number of spectra falsely identified as irradiated, and TN is the number of spectra truly identified as non-irradiated. The precision of a test, or its positive predictive value, gives the percentage of spectra identified as irradiated that are actually irradiated. It is given by (5.5):

$$PPV = \frac{TP}{TP + FP} \quad (5.5)$$

The NPV is the opposite of PPV. It gives the percentage of spectra identified as non-irradiated that are actually non-irradiated. The NPV of a test is given by (5.6):

$$NPV = \frac{TN}{TN + FN} \quad (5.6)$$

5.6 Smoothing

In all tests conducted, discussed below, the Raman spectrum was smoothed using a second order Savitzky-Golay filter with a window length, L_{SG} , of 5. Smoothing improves quality of the spectra by removing noise. The Savitzky-Golay window length is chosen by visually assessing the spectra to ensure that the noise reduction is achieved without spectrum distortion due to excessive smoothing.

Figure 5.2. Effect of spectrum smoothing with SG filter with different window lengths

Figure 5.2 shows one radiation free spectrum from Donor 125 smoothed using a Savitzky-Golay window length for $L_{SG} = 3, 5, 7, 11, 21, 31$. As the window size increases above 5, the Raman peaks begin to get distorted. This is more clearly visible in Figures 5.3 and 5.4 where the effects of smoothing are shown on two different Raman bands.

Figure 5.3. Effect of smoothing on Raman band at 727-780 cm^{-1} (pixels 528-581)

As the window length increased, the Raman peaks in the 727 cm^{-1} to 780 cm^{-1} region were squashed, and their profile broadened. As for the 1498 cm^{-1} to 1698 cm^{-1} region, multiple narrow bands were smoothed into one broader band.

Figure 5.4. Effect of smoothing on Raman band at $1498\text{-}1698\text{ cm}^{-1}$ (pixels 1299-1499)

5.7 Test 1: Effect of data normalization on classification

Raman spectra used in this study contain a degree of variance from each other due to differences in experimental conditions and instrumental parameter fluctuations. Normalization is applied to reduce variation that is uncorrelated with the chemical information of the spectra. In the dataset provided, there existed three main variations.

- Time variations: Spectra is acquired sequentially 20 times for every blood sample used (18 blood samples in total). As time passes, the acquired spectra may contain spectral variations from the first spectra acquired. As the drop of blood goes through the laser 20 times, the spectra acquired during the first exposure to laser may be different from the spectra acquired during the 20-th exposure.
- Day variations: Spectra from same donor collected on different days. Variations in the spectra may be expected from the same donor blood collected on different days.
- Donor variations: Spectra from different donors is expected to have differences since the amount of possible differences in the constituents in the blood.

A comparative study was done to see whether normalization affected classification or not.

In the study three different normalization methods were tested: vector, SNV, and area normalizations.

5.7.1 Experimental procedure

Before the experiments were conducted, all spectra were smoothed, cosmic spike removed, and background subtracted. After, the following experiments were considered:

- **Experiment 1:** In this experiment normalization tests were done on Donor 125 spectra. The spectra collected on Day 1 were used for training the classifier, while the spectra collected on Day 2 were used for testing.
- **Experiment 2:** In this experiment normalization tests were done on Donor 125 and Donor 52 spectra. Donor 125 spectra collected on both days were used for training the classifier, while Donor 52 spectra were used for testing it. Since Donor 125 was used for training, the normalization used for Donor 125 was applied to Donor 52.

Both experiments were carried out for the following cases:

- Case 1: Classification with vector normalization
- Case 2: Classification with area (AUC) normalization
- Case 3: Classification with SNV normalization
- Case 4: Classification without normalization (unnormalized spectra)

5.7.2 Parameters settings

For all experiments, the background was estimated using the WTQPLS algorithm with the following optimal parameters obtained by changing the parameters in an adhoc manner and picking the best amongst the set that was considered: $s=0.0035$, $\varepsilon=5$, and $p=0.2$. It should be noted that there was no other way to pick the parameters as there was no golden reference standard. For spike removal, the algorithm proposed by Tian was used.

5.7.3 Results

From the results in Table 5.1, it may be inferred that normalization is an important preprocessing step. In Experiment 1, not normalizing the data caused a specificity of 56 % which is close to random. In Experiment 2, the sensitivity was 21 % and the NPV was 56 %. From the normalization methods assessed, SNV normalization performed best.

Table 5.1. A summary of classification results for vector, AUC, SNV, and non-normalized spectra for **a)** Experiment 1 **b)** Experiment 2

Experiment 1					
		Sens	Spec	PPV	NPV
1	Vector Norm	0.997	0.294	0.809	0.973
2	AUC Norm	0.999	0.277	0.805	0.99
3	SNV Norm	0.988	0.77	0.928	0.958
4	No norm	0.974	0.56	0.869	0.878
Experiment 2					
		Sens	Spec	PPV	NPV
1	Vector Norm	0.818	0.613	0.679	0.772
2	AUC Norm	0.84	0.480	0.619	0.759
4	SNV Norm	0.767	0.736	0.744	0.760
5	No Norm	0.217	1	1	0.5611

5.8 Test 2: Effect of background subtraction on classification

A comparative study was done to see whether background subtraction affected classification or not. In the study, two different background removal cases were tested: visually optimal background subtraction, and visually sub-optimal background subtraction. For optimal background subtraction, a combination of WTQPLS parameters were selected which gave a visually acceptable background estimation (background estimation is deemed acceptable when it removes almost all the background or estimates almost all the background and does only minimally distort the peaks). For sub-optimal background subtraction, a combination of parameters were selected that intentionally under-estimated the background.

5.8.1 Experimental procedure

All experiments in this section were conducted on Donor 125 spectra. Spectra collected on Day 1 were used for training the classifier, while those collected on Day 2 were used for testing the classifier. Before the experiments were conducted, all spectra were smoothed, cosmic spike removed, and SNV normalized.

- **Experiment 3** : In this experiment, all the Raman spectra backgrounds were underestimated. The parameters of the WTQPLS algorithm used to estimate the background were set to $s=0.0035$, $\varepsilon=10^5$, and $p=0.2$.
- **Experiment 4** : In this experiment classification was done on visually optimal background subtracted spectra. The parameters of WTQPLS were set to $s=0.0035$, $\varepsilon=5$,

and $p=0.2$.

5.8.2 Results

Figure 5.5 shows the estimated background and background subtracted spectrum for one spectrum belonging to Donor 125.

(a) Optimal background subtraction

(b) Under estimating the background

Figure 5.5. a) Optimal background subtraction (Experiment 4). b) Under-estimating the background (Experiment 3)

Underestimating the background failed to remove all of the fluorescence. As a consequence,

the classifier learnt the fluorescence intensities alongside the smaller Raman peaks. This masked the chemical variations present between radiated and irradiated samples, since fluorescence is orders of magnitude higher than Raman intensities.

Table 5.2. A summary of classification results for **a)** Experiment 3 showing classification results using optimally subtracted spectra . **b)** Experiment 4 showing classification results using spectra contaminated with fluorescence

Optimal background subtraction				Under estimating the background			
Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
0.988	0.77	0.928	0.958	1	0	0.750	0

This resulted in inaccurate classification, as documented in Table 5.2 where the NPV and specificity were nil. When the parameters were adjusted to give a visually “optimum” background estimation and subtraction, little fluorescence remained in the spectrum. This enhanced the spectrum to background ratio, allowing the classifier to learn small differences present between non-radiated and irradiated spectra.

5.9 Test 3: Effect of cosmic spikes removal on classification

A comparative study was conducted to see whether cosmic spikes affected classification or not. In the study, all spectra were first smoothed, background subtracted, and SNV normalized. The following experiments were conducted to show the effect of cosmic spikes on classification:

- **Experiment 5** : In this experiment classification was done on cosmic spike-free, smoothed, SNV normalized, and background subtracted spectra.
- **Experiment 6** : In this experiment classification was done on smoothed, background subtracted, and SNV normalized spectra contaminated with cosmic spikes.

As with the background subtraction experiments, Donor 125 spectra collected on Day 1 was used for training the classifier, while Donor 125 spectra collected on Day 2 was used for testing.

5.9.1 Results

Not removing the cosmic spikes had a detrimental effect on classification. With spike-contaminated spectra, the ability of the classifier to correctly identify non-radiated spectra was nil. The NPV was also nil, meaning that the probability of spectra that were identified as non-radiated were most likely radiated.

Table 5.3. A summary of classification results for **a)** Experiment 5 showing classification results using spike-free spectra. **b)** Experiment 6 showing classification results using spike contaminated spectra.

Spike-free spectra				Spectra contaminated with spikes			
Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
0.988	0.77	0.928	0.958	0.997	0	0.745	0

5.10 Conclusion

In this chapter classification was used to highlight the importance of Raman preprocessing. The various experiments performed in this chapter led to the conclusion that preprocessing (normalization, cosmic spike removal, and background removal) is vital for any further data analysis using Raman spectra. This is highlighted in the classification metric results outlined in Tables 5.1 - 5.3 .

Chapter 6: Conclusion and Future work

6.1 Conclusion

To treat patients in case of wide scale exposure to radioactive chemical substances, medical authorities require a fast and reliable method in order to detect exposed citizens. The current gold standard, cytogenetic analysis, provides an accurate measure of exposure but requires 2-3 days, where medical intervention may be too late. Raman spectroscopy offers an alternative solution which can provide a quick, reliable method to detect exposure to radiation. This thesis provides the required signal processing to enable Raman spectroscopy based radiation exposure detection. Appropriate preprocessing of Raman spectra becomes a crucial component of the signal processing.

The work explored in this thesis described three novel algorithms for preprocessing Raman spectra. These are:

1. Cosmic spike detection using energy detection and ensemble empirical mode decomposition (EEMD).
2. Background estimation using weighted truncated quadratic asymmetric least squares (WTQPLS).
3. Background estimation using weighted Huber asymmetric least squares (WHPLS).

Chapter 1 provided a brief summary of the Raman spectroscopic technique and relevant information on the thesis structure.

Chapter 2 provided a summary of existing preprocessing methods (smoothing, cosmic ray spikes removal methods, fluorescence background removal methods, and data normalization) and their limitations.

Chapter 3 provided a novel method for the detection of cosmic spikes based on the principles of EEMD and energy detection. After detection, a variety of methods exist in literature that could be used to remove the spikes with minimum interference to the surrounding pure Raman profile. A qualitative analysis was done to compare the proposed algorithm with four other methods. The proposed algorithm best preserved non-spike regions while fully removing spikes in simulated and real spectra.

Chapter 4 provided two novel background estimation and subtraction algorithms. The

algorithms were built upon the works of [6] and [3]. The proposed algorithm combined the roughness penalty estimation of the AsLs algorithm, and the minimization of non-quadratic cost functions in order to estimate the background. The proposed algorithms were quantitatively compared to the AsLs algorithm, and both asymmetric cost functions using simulated datasets, and were found to produced lower MSE in all cases. Then the proposed algorithm was applied on real dataset to remove the background fluorescence.

In Chapter 5 several tests were run to highlight the importance of preprocessing Raman spectra for data analysis. The chapter explored the effects of data normalization, cosmic spike removal, and background subtraction on distinguishing between Raman spectra of irradiated (0, 1, 2 and 5 Gy) and non-irradiated (0 Gy) whole blood cells. The conclusion reached was that all three processing steps are required for any meaningful analysis to take place as confirmed by results in Tables 5.1 - 5.3 .

6.2 Limitations:

The main limitation in the work presented in thesis are as follows:

- There are no standardized databases of whole blood spectra unexposed and exposed to radiation to validate the resultant spectrum. Therefore, the work presented in this work could not be compared with other related work present in the literature.
- There are no gold standard noise-free spectra available to validate the proposed cosmic ray spike and fluorescence estimation and removal.
- An investigation was conducted to study how a non-irradiated sample changed over the course of 100s of exposure to various Raman laser powers. It was found that as the power of the laser increased, there was a corresponding increase in intensity of the spectrum over the course of a given acquisition. This indicated that the variations observed between each successive spectra acquisition may be due to possible degradation of the sample due to the 50 mW laser power used. However, this could not be confirmed firmly through repeated testing. The possibility of damage to the cells also exist.
- Amount of constituents in the blood change from donor to donor. For the same volume of blood collected, different donors have different white blood cell, and red blood cell counts. The composition of the cells themselves, such as the amount of hemoglobin in red blood cells, are different. The role-play of the blood components and the composition on the resultant Raman spectrum is unknown and needs to be addressed in order to take into account donor to donor variability in both unexposed and exposed samples. In this work, such nuisance factors have been accounted for.

6.3 Future Work

Future work includes the following:

- **Energy Detection and signal characteristics to detect spikes :**

The principle of energy detection discussed in the second chapter of this thesis could be applied on the raw Raman spectrum to detect both Raman peaks and spikes. After both peaks and spikes are detected, their characteristics could be extracted and leveraged to decide whether the detected peak is a Raman peak or a spike. This could be done by grouping all extracted characteristics by a clustering technique. Spikes will have similar characteristics and will therefore cluster differently than the broader Raman bands.

- **Windowed WTQPLS and WHPLS:**

Instead of fitting a single polynomial to the spectrum, the spectrum could first be divided into segments. Background estimation could then be done on each segment with the final estimate being the sum of all estimated backgrounds. This would be especially useful for complex backgrounds that could be viewed as the sum of different backgrounds. However, the length of the segment and number of segments need to be ascertained. This would require further research on fluorescent background.

- **Combining the non-quadratic cost functions with P-splines:**

B-splines of order k are basis functions that can be used to represent any spline function with the same degree. P-splines are the combination of fitting the data with B-splines and the roughness penalty proposed by Whittaker. A P-spline of degree k with a penalty order of degree $k+1$ can exactly fit a polynomial of degree k independent of ε . For background estimation, the minimized P-splines solution with the addition of the non-quadratic cost functions is given (6.1):

$$\hat{z} = (B'WB + \varepsilon D'D)^{-1} B'W(S + d) \quad (6.1)$$

where the WTQPLS and WHPLS are special cases where $\mathbf{B} = \mathbf{I}$ and the number of knots is equal to the number of data points [55]. P-splines allow for the independent selection of B-spline order. The penalty term allows the user selected number of B-splines to be much larger than the number of observations with the solution remaining well-conditioned. Similarly, a lower number of appropriately selected knots may be used, thereby, significantly reducing the dimensionality of the data. For large ε values, and a penalty of order k , the fitted estimate will tend towards a polynomial of $k - 1$. It would therefore be interesting to study different penalties other than the second order differences, and their effect on the proposed algorithms performance given that fluorescence backgrounds are usually modeled by polynomials.

- **Optimizing WTQPLS and WHPLS for each individual spectrum:**

Figure 6.1 shows the first 20 successively acquired spectra of the first sample of donor 125's non-irradiated whole blood cells after background removal using the following WTQPLS parameters: $s = 0.03$, $p = 0.03$, and $\varepsilon = 10^3$. The parameters were selected to produce a visually acceptable result using the first spectrum. Due to subtle differences in the fluorescence profile of each spectra, parameters may need to be adjusted in order for the same exact background to be removed. Since, however, datasets usually contain many spectra, it is advantageous to find an adaptive solution which optimizes all three parameters per spectra.

Figure 6.1. Twenty spectra background subtracted with parameters selected to produce a visually pleasing result using one spectrum

Bibliography

- [1] “Theory of Raman scattering.” [Online]. Available: <http://bwtek.com/raman-theory-of-raman-scattering>. [Accessed: 15-July-2018].
- [2] P. M. Ramos and I. Ruisánchez, “Noise and background removal in Raman spectra of ancient pigments using wavelet transform,” *Journal of Raman Spectroscopy*, vol. 36, no. 9, pp. 848–856, 2005.
- [3] V. Mazet, C. Carteret, D. Brie, J. Idier, and B. Humbert, “Background removal from spectra by designing and minimising a non-quadratic cost function,” *Chemometrics and Intelligent Laboratory Systems*, vol. 76, no. 2, pp. 121–133, 2005.
- [4] K. Kong, C. Kendall, N. Stone, and I. Notingher, “Raman spectroscopy for medical diagnostics - From in-vitro biofluid assays to in-vivo cancer detection,” *Advanced Drug Delivery Reviews*, vol. 89, pp. 121–134, 2015. [Online]. Available: <http://dx.doi.org/10.1016/j.addr.2015.03.009>
- [5] J. Castelino, P. Holland, G. Jacobs, M. Lapidot, and M. Markovic, “Effects of ionizing radiation on blood and blood components: A survey,” *International Atomic Energy Agency*, p. 42, 1997.
- [6] P. Eilers, “A perfect smoother,” *Analytical Chemistry*, vol. 75, no. 14, pp. 3631–3636, Jul 2003. [Online]. Available: http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers{}_2005.pdf<http://pubs.acs.org/doi/abs/10.1021/ac034173t>
- [7] A. Savitzky and M. Golay, “Smoothing and differentiation of data by simplified least squares procedures,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [8] S. A. Press, W.H. and W. T. Vetterling, *Numerical methods for Fortran 77: The art of scientific computing*, 2nd ed. New York, NY, USA: Cambridge University Press, 1987.
- [9] L. Komsta, “A comparative study on several algorithms for denoising of thin layer densitograms,” *Analytica Chimica Acta*, vol. 641, no. 1-2, pp. 52–58, 2009.
- [10] P. Eilers and D. Marx, “Flexible smoothing with B-splines and penalties,” *Statistical Science*, vol. 11, no. 2, pp. 89–121, 1996. [Online]. Available: <http://projecteuclid.org/euclid.ss/1038425655>
- [11] P. Eilers, B. Marx, and M. Durbán, “Twenty years of P-splines,” *SORT-Statistics and Operations Research Transactions*, vol. 39, no. 2, pp. 149–186, 2015.

- [12] P. Eilers and R. Bonifacio, "Automatic smoothing of remote sensing data," *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images, MultiTemp 2017*, June 2017.
- [13] F. Ehrentreich and L. Summchen, "Spike removal and denoising of Raman spectra by wavelet transform methods," *Analytical Chemistry*, vol. 73, no. 17, pp. 4364–4373, 2001.
- [14] S. Anthony and J. Timlin, "Removing cosmic spikes using a hyperspectral upper-bound spectrum method," *Applied Spectroscopy*, vol. 71, no. 3, pp. 507–519, 2017.
- [15] Y. Tian and K. Burch, "Automatic spike removal algorithm for Raman spectra," *Applied Spectroscopy*, vol. 70, no. 11, pp. 1861–1871, 2016.
- [16] D. Zhang, K. Jallad, and D. Ben-Amotz, "Stripping of cosmic spike spectral artifacts using a new upper-bound spectrum algorithm," *Applied Spectroscopy*, vol. 55, no. 11, pp. 1523–1531, 2001.
- [17] G. Phillips and J. Harris, "Polynomial filters for datasets with outlying or missing observations: Application to charge-coupled-device-detected Raman spectra contaminated by cosmic rays," *Analytical Chemistry*, vol. 62, no. 21, pp. 2351–2357, 1990.
- [18] A. Maury and R. Revilla, "Autocorrelation analysis combined with a wavelet transform method to detect and remove cosmic rays in a single Raman Spectrum," *Applied spectroscopy*, vol. 69, no. 8, pp. 984–92, 2015. [Online]. Available: <http://asp.sagepub.com/content/69/8/984.full>
- [19] Y. Katsumoto and Y. Ozaki, "Practical algorithm for reducing convex spike noises on a spectrum," *Applied Spectroscopy*, vol. 57, no. 3, pp. 317–322, 2003.
- [20] D. Zhang and D. Ben-Amotz, "Removal of cosmic spikes from hyper-spectral images using a hybrid upper-bound spectrum method," *Applied Spectroscopy*, vol. 56, no. 1, pp. 91–98, 2002.
- [21] L. Zhang and M. J. Henson, "A practical algorithm to remove cosmic spikes in raman imaging data for pharmaceutical applications," *Applied Spectroscopy*, vol. 61, no. 9, pp. 1015–1020, 2007.
- [22] P. J. Cadusch, M. M. Hlaing, S. A. Wade, S. L. McArthur, and P. R. Stoddart, "Improved methods for fluorescence background subtraction from Raman spectra," *Journal of Raman Spectroscopy*, vol. 44, no. 11, pp. 1587–1595, nov 2013. [Online]. Available: <http://doi.wiley.com/10.1002/jrs.4371>
- [23] W. Butler and D. Hopkins, "Higher derivative analysis of complex absorption spectra," *Photochemistry and Photobiology*, vol. 12, no. 6, pp. 439–450, 1970.
- [24] A. F. Fell, "Biomedical applications of derivative spectroscopy," *Trends in Analytical Chemistry*, vol. 2, no. 3, pp. 63–66, 1983.
- [25] T. C. O'Haver, A. F. Fell, G. Smith, P. Gans, J. Sneddon, L. Bezur, R. G. M., J. M. Ottaway, J. N. Miller, T. A. Ahmad, A. F. Fell, P. B. Chadburn, and C. T. Cottrell,

- “Derivative spectroscopy and its applications in analysis,” *Analytical Proceedings*, no. 22, pp. 22–46, 1980.
- [26] D. Zhang and D. Ben-Amotz, “Enhanced chemical classification of raman images in the presence of strong fluorescence interference,” *Applied Spectroscopy*, vol. 54, no. 9, pp. 1379–1383, 2000.
- [27] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, “Review of multidimensional data processing approaches for Raman and infrared spectroscopy,” *EPJ Techniques and Instrumentation*, vol. 2, no. 1, p. 8, 2015. [Online]. Available: <http://www.epjtechniquesandinstrumentation.com/content/2/1/8>
- [28] L. S. Mosier-Boss, P. and R. Newbery, “Fluorescence rejection in Raman spectroscopy by shifted-spectra, edge detection, and FFT filtering techniques,” *Applied Spectroscopy*, vol. 49, no. 5, pp. 630–638, 1995.
- [29] I. Daubechies, “The wavelet transform, time-frequency localization and signal analysis,” *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961–1005, 1990.
- [30] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, 1st ed. Wellesley-Cambridge Press, 1996.
- [31] Y. Hu, T. Jiang, A. Shen, W. Li, X. Wang, and J. Hu, “A background elimination method based on wavelet transform for Raman spectra,” *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 94–101, 2007.
- [32] C. M. Galloway, E. C. Le Ru, and P. G. Etchegoin, “An iterative algorithm for background removal in spectroscopy by wavelet transforms,” *Applied Spectroscopy*, vol. 63, no. 12, pp. 1370–1376, 2009.
- [33] Z. Zhang, S. Chen, and Y. Liang, “Peak alignment using wavelet pattern matching and differential evolution,” *Talanta*, vol. 83, no. 4, pp. 1108–17, 2011. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0039914010006259>
- [34] W. Kiefer, A. Mazzolini, and P. Stoddart, “Recent advances in linear and nonlinear Raman spectroscopy,” *Journal of Raman Spectroscopy*, vol. 38, no. April, pp. 1538–1553, 2007. [Online]. Available: <http://dx.doi.org/10.1002/jrs.1902>
<http://onlinelibrary.wiley.com/store/10.1002/jrs.1902/asset/1902{ }ftp.pdf?v=1{&t=h2ddhr1g{&s=77a1c3920a3ac3b4f0c58c3364221f79efd0cf32{ }5CnC:/Vasili/Laboratory/Articles/studyplasma.Data/PDF//Kiefer-2007-RecentAdvancesin>
- [35] M. Gao, G. Lewis, G. M. Turner, A. Soubret, and V. Ntziachristos, “Effects of background fluorescence in fluorescence molecular tomography,” *Applied Optics*, vol. 44, no. 26, p. 5468, 2005. [Online]. Available: <https://www.osapublishing.org/abstract.cfm?URI=ao-44-26-5468>
- [36] K. W. Short, S. Carpenter, J. P. Freyer, and J. R. Mourant, “Raman spectroscopy detects biochemical changes due to proliferation in mammalian cell cultures,” *Biophysical Journal*, vol. 88, no. 6, pp. 4274–4288, 2005. [Online]. Available: <http://dx.doi.org/10.1529/biophysj.103.038604>

- [37] T. J. Vickers, R. E. J. Wambles, and C. K. Mann, "Curve fitting and linearity : Data processing in Raman spectroscopy," *Applied Spectroscopy*, vol. 55, no. 4, pp. 389–393, 2001.
- [38] Z. Huang, H. Lui, D. McLean, M. Korbelik, and H. Zeng, "Raman spectroscopy in combination with background near-infrared autofluorescence enhances the in vivo assessment of malignant tissues," *Photochemistry and photobiology*, vol. 81, no. 5, pp. 1219–26, 2005. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15869327>
- [39] I. B. Gornushkin, P. E. Eagan, A. B. Novikov, B. W. Smith, and J. D. Winefordner, "Automatic correction of continuum background in laser-induced breakdown and Raman spectrometry," *Applied Spectroscopy*, vol. 57, no. 2, pp. 197–207, 2003.
- [40] M.-J. A. Lieber ,C, "Automated method for subtraction of flourescence from biological Raman spectra," *As*, vol. 57, no. 11, pp. 1363–1367, 2003.
- [41] J. Zhao, H. Lui, M. David, and H. Zeng, "Automated auto-fluorescence background subtraction algorithm for biomedical Raman spectroscopy," *Applied Spectroscopy*, vol. 61, no. 11, pp. 1225–1232, 2007.
- [42] T. Wang and L. Dai, "Background subtraction of Raman spectra based on iterative polynomial smoothing," *Applied Spectroscopy*, vol. 71, no. 6, pp. 1169–1179, 2017.
- [43] P. Eilers and H. F. M. Boelens, "Baseline correction with asymmetric least squares smoothing," *Life Sciences*, pp. 1–26, 2005. [Online]. Available: <http://www.science.uva.nl/{~}hboelens/publications/draftpub/Eilers{ }2005.pdf>
- [44] T. W. Randolph, "Scale-based normalization of spectral data," *Cancer biomarkers : section A of Disease markers*, vol. 2, pp. 135–144, 2006.
- [45] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 100–114, 2012. [Online]. Available: <http://dx.doi.org/10.1016/j.chemolab.2012.03.011>
- [46] A. Rinnan, F. Van den Berg, and S. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.
- [47] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 995, 903, 1998. [Online]. Available: <http://dx.doi.org/10.1098/rspa.1998.0193{ }5Cnhttp://rspa.royalsocietypublishing.org/content/454/1971/903>
- [48] D. Gabor, "Theory of communication," *Journal of the Institution of Electrical Engineers - Part I: General*, vol. 94, no. 73, pp. 58–, January 1947.
- [49] D. Kaslovsky and F. Myer, "Noise corruption of empirical mode decomposition and its effect on instantaneous frequency," *Advances in Adaptive Data Analysis*, vol. 02,

- no. 03, pp. 373–396, 2010. [Online]. Available: <http://www.worldscientific.com/doi/abs/10.1142/S1793536910000537>
- [50] T. Wang, M. Zhang, Q. Yu, and H. Zhang, “Comparing the applications of EMD and EEMD on time-frequency analysis of seismic signal,” *Journal of Applied Geophysics*, vol. 83, pp. 29–34, 2012.
 - [51] Z. Wu and N. Huang, “Ensemble empirical mode decomposition: a noise-assisted data analysis method,” *Advances in Adaptive Data Analysis*, vol. 01, no. 01, pp. 1–41, 2009. [Online]. Available: <http://www.worldscientific.com/doi/abs/10.1142/S1793536909000047>
 - [52] P. Rousseeuw and C. Croux, “Alternatives to the median absolute deviation,” *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993.
 - [53] J. Rukhshinda, “Automated baseline estimation for analytical signals,” pp. 75–80, 2016.
 - [54] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. New York: Wiley, 2001.
 - [55] G. Frasso and P. Eilers, “Smoothing parameter selection using the L-curve,” *Proceedings of the 27th International Workshop on Statistical Modelling*, pp. 135–140, 2012.