Methods for Improving Signal to Noise Ratio in Raman Spectra

Sinéad Barton, B.E.

A thesis presented to Maynooth University for the degree of

Doctor of Philosophy

Head of department: Prof. Ronan Farrell

Department of Electronic Engineering

Maynooth University

October 2018

Under the Supervision of Dr. Bryan Hennelly

Table of Contents

List of Figures

List of Tables

Abstract

Raman microspectroscopy is an optoelectronic technique based on the inelastic scattering of light. This technique has been demonstrated to have potential to identify different materials based on subtle differences in the Raman spectral profile using various multivariate statistical classification tools. However, Raman scattering is an inherently weak process. Low photon counts coupled with non-ideal collection efficiencies means that Raman spectroscopy is vulnerable to noise. This makes system optimisations, as well as efficient and reliable noise removal, a necessity in sensitive applications such as chemical classification or diagnostics. Provided in this thesis are software and experimental methodologies to evaluate system performance, predict system performance under various conditions, and to identify the optimal system configuration/set-up in order to achieve the highest possible signal to noise ratio. Modelling methodologies presented in this thesis allow the user to systematically evaluate minimum acquisition times, optimise camera read-out modes, and predict system behaviour with alternative optical elements in order to maximise signal to noise ratio. The denosing algorithms presented in this thesis have been shown to provide superior signal to noise ratio when compared with their traditional counterparts. When compared with the double acquisition method, the proposed cosmic ray removal algorithm resulted in a 10% improvement. An algorithm that enhances Savitzky-Golay smoothing with maximum likelihood estimation produced spectra with up to double the signal to noise ratio when compared to the raw spectra and consistently outperformed the algorithms it was compared to. The use of reflective substrates is also investigated and was shown to approximately triple the collected Raman scatter when compared with transparent substrates. By utilising the methodologies detailed in this thesis it is possible to improve the efficiency of the Raman system in question.

Acknowledgements

If someone had told me ten years ago that I would one day be Dr. Barton I'd have laughed, and yet, here we are. It's been one hell of a journey and I couldn't have done it without all the fantastic people who have supported me along the way.

Firstly, I would like to thank my supervisor, Dr. Bryan Hennelly, for the opportunity to work on such an exciting project and his guidance over the course of this PhD. Working with you, I got to experience a lot of things I might never have had the chance to.

To the staff, students, and interns in the Engineering Department at Maynooth who made sure each day had a ray of sunshine, I couldn't have asked for more. I'd never have made it through an undergrad, let alone, a PhD without all of ye. If I asked for help, I got it and if I needed encouragement, you gave it to me. I owe you all a lot more than chocolate, movies, or a pint but I hope that they, along with my gratitude, will suffice!

The universe has also seen fit to bestow upon me the best of friends. The last few years I have spent late nights chilling with a fountain of wisdom, singing with a lark, and playing chess with a husky. I also spent long days drinking coffee with a people whisperer, eating food with a force to be reckoned with, and trying to fill a dainty pair of Vans. I'd have scuppered my own boat without your example to follow. Keep it real, you magnificent creatures!

Finally, to my beloved father, I simply wouldn't be the woman I am today without you. I will always work hard so that you can continue to be proud of me. Lugs and thank you for everything.

Chapter 1

Introduction

Raman microspectroscopy is an optoelectronic technique based on the inelastic scattering of light that can be used to evaluate the chemical composition of various materials, including biological samples. This technique has been demonstrated to have potential to identify different materials based on subtle differences in the Raman spectral profile using various multivariate statistical classification tools. [1–6]

However, Raman scattering is an inherently weak process, with approximately 1 in 10^7 photons undergoing scattering. [7] Low photon counts coupled with non-ideal collection efficiencies, e.g. the numerical aperture of the microscope objective or the camera's quantum efficiency, means that Raman spectroscopy is vulnerable to noise. Noise will decide the detection limit of the recording process as well as the classification potential of multivariate statistical analysis that may be applied to a recorded dataset for classification purposes, making system optimisations, as well as efficient and reliable noise removal, a necessity in sensitive applications such as chemical classification or diagnostics. [8, 9] These optimisations will result in an enhancement of the signal to noise ratio that will translate into shorter acquisition times and higher classification sensitivity and specificity.

There are three primary approaches to minimising experimental noise in Raman systems; i) increasing acquisition times in order to collect more photons and, therefore, reduce the impact of shot noise, ii) optimising hardware in order to collect more scattered photons, and iii) denoising software. Software is a flexible, adaptable, and often cheap solution to this problem; this thesis aims to provide software and experimental methodology to evaluate system performance, predict system performance under various conditions, and to identify the optimal system configuration/set-up in order to achieve the highest possible signal to noise ratio. This thesis includes several contributions in the area of Raman spectroscopy that are all focused on improving the signal to noise ratio of the recorded Raman spectra:

- Rigorous methodology for estimating the signal to noise ratio in a given Raman spectrum.
- Methodology for estimating the 'system specific irradiances', i.e. estimating the irradiance of the spectrum in the plane of the detector. This facilitates the creation of simulated datasets that simulate the effect of a variety of experimental conditions.
- Software based methods that enables the identification of the recording parameters that provide the optimal signal to noise ratios for a given application that makes use of multivariate classification. These methods include; acquisition time and camera read-out parameters. This approach also makes use of simulated datasets.
- Software based methods for predicting the effect of altering optical elements in a given system on the signal to noise ratio of the recorded spectra.
- Investigation of the optimal substrate for Raman based cytology, particularly when laser excitation in the near infrared wavelength region is used.
- An algorithm to remove cosmic ray artefacts from spectra that yield spectra with significantly higher signal to noise ratio when compared to the commonly used double acquisition method.
- A denoising algorithm that enhances Savitzky-Golay smoothing with maximum likelihood estimation, which produces denoised spectra with significantly higher signal to noise ratio and peak fidelity than traditional Savitzky-Golay smoothing.
- The introduction of a novel metric known as the 'SNR product' that enables the user to evaluate denoising algorithms in terms of both peak fidelity and overall 'smoothness' in a single metric.

It is believed that the contributions listed above will be practically useful in many applications of Raman spectroscopy. By applying the methodology that is detailed in this thesis to optimise collected Raman scatter and increase the signal to noise ratio of spectra, it is possible to reduce acquisition time, increase the overall throughput of the system, and reliably record spectra that are suitable for classification purposes in the most efficient and cost effective manner possible, this includes the selection of the most cost-effective equipment (e.g. array detection) for a particular application. A detailed breakdown of the chapters contained in this thesis is provided in the following subsections.

1.1 Thesis Overview

This thesis is intended to be accessible to the engineer, biomedical physicist, and clinician. For this purpose a thorough background is provided that describes the principles on which the contributions in this thesis are based. This background material is provided in Chapters 2, 3, and 4. These background chapters provide no novel content with the exception of Section 4.4 that describes a novel method that can be used to estimate the irradiance that is incident on the array detector in a Raman spectrometer. This section, and the chapters that follow, all provide novel contributions that have either been published, submitted to a journal or are in preparation for submission. The details of this are provided at the beginning of each relevant chapter.

Overview of Raman spectroscopy, Chapter 2

In this chapter a classical description of the physics behind Raman spectroscopy is given. Raman spectroscopy may also be described using quantum mechanics; however, we do not explore this aspect in this thesis and focus only on the classical theory. Instead, references are provided for the reader should they require additional information. A description of the basic set-up of the Raman spectrometer is also provided in this section.

Numerical Methods, Chapter 3

The purpose of this chapter is to provide an overview of algorithms that are typically applied to Raman spectra. Multivariate statistical analysis must be applied to spectra in order to classify the data. This section details the typical signal correction techniques that must be applied as pre-processing measures in advance of multivariate classification, such as calibration and background subtraction as well as the multivariate statistical methods applied to the spectra to classify the cell types. These numerical methods are used throughout this thesis.

Signal to Noise Ratio in the Context of Raman Spectroscopy, Chapter 4

The purpose of this chapter is to mathematically define the noise sources present in Raman spectra so that they can be reliably modelled. There are several sources of noise present in experimental Raman spectra and all of these noise sources impact on the signal to noise of the Raman spectrum and the effectiveness of classification.

This chapter also details the methodology behind creating artificial datasets, which makes use of the noise models that are reviewed at the beginning of the chapter. Creation of artificial datasets relies on a procedure for extracting a reference spectral profile that is referred to in this thesis as a 'system specific irradiance'. The modelling of the noise sources, and the creation of artificial datasets, will be used throughout the thesis to analyse the signal to noise ratio of experimental spectra, perform system optimisation procedures, and test the performance of denoising algorithms.

The Effect of Signal to Noise Ratio on Classification, Chapter 5

Following the previous chapter in which the signal to noise ratio is investigated in a Raman spectrum, in this chapter the impact of low signal to noise ratios is investigated on the capability of Raman spectroscopy to provide classifiable spectra. Shot noise and dark current noise are time dependent and yet the length of the acquisition time affects these respective noise sources in the signal to noise ratio in an opposing manner. Dark current accumulates over time and, therefore, a long acquisition time will increase the dark current noise contribution in the spectrum. However, despite the fact that shot noise also increases with acquisition time, the impact of the shot noise on the signal to noise ratio will lessen with respect to acquisition time, because the signal intensity will increase at a faster rate than the shot noise. Therefore, a compromise must be made in terms of acquisition time. Another point of note is that longer acquisition times will result in more contamination with cosmic ray artefacts. This section shows that higher signal to noise ratios provide better classification accuracy using principal components analysis, which can provide information about minimum acquisition times for reliable classification for a given application.

Optimal CCD Read Modes, Chapter 6

The purpose of this chapter is to investigate the various camera modes that are available in typical array detectors used for Raman spectroscopy and how these settings can be optimised in order to maximise signal to noise ratio for a given application. Optimising the design and equipment used in the Raman system is an important step towards maximising the SNR of the resulting spectra. Assuming that the optical elements in the collection path of the spectrometer are suitable for the application, the most significant contributor of experimental noise is the array detector, for the purposes of this thesis, charge-coupled device (CCD). This piece of hardware is often cooled in order to reduce the rate of dark current generation. It is also often necessary to carefully choose read parameters under certain experimental conditions in order to minimise dark current and read noise contributions. Incorporating a CCD that has extremely low dark current, low read noise, and/or is resistant to cosmic rays may not always be feasible. In such cases, it is necessary to make the most of the system available by modelling it's behaviour and selecting the optimal CCD settings.

Predicting the Effect of Changing Optical Elements in a Given System, Chapter 7

By expanding the model described in the previous chapter, it is possible to predict how the system may behave after any optical element has been replaced, for example the CCD, so that the user can deduce with a certain level of confidence, that a piece of expensive new equipment will or will not improve the classification accuracy of their system. This chapter expands on the previous model by including the quantum efficiency of the CCD so that datasets may be simulated in a particular system for any given CCD. We briefly discuss how this methodology may be expanded on to include other optical elements, such as the microscope objective or diffraction grating.

An Algorithm for the Removal of Cosmic Ray Artefacts, Chapter 8

In the event that no further enhancement in signal to noise ratio can be achieved through optimisation of the hardware, software post-processing methods may still be used to further enhance the signal to noise ratio. It is necessary to approach the removal of cosmic ray artefacts differently to that of CCD induced noise due to their disparate natures. A number of methods have previously been proposed for the detection and replacement of pixels contaminated by cosmic ray artefacts (CRAs), which result from the random interaction of high energy particles with the detector sensor during the acquisition. Arguably the most commonly used technique is the double acquisition method [10], whereby two consecutive spectra are recorded, compared, and averaged together in order to remove the CRA contamination. However, multiple acquisitions will degrade the SNR of the resulting spectrum due to the inclusion of multiple instances of read noise being introduced into the spectrum. Therefore, a new method is proposed in this chapter whereby the principles behind the double acquisition method are modified in order to process databases of single acquisition spectra such that the SNR of the spectra may be optimised.

Algorithm for Optimal Denoising of Raman spectra, Chapter 9

In this chapter, a novel software approach to enhance the signal to noise ratio in a Raman spectrum is described. CCD induced noise is typically minimised by using smoothing or filtering techniques, typically Savitzky-Golay smoothing. [11] The SNR of a spectrum fluctuates across the wavenumber range of the signal. For this reason, treating the entire spectrum to the same smoothing process can result in undesirable effects such as the alteration of the underlying spectral features. An alternative algorithm is proposed in this chapter that enhances the Savitzky-Golay method with Maximum Likelihood Estimation in order to limit the smoothing effect in peak regions, thus preserving the spectrum's features while effectively smoothing areas of no diagnostic interest.

Reflective Substrates, Chapter 10

In this chapter, a software independent method of improving the signal to noise ratio of cell spectra is investigated. While not strictly part of the recording system, the substrates that the samples are deposited on can have a significant impact on the performance of the systems. [12] Near infrared (NIR) lasers are a preferable wavelength choice for the inspection of biological samples. Glass is a preferable substrate for use in clinics since it is cheap and disposable. However, when NIR lasers interact with glass they produce large background signal that overpowers the Raman signal produced by the sample. In this chapter, the proposed solution is to use a reflective substrate i.e. a glass slide with a 100 *nm* film of gold on it's surface. These substrates are a relatively cheap alternative to Raman grade calcium fluoride (approximately 2% of the cost) and they also significantly boost the collected Raman scatter by introducing a double-pass of the laser through the sample as well as collecting the forward scattering, which would otherwise be lost. An investigation into the viability of a glass substrate with a gold film is investigated in terms of incorporating this substrate into the sample preparation process for clinical applications and it's affect on the classification accuracy of multivariate statistics. It is shown that this approach can reduce the acquisition time required to achieve reliable classification by as much as a half, due to the improved signal to noise ratio of the recorded spectra.

Conclusion, Chapter 11

The work presented throughout the thesis is reviewed in this chapter and the various contributions are assessed. Suggestions for future work that build on the contributions in this thesis are also provided.

Chapter 2

An overview of Raman spectroscopy

2.1 Introduction

Presented in this chapter, is the basic theory of Raman spectroscopy and the physics behind the phenomenon of the inelastic scattering of light on which Raman spectroscopy is based. Raman spectroscopy is used to record all the spectra presented in this paper and the various contributions presented in this thesis are designed to optimise the signal to noise ratio of data collected using this particular method.

The principle behind Raman spectroscopy was first theorised in 1921 by Sir Chandrasekhara Venkata Raman on an ocean voyage when he observed the scattering of blue light from the ocean and the surrounding glaciers. He published a paper in Nature [13] following a number of experiments to elaborate on the light scattering explanation proposed by Lord Rayleigh. In 1930 he was awarded the Nobel Prize in Physics for his use of a spectrograph to detect what became known as 'Raman lines', resulting from monochromatic light passing through transparent materials. [14]

The invention of the laser, array detectors, and the subsequent price drop as technology developed enough to make these items commercially viable allowed the method to be explored for a number of applications. Raman spectroscopy has been shown to be an effective tool for quality control in the pharmaceutical industry [15], analysis of sample purity in semi-conductor industry [16], and disease diagnosis. [17, 18] This technique does not require the use of additional reagents and does not affect the chemical composition/structure of the sample, making it a desirable tool for diagnostics. The analysis provided in this thesis is primarily to optimise Raman systems and the software used in the post-processing of the resulting spectra for diagnostic purposes in order to boost SNR, reduce acquisition time/increase throughput, and, in some

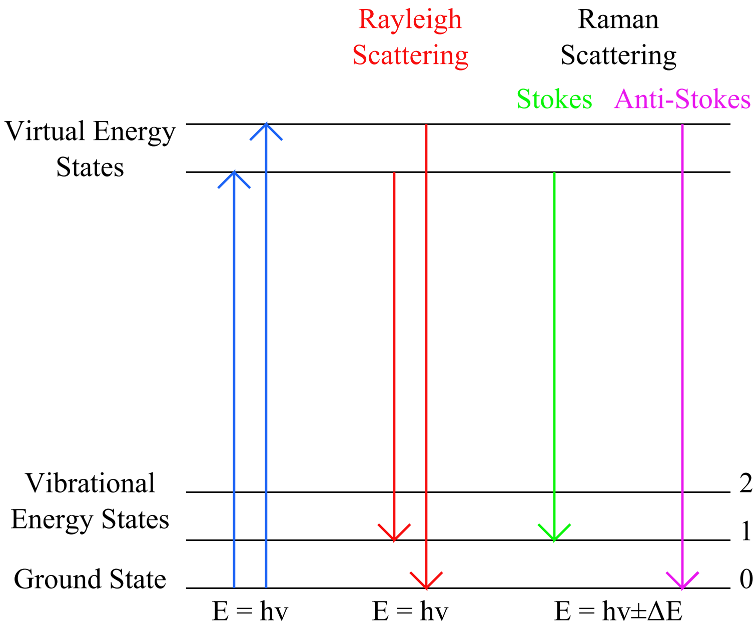


Figure 2.1: Jablonski energy level diagram of Rayleigh and Raman scattering

cases, reduce cost. However, this analysis can arguably be applied to any system that exhibits similar spectral features and noise behaviour. Raman spectroscopy can be described in classical terms or in terms of quantum mechanics. The following work is based on a classical understanding of Raman scattering, as a quantum mechanical interpretation is unnecessary for the scope of this work. [19–23]

2.2 The physics of Raman spectroscopy

The technique of Raman spectroscopy is a powerful tool that can be used to investigate the chemical composition of a biological sample. It is based on the inelastic scattering of monochromatic light that occurs when a photon interacts with a molecular bond. These molecular bonds may exhibit a higher energy vibrational state that possesses a change in polarisability between each of the states. This is known as a selection rule for Raman scattering. These higher energy states and changes in polarisability do not occur in Rayleigh scattering. The mechanics of this phenomenon is illustrated by a Jablonski diagram (Fig. 2.1).

Raman scattering is an inherently weak process with approximately 1 in 10^7 photons undergoing the process. [7] Raman scattered photons will either lose energy or gain energy, the former is referred to as Stokes and the latter as Anti-Stokes. The magnitude of the change in energy, referred to as the Raman shift, is described by the following equation:

$$\Delta\tilde{v}(cm^{-1}) = \left[\frac{1}{\lambda_{incident}(nm)} - \frac{1}{\lambda_{scattered}(nm)}\right] \times \frac{10^9(nm)}{10^2(cm)} \tag{2.1}$$

Where $\Delta\tilde{\nu}$is the change in energy in terms of Raman shift, i.e. wavenumbers (cm^{-1}), and λ is the wavelength (nm) of the incident and emitted photons. This observed shift in magnitude depends on the rotational and vibrational energies of the interrogated molecules in the sample. [24] This will be further examined in Section 2.3 and Section 2.4.

By collecting and analysing the Raman scatter produced by a sample, it is possible to identify and quantify the chemical bonds, which in aggregate allow a chemical 'fingerprint', to be obtained for the sample. Symmetric spectra are produced corresponding to a positive and negative Raman shift, which are referred to as the Stokes and Anti-Stokes spectra. Typically, only the Stokes photons are collected since they are numerous in comparison to their Anti-Stokes counterparts. This is due to the intensity of each spectrum being dependent on the population density of the initial energy state. For any state in thermal equilibrium the ground state will be more populous than the first excited state and as such the Stokes spectra will be more intense. The relative intensities of the Anti-Stokes and Stokes spectra may be used to calculate the temperature using Boltzmann's distribution. [25]

2.3 Rotational Raman spectroscopy

A molecule must be anisotropically polarisable in order to undergo rotational Raman scattering. [26] By applying an electrostatic field to the molecules being analysed and examining it's impact on it's electric charge distribution it is possible to determine the polarisability of the molecule in question. This property indicates how readily it's electron distribution can be distorted by the applied field. A first order dipole moment, described by Equation 2.2, is induced.

$$\mu_{induced} = \alpha E \tag{2.2}$$

where $\mu_{induced}$ is the dipole moment induced, α is the polarisability of the molecule, and E is the amplitude of the applied electrostatic field.

However, an external electric field resulting from a laser will induce a second order dipole moment given by:

$$\mu_{induced} = \alpha E(t) = \alpha E_0 cos(\omega_i t) \tag{2.3}$$

where t is the time and ω_i is the angular frequency of the electric field. Second order dipole moments are very weak but detectable. Higher order dipole moments can be observed in very intense electric fields. These are referred to as nonlinear optical phenomena.

Variations in polarisability ($\Delta\alpha$) are time dependent for anisotropic molecules that are rotating with an angular frequency of ω_r. α will rotate through 2π taking on a range of values in the range $-\Delta\alpha < \alpha_0 < \Delta\alpha$, resulting in Equation 2.4.

$$\alpha(t) = \alpha_0 + \Delta\alpha cos(2\omega_r t) \tag{2.4}$$

By combining equations 2.3 and 2.4 together it is possible to derive an expression for Rayleigh, Stokes, and Anti-Stokes scattering.

$$\mu_{induced} = \alpha_0 E_0 cos(\omega_i t) + \frac{1}{2} E_0 \Delta\alpha[cos(\omega_i t + 2\omega_r t)] + \frac{1}{2} E_0 \Delta\alpha[cos(\omega_i t - 2\omega_r t)] \tag{2.5}$$

The first component describes the Rayleigh scattering element, the second describes Anti-Stokes Raman scattering, and the third Stokes Raman scattering. Once a 180° rotation has been completed the distortion of the molecule by the applied electric field dissipates, returning the molecule to it's initial state. If $\Delta\alpha$ is equal to zero then the molecule is not anisotropically polarisable and therefore the Stokes and Anti-stokes elements of Equation 2.5 disappear.

2.4 Vibrational Raman spectroscopy

All diatomic molecules are vibrationally Raman active due to how their polarisability changes as they vibrate. [27] It is necessary to apply group theory in the case of polyatomic molecules to determine if it is vibrationally Raman active. [28] The theory behind vibrational Raman and rotational Raman is similar. However, in the electron distribution there also occurs a periodic compression and stretching, that causes the polarisability of the molecule to oscillate in the direction of the electric field.

Equation 2.4 can be modified to demonstrate that the induced dipole oscillates in phase with the molecules vibrational motion.

$$\alpha(t) = \alpha_0 + \delta\alpha cos(2\pi v_{vib} t) \tag{2.6}$$

where $\delta\alpha$ is the change in amplitude of the polarisability over one vibrational cycle, v_{vib} is the vibrational frequency in Hertz (Hz), and all other values are as previously described. In a similar fashion to that described in the previous section, if the applied electric field results from a laser source, Equations 2.4 and 2.6 may be combined, as shown in Equation 2.7, to define an expression for the induced time dependent dipole moment.

$$\mu_{induced} = \alpha_0 E_0 cos(2\pi v_i t) + \frac{1}{2} E_0 \delta\alpha [cos(2\pi\{v_o + v_{vib}\}t)] + \frac{1}{2} E_0 \delta\alpha [cos(2\pi\{v_o - v_{vib}\}t)] \tag{2.7}$$

where v_i is the frequency of the incident laser and v_o is the oscillating frequency. As in Equation 2.5, the first, second, and third section of this equation is relevant to Rayleigh, Anti-Stokes, and Stokes scattering respectively.

2.5 Roto-vibrational Raman spectroscopy

Vibrational transitions can often induce a change in the rotational quantum number, as the angular momentum must be conserved when photons are either emitted or absorbed. Raman spectra infrequently reveal this phenomenon, unless the molecule in question is diatomic. This process can be represented spectrally, in terms of quantum mechanics, using the Born-Oppenehimer approximation. [29]

2.6 Raman spectrometer for experimental purposes

A simple Raman spectrometer consists of a monochromatic excitation source (i.e. a laser), a set of lenses and mirrors to deliver the laser to the sample, a set of lenses and mirrors to collect the Raman scatter while filtering out any Rayleigh scatter, and finally a detection module (i.e. an array detector and spectrograph combination). Raman systems will also typically have an imaging system connected to the spectrometer in order to focus the laser onto a sample such as a slide with cells deposited on it. A more complex version of the setup described here can be seen in Fig. 2.2. This system is designed in back scattering configuration where the laser is delivered and collected through the same microscope objective i.e. part of the collection path is shared with the delivery path. The system schematic depicted is typical of the systems used to collect data in this thesis. Variations in optical elements in the systems, e.g. CCD or microscope objective, are detailed or referenced in their particular sections.

Each optical element in the spectrometer must be chosen and aligned for optimal laser throughput and to collect the maximum scatter produced by the sample. The reason for this is to ensure there is maximal SNR,

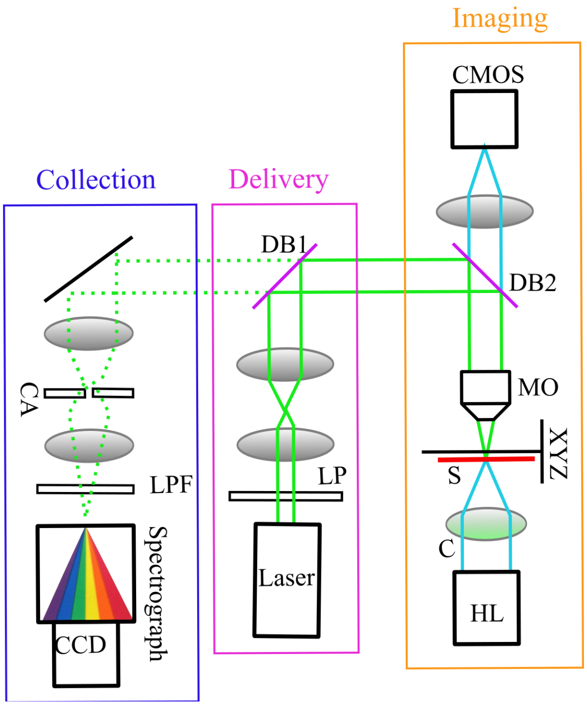


Figure 2.2: Diagram of the basic confocal Raman micro-spectrometer set-up. Similar in configuration to all systems used to collect the data presented in this thesis.

Abbreviation	Component	Function
CCD	Charge Coupled Device	Collect Raman scattered photons
LPF	Low Pass Filter	Remove Rayleigh scatter
CA	Confocal Aperture	Define where scatter is collected from
DB	Dichroic Beamsplitter	Reflects certain wavelengths and transmits others
LP	Line Pass Filter	Remove laser sidebands
CMOS	CMOS camera	View sample topography
MO	Microscope Objective	Deliver laser and collect scatter
XYZ	Translation stage	Move sample relative to the laser spot
S	Biological sample	Object to interrogate
C	Colour filter	Remove interfering wavelengths from the hot lamp
HL	Hot lamp	White light to view the sample

Table 2.1: List of components depicted in Fig. 2.2, from left to right and top to bottom of the diagram.

which can be challenging due to the weak Raman signal. This is further explored in Chapter 4. A basic diagram of a confocal Raman system can be found in Fig. 2.2 and descriptions of the system components can be found in Table 2.1.

Once a Raman spectrum has been collected from the sample, there are a number of mathematical procedures that must be applied to the data in order to ensure reproducibility across datasets and train multivariate statistical models to classify the data. A number of numerical methods used for this purpose are described in the following chapter. These methods are used throughout the thesis to process the recorded Raman spectra.

Chapter 3

Numerical methods for post-processing of Raman spectra

3.1 Introduction

Once a Raman spectrum has been recorded using the methods described in the previous chapter, there are a number of procedures that are necessary in order to ensure comparability of spectra across experiments and to evaluate the viability of the data for classification purposes. This chapter describes the algorithms that are typically applied to Raman spectra and are used throughout this thesis for the purposes of signal correction and the multivariate statistical analysis that are ultimately applied to the data in order to classify the spectra. Noise present in the experimental spectra can affect these procedures. In Chapter 5 and 10 in particular, the effect of the signal to noise ratio on the multivariate statistical analysis presented in this chapter is examined.

Three distinct components exist within a Raman spectrum; i) Collected Raman scatter, ii) baseline signals, and iii) noise. Noise and unwanted baseline signals can have their presence reduced through the use of software processing techniques in order to prevent the possibility of increased diagnostic error during the classification process. Various procedures for reducing the presence of noise will be discussed in the following chapters of this thesis. All other procedures are described in this chapter. After the pre-processing has been applied to the spectra, the data can be used to train and test classifiers created using multivariate statistical analysis, in particular the methods described in Sections 3.5.1, 3.5.2, and 3.5.3.

3.2 Calibration

During the course of a lengthy set of experiments, various components in the Raman spectrometer may 'drift' due to thermal or vibrational effects. This slight alteration to the system can produce variations in the Raman spectrum relative to the expected wavenumbers of particular peaks. Variations in optical components, e.g. optical filters and diffraction grating, may also have an impact on the intensity of the collected Raman lines from a given Raman spectrometer. It is therefore necessary to accurately calibrate the system to ensure a meaningful comparison of spectra obtained from different Raman systems. Calibration can either be applied to the system itself or retro-actively applied to recorded spectra. In order to calibrate a dataset the wavelength axis should be calibrated, followed by intensity calibration, and then wavenumber calibration.

Each column of pixels in the array detector is assigned exact wavelength positions to perform wavelength calibration. Non-linearity in the wavelength dispersion across the CCD caused by the spectrograph grating can result from two distinct causes; i) the sine function that describes wavelength dispersion, and ii) distortions that are present in all spectrograph optical designs. In principle, a one- or two-point calibration could be sufficient to compensate for the former cause of non-linearity. However, multi-point calibration is required to correct adequately for distortion and permit spectrograph-to-spectrograph comparisons. [30] A spectrum from a known sample that contains many spectral peaks, e.g. a Neon lamp, is measured. The position of each peak is calibrated against a reference spectrum whose spectral lines contain an exact wavelength position. [31] Wavelength-dependent transmission efficiency of the optical elements within the system, as well as the sensitivity of the detector, are then corrected by intensity calibration. This form of calibration may be accomplished using a National Institute of Standards and Technology calibrated white light source. [32] Wavenumber calibration is particularly important if the source wavelength is not accurately known. This type of calibration assigns a particular wavenumber shift to each column of pixels in the CCD camera. [33] A known sample, e.g. silicon, is recorded, and the position of the peak(s) in the resulting spectrum are calibrated to coincide with the precisely known Raman peak positions for the relevant calibration sample.

3.3 Normalisation

Normalisation is performed in order to provide a common scale for comparing Raman peaks across a range of spectra. This can be achieved by dividing each sample in a spectrum by some constant. Normalisation methods that can be applied to Raman spectra include; peak normalisation, vector normalisation, area normalisation, and min-max normalisation. [34] Peak normalisation is achieved by measuring the constant as the height difference between the baseline and the maximum point of a chosen peak; vector normalisation obtains the constant value by calculating the sum of the intensity values for each variable in the spectrum, and finding the square root of this value; for area normalisation, the constant corresponds to the sum of the intensity values for each variable in the spectrum; and min-max normalisation involves subtracting the minimum and dividing by the maximum so that the spectrum is scaled between 0 and 1. [35,36] Min-Max normalisation is primarily used in this thesis in advance of multivariate statistical analysis.

3.4 Background subtraction

Various phenomena such as fluorescence induce uneven amplitude shifts across different wavenumbers. Before proceeding with further analysis it is necessary to compensate for these amplitude shifts. There are a number of proposed procedures to do this [37–40], however, the method used in this thesis is a least squares fit of a background spectrum and/or polynomial to remove the effect of varying baseline signals known as an extended multiplicative signal correction (EMSC). [41–43]

Explicitly, the EMSC algorithm computes an optimum baseline made up of an N order polynomial and a weighted background signal, typically from the substrate, that is recorded at the beginning of each experiment. The algorithm applies a least squares fit to (i) a reference Raman spectrum (in this thesis, the reference spectrum is typically the mean of the dataset being fitted), (ii) the contaminant signal, and (iii) an N order polynomial. Following an Ordinary Least Squares (OLS) fit, the weighted components of (ii) and (iii) are subtracted from the raw spectrum. The value of N is dataset dependent and it has been shown elsewhere that the use of high values of N does not result in over-fitting [38]. Mathematically, EMSC can be defined through the following equations. [42] It is assumed that the raw Raman spectrum is comprised of three distinct components.

$$X_0 = R + B + P \tag{3.1}$$

where X_0 is the raw spectrum that can be described as a sum of R, the Raman spectrum of interest, the baseline signal, P, and the contaminant signal, B. The principle goal of the EMSC algorithm is to accurately estimate the concentration of B and P present in the spectrum so that they can be reliably separated from R. It can be assumed that R can be approximated by scaling the reference spectrum, r, by a certain weight.

$$R \approx c_r \times r \tag{3.2}$$

where c_r is the scaling factor for a particular spectrum. In a similar manner, it may be assumed that spectral contribution of the background signal, B, from the substrate may be approximated by the product of the recorded contaminant signal, b, by a particular weight, c_b.

$$B = c_b \times b \tag{3.3}$$

P, the baseline signal, can be represented by a polynomial of an appropriate order, N.

$$P_N = c_0 + c_1 + c_2^2 ... c_N^N \tag{3.4}$$

where c_m for $m = 1, 2..N$ are the polynomial coefficients. [44]

X_0, r, b, and N are inputted into the EMSC algorithms and estimates for c_r, c_b, and c_m, for all m, are returned based on an optimal fit using OLS. [38, 43]

$$X_0 \approx [c_r \times r] + [c_b \times b] + \sum_{m=0}^{N} c_m^m \tag{3.5}$$

The final background corrected signal is given by:

$$X_{final} = \frac{X_0 - [c_b \times b] - [\sum_{m=0}^{N} c_m^m]}{c_r} \tag{3.6}$$

It should be noted that in a number of areas in the thesis, recording a background contaminant spectrum from a substrate was not possible as the spectra were recorded from a polymer spectrum. In these cases an EMSC fit and polynomial were performed without the use of a background signal. This does not affect the functionality of the EMSC algorithm described here. This algorithm is used to process spectral datasets in Chapter 10, 8, and 9.

3.5 Multivariate statistical analysis

This section provides information on the multivariate statistical analysis that is commonly applied to Raman spectroscopic data for classification. [45–48] This involves the application of pattern recognition techniques, such as Principal Components Analysis (PCA) or Linear Discriminant Analysis (LDA), in order to identify subtle changes across datasets that can be used to accurately differentiate between different pathological groups and subgroups. This typically involves training a multivariate statistical algorithm on known samples and using the resulting model to classify unknown samples. The algorithms described in the following subsections are applied in Chapters 5 and 10, in particular PCA.

3.5.1 Principle component analysis

PCA is an unsupervised statistical method that is based on the transformation of spectral data into a set of variables commonly referred to as principle components (PCs), in order to reduce the number of variables within a dataset. All PCs are orthogonal to one another and are generated such that they represent the maximum amount of variance possible within the dataset. To do this, the line of best fit through the plotted data points is found i.e. the direction along which the maximum variance is explained. This is known as the first PC coefficient. [49] Subsequent PCs are orthogonal to the previous PCs yet still describe the maximum remaining variance in the data.

PCs are found computationally by calculating the eigenvectors and eigenvalues of the covariance matrix of the dataset. The eigenvalues are sorted by decreasing magnitude and their corresponding eigenvectors represent the first PC, second PC, third PC etc from largest eigenvalue to smallest. Following this it is possible to consider each sample dataset as the product of a linear sum of PCs multiplied by a scalar, referred to as a score. [50] This technique enables the user to minimise the variables present in a given dataset while retaining almost all of the spectral information. The advantage of this tool is that it allows faster classification algorithms to be designed since it creates a simpler representation of the data.

Further analysis can be applied to the PCs to cluster the data into groups that represent different biological samples e.g. LDA, that is described in the following section.

3.5.2 Linear discriminant analysis

Also referred to as Fisher's discriminant analysis, LDA is a supervised multivariate technique. It operates on the basis of finding the direction of variance that provides the best separation for multiple groups of data in order to optimise class separability. This technique is primarily applied to PC scores so as to further reduce the dimensionality of the dataset. This is achieved by finding a linear combination of vectors that maximise the variance within the datasets, in a similar manner to PCA. However, LDA also determines the component vectors that maximise the separation between classes, therefore, improving the classification results.

	Classified as:	
	Cancerous	Healthy
Cancerous samples	TP	FN
Healthy samples	FP	TN

Table 3.1: Example of how samples may be classified using PCA-LDA followed by a LOO cross validation. The abbreviations represent; TP, true positive; FN, false negative; FP, false positive; TN, true negative.

Dimensionality reduction, such as the processes described here, can help minimise the error in parameter estimation, thus, avoiding over-fitting. It can also significantly reduce the computational cost of classifying data.

3.5.3 Cross validation

The accuracy of a classification model can be estimated by using cross validation to assess the results of the statistical algorithm that has been applied to data. Leave-one-out (LOO) is the most common form of cross validation. It is based on isolating a spectrum from a known dataset of spectra, using the remaining data to train the statistical algorithm, and validating the trained algorithms performance by applying it to the isolated spectrum. This process is iteratively applied, by isolating each spectrum in turn in the dataset. The classification results can be defined using four labels, as described in Table 3.1 using 'cancerous' and 'healthy' as an example for an arbitrary dataset containing spectra recorded from both types of cells.

The four labels described in Table 3.1 can be used to compute the sensitivity and specificity of the classification algorithms performance. The sensitivity of the algorithm can be simplistically defined as how good the algorithm is at being correct and can be defined in mathematical terms as:

$$sensitivity = \frac{TP}{TP + FN} \tag{3.7}$$

Specificity can be defined in simplistic terms as how good the algorithm is at excluding the incorrect parameters. Mathematically defined as:

$$specificity = \frac{TN}{TN + FP} \tag{3.8}$$

where the variables in Equations 3.7 and 3.8 are as defined in Table 3.1.

3.6 Summary

In this chapter the typical numerical methods for the processing of Raman spectra are described. It is necessary to apply these procedures in a consistent manner to ensure reproducible results across experiments, especially those that are performed across different Raman systems. These procedures can be affected by the noise present in the system, in particular the results of multivariate statistical analysis. In cases where systems are performing sub-optimally it is advantageous to model the sources of noise present in the system to analyse how these noise sources may be minimised. The principle sources of noise in a Raman system and their associated probability distributions are presented in the following chapter. Also presented in the next chapter is methodology for discerning the irradiance collected by a specific system from a specific sample. These irradiances allow the experimentalist to create artificial datasets that can be used for a variety of optimisation procedures that are detailed in Chapters 5-9.

Chapter 4

Signal to Noise Ratio in the Context of Raman Spectroscopy

4.1 Introduction

This chapter presents the methods for modelling each specific noise source in a Raman spectroscopy system, which then forms the basis of the methodologies presented in the following chapters is based. The methods for modelling the noise that are presented in the following sections can be used to create artificial datasets that can be used to simulate spectra recorded under a variety of experimental conditions. Consequently, the effect of system parameters, experimental procedures, and denoising algorithms on the signal to noise ratio (SNR) may be examined without the need for performing the extensive experiments that would otherwise be required for validation.

There are a number of principle noise sources present in a Raman spectrometer. The following sections will explore these noise sources and how they affect the SNR of the recorded spectra. Shot noise, the quantum efficiency of the camera, dark current, read noise, and cosmic rays are noise sources that can create challenges for recording spectra that are suitable for classification. In the following sections, the probability distributions that are used to model these types of noise as well as the basic procedures that may be used to minimise their impact on the SNR. Section 4.3 details the two methods employed in this thesis for calculating SNR in experimental spectra and the benefits of each method. Section 4.4 elaborates on the theory discussed in the previous sections to show how the theory may be used in order to simulate artificial Raman spectra.

4.2 Fundamentals of Noise

4.2.1 Shot Noise

Shot noise is the name given to the difference between the mean number of photons that are collected per unit area (i.e. per pixel) per unit time and the true irradiance. A charge-coupled device (CCD) is a device capable of converting photons to electrons and storing/transferring the accumulated charge within a sequence of capacitive 'wells' or pixels. CCD pixels are metal oxide semiconductor capacitors located on a photo-active layer of silicon. When a pixel is exposed to light, electrons accumulate in a potential well, i.e. a pixel, in an amount that is proportional to the intensity of the light that is incident on the pixel. Each pixel is associated with a specific Raman wavenumber, as the irradiance intensity varies across the entire spectral range and so for a single pixel the incident photons are converted into electrical charge at a mean rate that is proportional to the mean flux of light per unit area or 'irradiance', i. [51] Please note that while the convention is to express irradiance in Watts per square metre that this work will solely be considering photons. For a given value of i the probability, p, of collecting n_{ph} photons per unit area, per second is described by a Poisson distribution as follows:

$$p(n_{ph};i) = P(n_{ph};\mu = i) \tag{4.1}$$

where $P(n_{ph};\mu = i)$ is explicitly defined by:

$$P(n_{ph};i) = \frac{i^{n_{ph}} \exp(in_{ph})}{n_{ph}!} \tag{4.2}$$

A Poisson distribution is defined only by the mean of the distribution, μ, which in this case is equal to the mean irradiance per unit time, i. A key property of the Poisson distribution is that the variance and mean are equivalent, and therefore the standard deviation is given by a square root relationship to the mean. Shot noise can be quantified by the standard deviation of the Poisson distribution above, $\sqrt{i}$. [52, 53]

4.2.2 Quantum Efficiency

The quantum efficiency (QE) of a pixel is the ratio of photons that are successfully converted to electrons in the array detector and is wavelength dependent. The QE of a given CCD is usually defined as a function of wavelength in the camera specifications. The physical effect of the QE for a particular wavelength, $q(\lambda)$, is mathematically described by the binomial distribution. The probability of converting n_{ph} photons into n_{pe}

electrons is given by the following binomial distribution:

$$B(n_{pe};n_{ph},q(\lambda)) = \binom{n_{ph}}{n_{pe}}[q(\lambda)]^{n_{pe}}[1-q(\lambda)]^{n_{ph}-n_{pe}} \tag{4.3}$$

where:

$$\binom{n_{ph}}{n_{pe}} = \frac{n_{ph}!}{n_{pe}!(n_{ph}-n_{pe})!} \tag{4.4}$$

The QE limits the ability of the CCD pixel to detect the incident mean flux. The overall probability of detecting n_{pe} electrons (for a given i) can be determined by first considering the probability of collecting n_{ph} in the pixel, followed by the probability of converting these n_{ph} photons into n_{pe} electrons. The overall probability is given by convolving the two independent probability distributions defined in Equations 4.1 and 4.3:

$$\begin{aligned} p(n_{pe};i,q(\lambda)) &= \sum_{n_{ph}=n_{pe}}^{\infty} P(n_{ph};i)B(n_{pe};n_{ph},q(\lambda)) \\ &= \frac{[iq(\lambda)]^{n_{pe}}e^{-iq(\lambda)}}{n_{pe}!} \\ &= P(n_{pe};iq(\lambda)) \end{aligned} \tag{4.5}$$

4.2.3 Dark Current Noise

Dark current is the name given to thermally generated electrons at surface states, within the bulk silicon, and in the depletion region of the CCD. Dark current can also be described by a Poisson process with mean rate of generation, c, per unit area per unit time. [51,52] c, in electrons per pixel per second (e/p/s), depends primarily on temperature and can be reduced by cooling the CCD. [54] The amount of dark current present in the signal is also dependent on the area of the photo sensor therefore smaller pixels will exhibit a lower mean rate of dark current generation with a caveat that they will also have a reduced signal irradiance since this is also a function of area. This source of noise is also time dependent, thus, increasing with exposure time. Dark current noise will increase with integration time with an initial value of zero as the CCD is cleared of spurious charge immediately before an image capture. Based on the reproductive properties of the Poisson distribution, independent Poisson events occurring concurrently can be modelled by a single Poisson distribution. [55] The sum of two independent Poisson distributions with mean values, μ_1 and μ_2, respectively may be described as a single Poisson distribution with mean of $\mu = \mu_1 + \mu_2$, which can be proved using convolution as described by Equation 4.6.

$$\begin{aligned} p(n_{pe};i,q(\lambda),c) &= \sum_{i=0}^{n_{pe}} P(n_{ph};iq(\lambda))P(n_{dc};c) \\ &= \frac{(iq(\lambda)+c)^{n_{pe}} e^{-(iq(\lambda)+c)}}{n_{pe}!} \\ &= P(n_{pe};iq(\lambda)+c) \end{aligned} \tag{4.6}$$

where $p(n_{pe};i,q(\lambda),c)$ denotes the probability of accumulating n_{pe} electrons in a pixel well per unit time given a mean rate of flux, i, mean rate of dark current generation, c, and the QE at the photons wavelength, $q(\lambda)$. All of the analysis thus far is based on a single pixel for a unit of time, 1 second. However, given the nature of shot noise and dark current they increase over acquisition time, therefore, it is necessary to amend the analysis to represent this. The integration time, t, can be included in the combined Poisson distribution described in Equation 4.6 as a linear scaling factor:

$$p(n_{pe};i,q(\lambda),c,t) = P(n_{pe};(iq(\lambda)+c)t) \tag{4.7}$$

The standard deviation of the noise thus far is given by $\sqrt{(iq(\lambda)+c)t}$. The primary method used, aside from manufacturing, to reduce the amount of dark current generated by the CCD is to cool the device. This may be done by air cooling, water cooling, or in some cases by using liquid nitrogen, depending on the make and model of CCD in question.

4.2.4 Read Noise

Read noise is largely generated by the electronics of the CCD when shifting the charge carriers out of their pixels and converting them into a quantifiable voltage. Read noise can be described as a 'catch-all' term for a combination of various inherent noise sources from the photo-receptor to the Analogue to Digital Converter (ADC), including noise contributions from the shift register (charge transfer efficiency), pixel reset noise, clock induced charge, as well as quantisation noise. [56] Charge Transfer Efficiency (CTE) estimates the percentage of electrons that will successfully be transferred from one pixel to the next during readout. [52, 57] Low CTE will result in residual charge remaining in the pixels and a reduction of signal. CTE relies largely on the manufacturers production process, however, fast pixel shifts when moving charge from the pixel wells to the readout register can cause the CTE to degrade. Reset noise is a form of temperature dependent noise that is caused by the resetting of the CCD to a reference voltage at the beginning of each acquisition. Quantisation noise occurs due to the digitising process of the ADC. When a pixel value is digitised the ADC approximates the analogue voltage to the nearest quantisation level and this introduces an uncertainty into the value known as quantisation noise. [58] Clock induced charge (CIC) is generated when a CCD is clocked into inversion and holes become trapped in the silicon interface of the CCD. When the clock reverts to the non-inverted state, the holes are accelerated away from the interface, and may generate spurious electrons through impact ionisation. Fast clock rise times, slow clocking, and high clock swing amplitudes all contribute to CIC generation. [59] Minimisation of this phenomenon is primarily restricted to manipulating the manufacturing process of the camera and is considered to be the limiting factor of single photon detection.

Read noise is both time and signal independent and determines the fundamental limit of detecting a single photon where effective read noise is a significant contributing factor i.e. greater than one electron per pixel. Read noise is modelled by a normal distribution. [51] The probability of registering n_{sig} electrons in a pixel is defined as follows:

$$p(n_{sig}; i, q(\lambda), c, t, \mu, \sigma) = P(n_{pe}; [iq(\lambda) + c]t) + N(n_r; \mu, \sigma) \tag{4.8}$$

where the normal distribution is explicitly defined as:

$$N(n_r; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)}{2\sigma^2}} \tag{4.9}$$

While the number of electrons generated by readout, n_r, is unaffected by time and signal independent it is strongly read-out rate dependent as well as gain dependent [60] and these parameters define the mean of the normal distribution, μ, and the standard deviation, σ.

It is possible to minimise read noise through use of low read out rates or through use of Electron Multiplying (EM) CCDs. [61] EMCCDs are designed to make read noise negligible through the addition of a gain register prior to the electronic readout. While the gain registers produce their own noise, due to the imperfect scaling of the signal [62, 63], it is possible to render the effective read noise to be one electron per pixel by setting the gain to be equal to the expected number of read noise electrons produced.

4.2.5 Cosmic Ray Artefacts

Cosmic Ray Artefact (CRA) contamination is a non-optical noise source that occurs when recording spectra using any photo-electric device, such as a charge-coupled device (CCD). These intermittent events are caused by high energy particles interacting with the detector [64], the effect of which is to release large numbers of electrons that are indistinguishable from photo-electrons. These 'packets of energy' primarily consist of high speed protons which produce secondary particles on collision with Earth's stratosphere. Both the primary and secondary artefacts manifest themselves as ionisation trails on array detectors. [54] CRAs are randomly distributed in time and intensity and are generally localised to a small number of adjacent pixels in an array detector, although, they may in some cases have a broader width. CRAs can be especially prominent when the spectral irradiance is weak, such as for the case of Raman spectra recorded from biological samples, which necessitates a detector that is sensitive to low photon counts and the utilization of long camera integration times. Preventative measures can be taken to reduce the possibility of CRA interference by reducing the effective CCD aperture and/or reducing the exposure time. Reducing the effective CCD aperture, or pixel region, from which the spectrum is extracted is possible if the spectral irradiance is incident on a portion of the available pixels. In this case, the spectrum may be read out as an image or in crop mode i.e. a region of interest read out. The different types of camera read mode and how they affect the SNR of spectra is discussed in greater detail in Section 6.2. Reducing the integration time limits the window of time in which a cosmic ray can manifest. Long exposure times can result in multiple cosmic rays contaminating spectra, however, this is not always feasible.

4.2.6 Other Effects

There are four other effects of optical elements that can impact the SNR of the spectrum being collected; charge traps, source wavelength, microscope objective, and the spectrograph grating. The first source of noise does not affect the collected irradiance but the latter three do. For the purposes of this thesis, these sources of noise are not considered as there are no charge traps in the CCD used in the experiments and the majority of modelling is based on estimates of system specific irradiance, see Section 4.4.1. However, it is necessary to acknowledge these potential sources of noise, especially those that affect the collected irradiance, as the modelling presented in Chapter 7 may be extended to include the effect of these optical elements.

4.2.6.1 Charge Traps

A charge trap is introduced through the manufacturing process of the CCD caused by lattice defects in the silicon wafer or impurities that can result in a section of the CCD where the CTE is significantly lower than expected. Charge traps can be localised to pixels defects, such as hot pixels, or can be associated with a particular column/row of the CCD. Another notable effect is blooming which is a phenomenon which occurs when a pixel is saturated (exceeds it's full well capacity), often as a result of charge trapping, and begins to leak charge into adjacent pixels causing a white spot to 'bloom' on the CCD. [65] This phenomenon is generally removed manually in post-processing by averaging across the affected region or by using an image read out mode and excluding the affected area.

4.2.6.2 Source Wavelength

The source wavelength is an important consideration when obtaining Raman spectra from biological specimens. Tissue and bio-fluid are preferably examined in the NIR region as there is greater penetration depth into samples at this wavelength due to the lower levels of absorption. [66] However, lasers with longer wavelengths produce significantly less scattering photons, which is equivalent to reducing the power of the laser. [67] The intensity of Raman lines are proportional to the fourth power of the laser frequency, as detailed in Equation 4.10. [68]

$$\left(\frac{v_{532}}{v_{785}}\right)^4 = \left(\frac{c}{\lambda_{532}} \times \frac{\lambda_{785}}{c}\right)^4 = 4.74 \tag{4.10}$$

where v represents the laser frequency, c represents the speed of light, and λ represents the laser wavelength. The above equation shows that for a given acquisition time the Raman lines collected from a 532nm source laser will be approximately 5 times as intense as those collected from a 785nm source. Therefore the scattering efficiency is greater at the lower wavelength. This reduction in photon counts from the visible region to the NIR region necessitates longer acquisition times and therefore a trade-off between the collected irradiance and the production of dark current must be made. QE can also vary greatly with source wavelength and therefore an array detector that has been optimised for the laser's corresponding wavelength region will provide spectra with the highest SNR in the shortest acquisition time.

4.2.6.3 Numerical Aperture and Microscope Objective

The Numerical Aperture (NA) of a Microscope Objective (MO) is a measure of the range of angles over which light is collected by the objective and effects the irradiance ultimately detected by the CCD. Consequently, this component affects the shot noise and, thus, the SNR of the system. Equation 4.11 is given in terms of NA, refractive index, and the solid angle.

$$NA = nSin(\theta) \tag{4.11}$$

Where *n* is the refractive index of the medium between the sample and the objective (air if a dry objective, otherwise oil or some immersion medium with a refractive index greater than 1) and θ is the angle, from the normal, of light that is collected by the objective i.e. the solid angle. The maximum NA achievable for an objective in air is 1, however, an immersion lens can exceed this due to the refractive index of the immersion medium increasing the solid angle. The role of the MO, in terms of NA and magnification, as well as its relationship with the diameter of the confocal aperture, on SNR, will be discussed in more detail in Chapter 7, in which we consider the effect of replacing different elements with the spectrometer and estimating the effect on the SNR.

4.2.6.4 Diffraction Grating

Finally, the grating can also negatively affect the collected irradiance, depending on it's blaze angle and mounting configuration. [69] The grating has a periodic structure that splits and diffracts light into multiple beams travelling at different angles using Huygens-Fresnel principle. [70] The directions of the diffracted beams depends on the spacing of the grating lines and on the wavelength of the light. Arguably, the most commonly used method of estimating the grating efficiency is by calculating the blaze condition of the grating. Many grating specification sheets will define the first-order Littrow blaze wavelength for which the efficiency is maximal for the grating mounted in the Littrow configuration. [71]

4.3 Calculating Signal to Noise Ratio

From the noise sources outlined in the previous sections, the SNR for a single pixel can be expressed as the following:

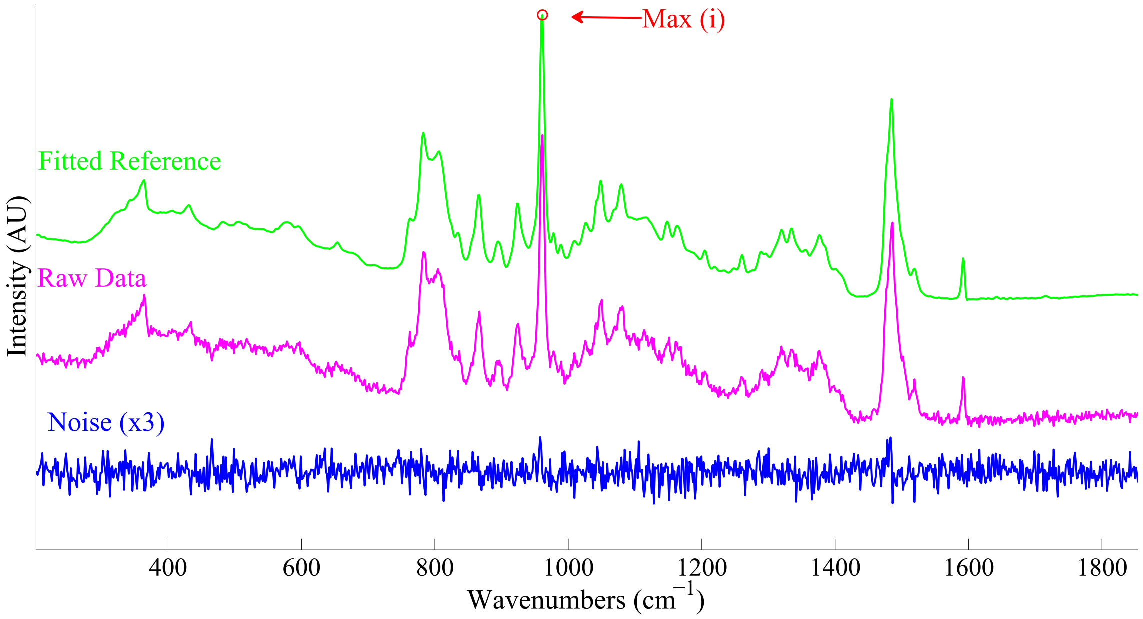


Figure 4.1: Estimating SNR in experimental data

$$SNR_{pix} = \frac{iq(\lambda)t}{\sqrt{[iq(\lambda)+c]t+n_r^2}} \tag{4.12}$$

where, more specifically, i represents the mean irradiance that is incident on the pixel, $q(\lambda)$ represents the QE at the wavelength (λ) in question, and c is the mean rate of dark current generation, t is the exposure time of the capture, and n_r is the standard deviation of the read noise.

4.3.1 Estimating SNR in Experimental Spectra

Accurately modelling the noise in experimental Raman spectra can be challenging as not only does the SNR fluctuate across the wavenumber range being examined but it is challenging to accurately measure the shot noise present in the spectra as this changes from spectrum to spectrum. The same method depicted in Fig. 4.1 [72] may be applied in two separate contexts.

The first method is to create a reference spectrum by applying Savitzky-Golay (SG) smoothing [11], typically a polynomial order of 3 and a window size of 9. This filtered spectrum may then be subtracted from the experimental spectrum to create the noise spectrum shown in Fig. 4.1. The maximum value in the spectrum is the divided by the standard deviation of the noise signal.

The second method may be applied if there is a suitably large dataset of spectra with little variance in spectral profiles e.g. polymer spectra. Once a reference spectrum has been established by taking the mean of the dataset, a least squares fit [42] of the reference spectrum to the data can be performed. The fitted reference can then be subtracted from the data to create the noise signal and the same procedure as described in the previous paragraph. In the case of a dataset that has significant variances across the dataset due to sample heterogeneity or morphology it is recommended to use the first approach as it will not be significantly affected by differences in intensity of spectral profiles.

4.4 Creating Artificial Datasets Based on Experimental Data

When attempting to predict system response for changing system parameters or the effectiveness of a denoising algorithm on collected Raman spectra, artificial datasets are very useful. They are relatively uniform i.e. there are no varying baselines, they are not affected by changing sample morphology e.g. thickness of the cells or photo-bleaching, and the datasets can be created to the users specifications i.e. specific SNRs and numbers of spectra. Another important advantage of these types of datasets is that they can be compared to the reference spectrum that the dataset is based on. This allows the user to reliably measure the effect of the algorithm and/or system conditions in terms of SNR. In this section, a method for estimating system specific irradiance, i.e. the irradiance that may be expected for a given spectrometer configuration over 1 second acquisition time, is described. It is then shown how this irradiance may be used to simulate experimental conditions based on camera parameters and acquisition time.

4.4.1 Estimating System Specific Irradiance

In order to create suitable datasets, it is necessary to first create a reference spectrum based on a typical sample that the system is being used to evaluate. The first step is to take a large dataset, X, of the target sample collected on the same system. Provided that the system is regularly maintained to the same standards, a previously recorded dataset may be used. By averaging together this dataset, a low noise 'ideal' spectrum can be established, x_{mean}. In order to remove the mean dark current and read noise a suitably large dataset of dark current backgrounds is also averaged together to create DC_{mean} that is subtracted from x_{mean} to isolate the irradiance signal from the rest of the spectrum. The irradiance signal then scaled down to 1 second by dividing by the acquisition time for which the signal was recorded for. This signal is in units of 'counts'. Counts are the units used by the CCDs ADC. The ADC will register a number of electrons in the shift register as a single count. Therefore, in order to relate the spectral intensity to collected photons, it is necessary to convert the irradiance back into electrons using the values cited in the CCD specification sheet for the ADC at a given read-out rate and mode. Values for dark current and read noise are generally given in terms of electrons and so the units with which the simulated spectrum is described must match. Throughout this thesis this reference electron spectrum will be denoted x_{ref}.

4.4.2 Simulating Experimental Conditions

In this section, the basic process for simulating spectra is described. This process is valid for a single column of a standard read-out mode known as a Full Vertical Bin (FVB). An FVB read-out involves shifting all pixels in a column into the shift register and a single spectrum of intensity values is then read out from the CCD. There are other read-out modes that will affect this process but these will be discussed in greater detail in Chapter 6.

The initial step is to take a sample from x_{ref}, denoted $x_{ref}(i)$, which is then scaled by the target acquisition time, t, that is being simulated. $x_{ref}(i)t$ is taken to be the mean of a Poisson distribution and a random value from the distribution is taken to simulate the effect of experimental shot noise. This is denoted $x_{sim}(i)$.

Dark current is modelled in a similar fashion to shot noise in that the mean rate of dark current per pixel per second, c, given in the CCD specification sheet is scaled to the acquisition time. However, for a given $m \times n$ matrix of CCD pixels the m pixels in column i will contribute dark current and so ct must also be scaled by m. $ct \times m$ is used as the mean of the probability distribution and a random value is selected to simulate the effect of dark current noise, $DC_{sim}(i)$.

A single instance of read noise is associated with each pixel that is being read out from the CCD. As previously stated, read noise is modeled by a Normal distribution. The mean and standard deviation of this distribution is typically given in the CCD specification sheet. A random value of this distribution is then selected, $R_{sim}(i)$.

Sample i from the simulated spectrum, S_{sim}, is given by Equation 4.13

$$S_{sim}(i) = x_{sim}(i) + DC_{sim}(i) + R_{sim}(i) \tag{4.13}$$

Thus the total noise of a sample in a spectrum recorded in FVB mode for a given biological specimen and a given Raman system is modelled.

4.5 Summary

Noise in a Raman system may be modelled using the probability distributions described in Section 4.2. Details of the noise sources originating within the array detector can be obtained from the detector's specification sheet and, consequently, can be used to simulate the effect of experimental noise in a spectrum. However, irradiance will change depending on sample type and system configuration. Therefore, in order to simulate the performance of a particular Raman system for a particular sample, it is necessary to calculate a system specific irradiance from an experimental dataset on which to base the simulation of experimental shot noise. System specific irradiances can be used to simulate a variety of experimental conditions in order to provide information on the optimal parameters for acquiring spectra such as the minimum acquisition time for reliable classification (Chapter 5), optimal read-out parameters (Chapter 6) , and how a system may behave with an alternative optical element (Chapter 7).

Chapter 5

The Effect of SNR on Classification

The work in this chapter is related to the following publication:

Sinéad Barton, Laura Kerr, Katarina Domijan, and Bryan Hennelly . "On the effect of experimental noise on the classification of biological samples using Raman micro-spectroscopy." Biophotonics: Photonic Solutions for Better Health Care V. Vol. 9887. International Society for Optics and Photonics, 2016.

I would like to thank Dr. Shirley O'Dea from the Biosciences Department in Maynooth University for cultivating and depositing the cells used to record data. I would also like to extend my gratitude to Dr. Laura Kerr for recording the datasets from which the system specific irradiances of both cell lines were obtained.

5.1 Introduction

The goal of this thesis is to develop a set of methods to improve the SNR in a Raman spectrum. The methods for modelling the SNR in spectra were described in the previous chapter, as well as the method for creating artificial datasets. The specific goal of this chapter is to use this methodology to simulate spectra collected over different acquisition times. By doing this, it is possible to establish the specific SNR required for reliable classification using the numerical methods that have been previously described. Consequently, the minimum acquisition time required to collect spectra can also be established. This work provides a baseline reference for minimum acquisition time and SNR for reliable classification. In later chapters it will be shown how to simulate datasets for a given system, when different elements are swapped into the system. In this context, the method described in the current chapter could be used to interpret the effect of these different elements on classification and the minimum acquisition time of this 'new' system for reliable classification of a particular set of spectral datasets.

The successful classification of cell lines based on subtle spectral differences in recorded Raman spectra necessitates optimal acquisition parameters. [73] This chapter is an investigation into the relationship between SNR and the diagnostic potential of a Raman micro-spectroscopy system with PCA-LDA classification, specifically for the case of urine cytology. The sensitivity and specificity of Raman based multivariate classification of two cell line datasets is evaluated using PCA-LDA classification and Leave-One-Out cross validation, as described in Chapter 3.

Preliminary recordings of two cell lines, T24 and RT112 (high and low grade bladder cancer respectively, which are representative of the types of cells that are obtained from the clinical procedure of urine cytology for bladder cancer diagnosis [68]) are used to create a system specific irradiance that is representative of both cell lines. These system specific irradiances are then used to simulate experimental conditions as per the method described in Section 4.4. Any additions to this procedure will be briefly discussed in Section 5.2.1. These datasets are then input to the PCA-LDA leave one out cross validation procedure in order to evaluate classification sensitivity/specificity as described in the previous paragraph. Through this analysis it is illustrated that the SNR does have an impact on the reliability of classification and that it is possible to estimate the minimum exposure time required for reliable classification for a given experimental system.

This chapter is divided into two main sections. Section 5.2 describes the design of the simulation experiment, the generation of the data used in the simulation, and the statistical analysis applied to the spectra. In Section 5.3 the results are presented including PCA scatter plots for various simulated exposure times and a sensitivity/specificity graph are used to illustrate the impact of noise on classification accuracy.

5.2 Simulation Design

The main types of noise have been discussed previously in Chapter 4 and can be affected by a number of system components: source wavelength [12], source power, Quantum Efficiency (QE), microscope objective (varying numerical apertures), time, temperature, spectrograph grating, and camera type; the first four directly affect the spectral irradiance, which impacts upon the shot noise and consequently the SNR. Time can affect both shot noise and dark current. Temperature primarily affects the dark current production within the CCD. The type of camera used affects all types of noise discussed since the camera determines the QE relative to the spectral wavelengths, the average rate of dark current generation at given temperatures, and has specific amounts of read noise associated with the ADC. Since the standard deviation of read noise is relatively small when compared to the overall signal energy it is unlikely that it will have a significant effect on the classification process and so the main focus of this chapter is on shot noise and dark current. To simplify matters, acquisition time was the only parameter varied in our simulations as it is a parameter that is frequently varied in experiments as well as being a significant factor in the noise produced, particularly dark current and shot noise. Reducing the acquisition time per spectrum is often a motivation in Raman based classification studies, since long exposure times can lead to time consuming experimental conditions, particularly when a high volume of samples must be analysed.

5.2.1 Data Generation

A dataset of T24 and RT112 (high and low grade bladder cancer respectively) cultivated under the same conditions were recorded under ideal camera conditions and acquisition time. A full description of the experimental set up used, as well as the cell preparation, is provided in reference [68]. Double acquisitions of 50 cells were obtained for cosmic ray artefact removal. After spike contamination was eliminated in all spectra, they were averaged to create a smooth, relatively noise free, signal. Mean dark current and mean read noise was removed via subtracting a dark current background obtained under the same acquisition parameters. The remaining signal was converted back into electrons rather than ‘counts’ (a unit employed by Andor’s ADC) by multiplying the recorded signal values by the number of electrons the ADC registers as a count. This value can be obtained from the specifications sheet of the CCD and varies for each read-out rate. The signal is then divided by the acquisition time to estimate the spectral irradiance at 1s. Rubber-band background subtraction [40] was applied to the true irradiance signal using a 7^{th} order polynomial, this particular polynomial was chosen empirically, in order to reduce the presence of any background contribution; the resultant spectral irradiance signals are illustrated in Fig.5.1 as the base signals. These two irradiances were used as the basis of the datasets that were generated to simulate different experimental conditions. Each dataset was generated using the simulation model described in the Section 4.4, by modelling the respective distributions based on the irradiance signal, acquisition time, dark current parameters estimated from the camera specifications, and the read noise contribution. The system modelled is an ideal system (i.e. no system drift resulting in calibration issues or changes in irradiance), delivering 5mW to a $1\mu m$ spot on the sample, and the resulting scattering is transmitted to the Andor iDus BR-DD operating with optimum read parameters, i.e. cooled to $-80°C$ and a $33kHz$ read out rate, to minimise the dark current and read noise contributions from the CCD.

Fig.5.1 shows the improvement of spectral quality from a one second acquisition on the system de-

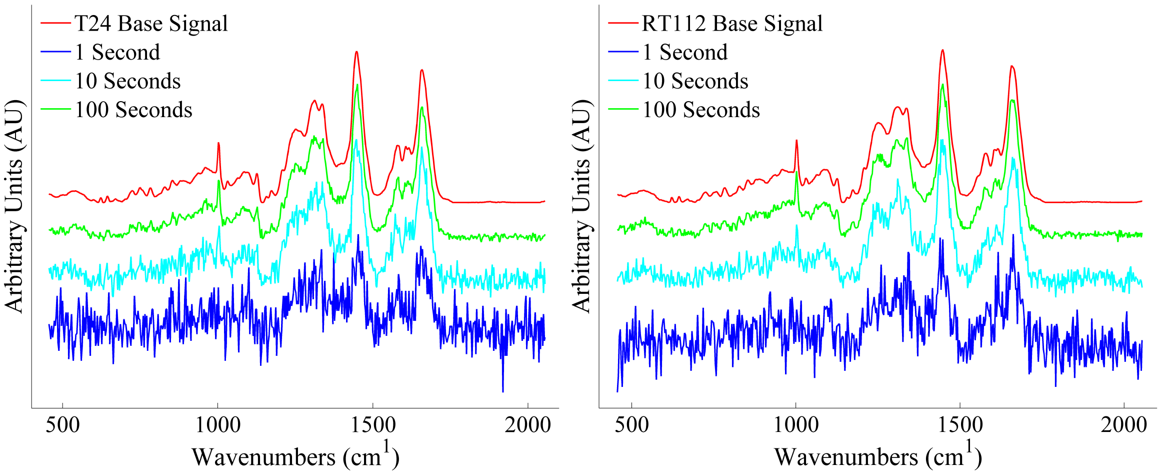


Figure 5.1: An illustration of simulated spectral quality over different integration times compared to the base signal used to generate them (spectra have been area normalised for illustrative purposes).

scribed to one hundred seconds for each of the cell lines chosen. Spectral features are obscured in low acquisition times and so it is expected that the sensitivity and specificity of that dataset will be low where-as over longer integration times peaks are sharper and increasingly well defined. It would be expected that ten seconds will give good clustering and reasonably reliable classification and one hundred seconds will provide 100% sensitivity and specificity.

5.2.2 Statistical Analysis

PCA and LDA are described in detail in Section 3.5. We recall that PCA is a multi-variate statistical technique that reduces the dimensionality of datasets by linearly transforming a set of variables into a set of eigenvalues and eigenfunctions, that are ordered in such a way that the first component accounts for the greatest variation across the dataset and the last component accounts for the least. [74] LDA is a supervised classifier that determines linear combinations of spectral features in order to characterise two or more datasets. PCA is a popular and powerful classification tool in its own right but can be used in conjunction with LDA to provide an automated ‘unsupervised’ classification tool. [75] A PCA-LDA model with Leave-One-Out cross validation was implemented to determine the classification sensitivity and specificity of the simulated datasets. All spectra were area normalised before being classified. Only the principal components that explained 95% of the variance were input to the LDA function.

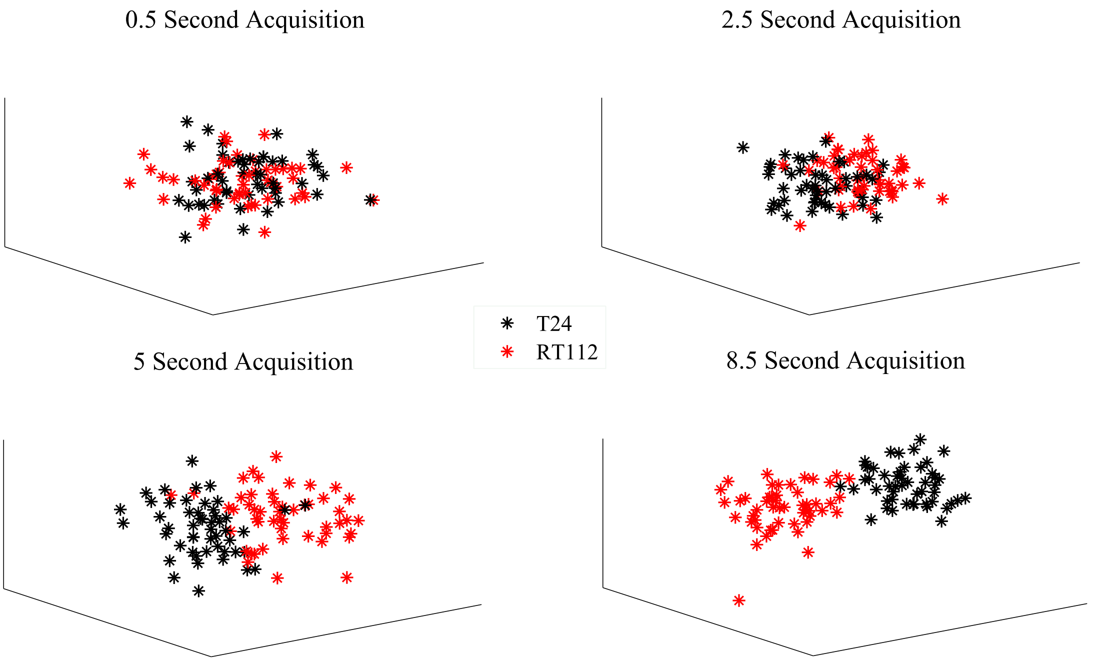


Figure 5.2: PCA scatter plots of the simulated data at different exposure times.

5.3 Results

Initially 20 datasets of 100 spectra (50 each of T24 and RT112) were created simulating the expected noisy spectra recorded from a FVB of the Andor iDus 420 BR-DD operating under ideal conditions, in one second increments from 1 to 20 seconds. It was noted that there was sufficient clustering in the scatter plots of the first three scores after 8.5 seconds to indicate good classification, as depicted in Fig.5.2, which shows the improved clustering over a smaller range of times.

These results indicate that after approximately 10 seconds reliable classification is possible. Theoretically, this can be linked with the noise signal interfering with spectral features that the PCA-LDA model use for classification. To test this hypothesis three datasets of 1 second, 5 seconds, and 10 seconds were examined to see if the maximum difference between the base signals would be obscured by noise. The noise signals were estimated by subtracting the mean of the respective datasets from the first spectrum in the dataset.

The maximum difference between the base spectra is approximately 7 electrons, which scales with time, therefore for 5 and 10 seconds the maximum difference is estimated at 35 and 70 electrons respectively. The standard deviations of the noise signals for 1s, 5s, and 10s were calculated as 9.5, 17.91, and 26.2 respectively. For a 1 second acquisition one standard deviation of noise will routinely be greater than the maximum difference in the signals meaning that the classification cannot reliably identify spectral patterns for classification. 5 second acquisitions require two standard deviations to mask the maximum difference

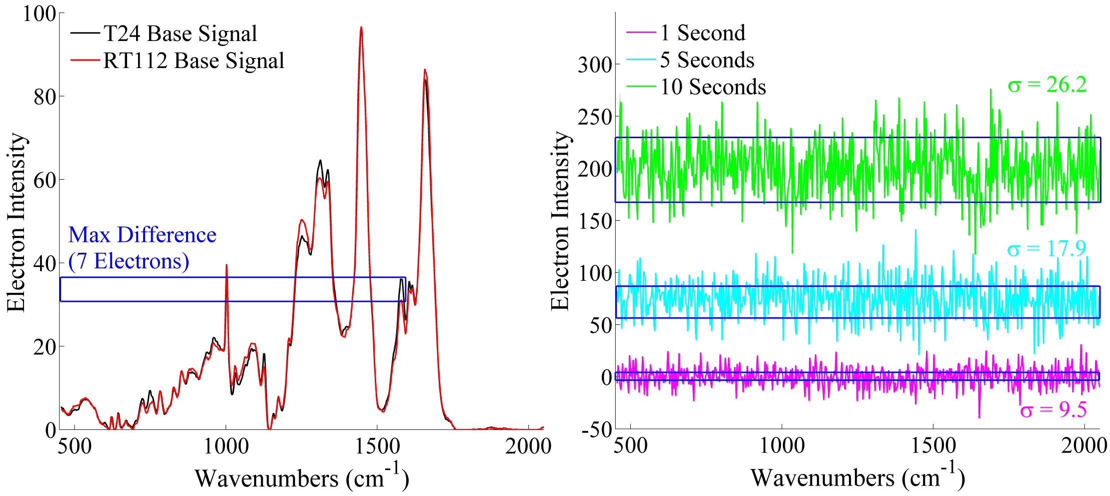


Figure 5.3: An illustration of the spectral features from the base signals used to create the datasets compared to the estimated noise signals for three different acquisition times. The noise signals have been offset for illustrative purposes but no normalisation or scaling has been applied.

in the signals, indicating that the likelihood of obscuring the most significant spectral feature reduces, this level increases to three standard deviations at 10 seconds further reducing the possibility.

Following on from this result a smaller range of acquisition time was examined; 26 datasets of 100 spectra (again 50 of each cell type) were simulated over a range of 0.5s to 13s in 0.5 second intervals. These datasets were classified using the PCA-LDA model described in Section 5.2.2 and the sensitivities and specificities of T24 with respect to RT112 were calculated. The results of this are graphed in Fig. 5.4, in which it can be clearly seen that the sensitivity and specificity increase non-linearly with respect to acquisition time before equaling 100% after 10 seconds; therefore, for this system we predict that a 10 second exposure is the minimum time required to produce reliable classification results despite significant noise levels still being evident in the spectrum, as shown in Fig.5.1. This analysis can be extended to other systems with different equipment, so long as the camera specifications are available.

5.4 Summary

The simulation model that is presented in this chapter demonstrates that it is possible to reliably model the noise present in a Raman spectrum collected from a biological cell using a standard Raman micro-spectrometer optical set-up. This model takes into account *a priori* knowledge of the source laser wavelength and power incident on the cell, as well as the camera specification in terms of dark current and read noise. The true irradiance of the Raman spectrum obtained from a biological cell can be estimated from spectra collected under ideal conditions over very long exposure times. The simulation has shown that

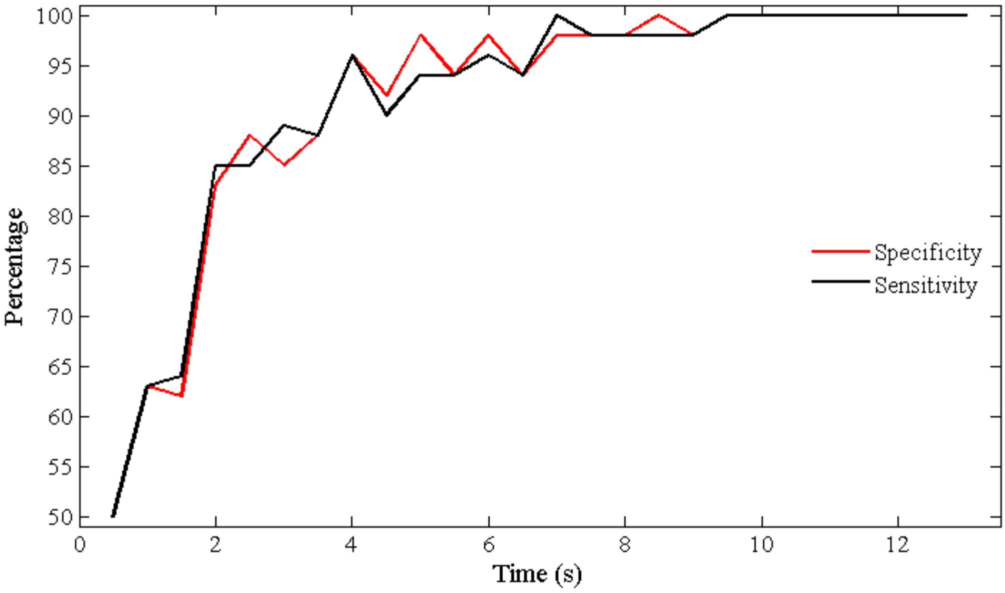


Figure 5.4: A graph of sensitivity and specificity vs time.

PCA-LDA classification is reasonably robust and resistant to noise at low power and integration times. This observation indicates that it may be possible to obtain good classification results using less expensive (noisier) CCDs or under non-ideal conditions by optimising the read parameters of the CCD to boost the SNR of the recorded signals. Optimising the read parameters of the detector can significantly increase the SNR of the recorded spectra. Parameters such as temperature, read-out rate, and read mode affect the noise introduced into the spectrum by the CCD. Multiple acquisitions, e.g. for the purposes of cosmic ray removal, also affect the noise introduced into the spectra. Both of these aspects are investigated in the following chapter to quantify how these parameters affect the SNR of the recorded spectra.

Chapter 6

Optimal CCD Read Modes for Maximising Spectral SNR

The work in this chapter is related to the following publication:

Sinéad Barton and Bryan Hennelly. "Signal to noise ratio of Raman spectra of biological samples." Biophotonics: Photonic Solutions for Better Health Care VI. Vol. 10685. International Society for Optics and Photonics, 2018.

6.1 Introduction

The principle goal of this chapter is to improve the SNR of the recorded Raman spectrum recorded by a given system by optimising the read-out parameters of the CCDs. An investigation is performed into the effect of different read out methods (full vertical bin, cropped bin, or full image readout) of the CCD on the SNR of a Raman spectrum. The methodology used for creating experimental datasets that was defined in Section 6.3.1 is expanded to account for the different levels of noise resulting from different read-modes that are used to record the spectra. This work allows the user to determine the optimal read parameters required for a given application in order to maximise SNR in the recorded spectra. Based on the analysis provided in this chapter, methodology for reliably simulating the effect of read-out parameters is established and is used in the following chapter to predict the effect of changing optical elements in a given Raman system. Also presented in this chapter is evidence that the SNR of a spectrum acquired using a single acquisition is higher than spectra obtained from multiple acquisitions averaged together. This conclusion provides the motivation for Chapter 8.

The principle source of noise in a Raman spectroscopy system is the noise from the array detector, which is typically a CCD. In this chapter, practical read out methods in high dark current situations for the enhancement of the SNR are investigated. As previously discussed, high dark current conditions result from either insufficient cooling of the camera/array detector or from long exposure times, or they may simply be inherent to the camera. The experimental and simulated results show that a spectrum collected on a confocal Raman system in a high dark current environment can differ significantly depending on the capture mode; in this case, reading out the entire image and extracting the spectrum from a row of pixels can result in significantly improved SNR compared to the spectrum recorded from full vertical binning. Crop mode provides the best results, by limiting both the dark current and read noise contributions.

The effect of multiple acquisitions, i.e. for X acquisitions of time t/X averaged together as opposed to a single acquisition of time t, on the SNR is experimentally quantified. This scenario might arise through averaging together for cosmic ray removal or through attempting to overcome the heterogeneity of cells. It is demonstrated in this chapter that multiple acquisitions negatively impact SNR. While there is much attention in the literature to model the sources of noise, this chapter aims to be one of the first to provide a comprehensive overview of all the main noise contributors, in terms of practical recording modalities and their effect on SNR, and to analyse the optimal camera configuration to record weakly scattering samples, such as biological samples.

This chapter is divided into three main sections. Section 6.2 provides an overview of the three principle recording modes and theorises the SNR for a single camera pixel, as well as for the case of a row of pixels that are binned together. Section 6.3 mathematically models SNR in the context of these different recording modes with a particular focus on weakly scattering samples such as a biological cell. Lastly, Section 6.4 provides the experimental results that validate the mathematical model developed in the previous section.

6.2 CCD Read Modes and Parameters

Read modes are methods by which the spectrum is extracted from the CCD and can be specific to individual spectrometer architectures and manufacturers e.g. the Renishaw systems that perform line scanning are designed for an array of spectra to be read as an image; this type of architecture is discussed in detail in Ref [76]. The following sections have been written in the most general sense possible with specific reference to a confocal microscopy set-up such as that produced by Horiba and discussed in detail in references. [12,68]

6.2.1 Full Vertical Binning and Crop Mode

Reading a Full Vertical Bin (FVB) of the CCD involves pooling each column of pixels into their respective shift register and reading out only a single value for each column. In a confocal microscopy [12, 68], only a relatively small subsection of the camera will be illuminated by Raman scattered photons. This indicates that in FVB mode the majority of the pixels are solely contributing DC noise to the spectrum although there is only a single instance of read noise associated with each column. The main advantages of this mode are the minimal read noise and fast cycle time. This is due to the method performing a single read-out from the shift register, as opposed to reading out individual pixels. Crop mode is similar to FVB, except that the spectrum is extracted from a particular subset of pixels on the CCD face and can also be referred to as a Region of Interest (ROI) extraction. While crop mode is generally put forward as a method for increasing frame rates it provides the added benefit of eliminating the dark current contribution from any pixel outside the ROI. Theoretically, crop mode will provide the highest SNR yield as it combines the advantages of both FVB mode and image read out.

6.2.2 Image Mode

As opposed to FVB mode, each individual row of the CCD is moved into the shift register and read out into a matrix of values representing the electron count registered in each individual pixel and a single instance of read noise is added to each of these. The advantage of this is that a large number of pixels, which only contribute DC noise, may be excluded from the final spectrum by extracting only the illuminated row(s) of pixels. This will significantly reduce the impact of DC noise, in the same way as is achieved by crop mode described above, with the caveat that multiple instances of read noise will be introduced. The other major disadvantage of this read method is the time involved since this takes a significantly longer time to extract from the CCD than binning.

6.2.3 Read Rates

Separate to the discussion on read out modes discussed in the previous subsections, is the subject of read out rates. [60] Read out rates relate to the speed at which the FVB or image is read out and directly affect the CTE of the CCD. Higher read out rates equate to a lower CTE i.e. rapid shifting of electrons from one pixel to another is less stable and so less efficient. Read out rates will, therefore, obviously affect the overall read noise contribution and are defined in the CCD specification sheet for given camera types.

6.3 Modelling SNR for Different Read Modes

6.3.1 Mathematical Model of How Read Modes Affect SNR in a Recorded Spectrum

As previously discussed, the SNR for a single pixel is given by Equation 4.12. This definition of SNR will vary slightly depending on the read methods selected i.e. FVB or image. Equation 4.12 is valid if the entire image has been read out, and therefore, each pixel has an instance of read noise. As discussed earlier, the spectrum will be recorded by one of three possible modes: (i) full image read out followed by extracting and summing a small number of rows to produce one spectrum; (ii) full vertical binning; and (iii) cropping, which is equivalent to full vertical binning over a reduced number of rows. For all three cases, a single general equation, Equation 4.8, can be extended to model the SNR for a single spectral component as follows:

$$SNR = \frac{iq(\lambda)t}{\sqrt{[iq(\lambda)+cp_{dc}]t+p_r n_r}} \tag{6.1}$$

where p_{dc} is the number of pixels contributing dark current noise and p_r is the number of pixels contributing read noise to the spectral component. In FVB mode, p_{dc} will be the same as the number of pixels in the full column and p_r will be one as there is only one instance of read noise. For an image read out, both parameters are dependent on the number of rows that are extracted and added together to create the spectrum. In this case, selecting the optimal number of rows to extract presents a challenge for maximising the SNR as there is no trivial solution for determining whether inclusion of an additional pixel will increase or decrease the SNR. However, it is possible to reliably identify areas of the CCD that have been illuminated by Raman scattered photons by performing a grid search and isolating rows where the majority of the pixels register as higher than the mean dark current contribution plus some multiple of the standard deviation of the dark current contribution; in the experiments below we choose this multiple to be six. It should be noted that the last row to be read in image mode will have higher levels of dark current than the first row due to the time lapse in moving the rows into the shift register.

6.3.2 Model for Emulating Experimental Conditions in a Recorded Spectrum

As already discussed in Chapter 4, shot noise cannot be estimated from camera parameters and therefore, it must be estimated from the spectrum itself. In order to accurately simulate the spectral intensity arriving

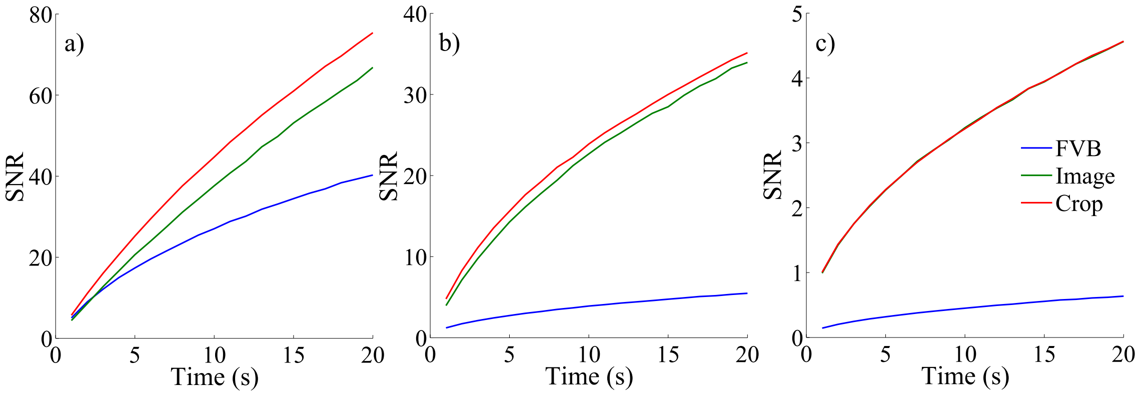


Figure 6.1: Illustration of the effect of increasing CCD temperature, and thus dark current production, on SNR over time for FVB, crop, and image read mode on an iDus 420 BR-DD. a) -80°C, b) -50°C, c) -20°C. The irradiance used in this simulation is equivalent to the Raman irradiance expected from an epithelial cell , based on thousands of such recordings using a commercial Raman microspectrometer.

at a detector over a given acquisition time. Knowledge of the irradiance then permits a spectral intensity to be simulated, which includes a shot noise term, for any acquisition time. This reference spectrum was provided by recording a high SNR spectrum of a polymer reference material, that was later used for a series of experiments, to statistically analyse SNR experimentally for comparison with the simulations. The mean dark current noise and mean read noise baseline is subtracted from the reference spectrum and counts are converted to electrons as discussed earlier. Poisson noise is then generated for each spectral component in order to create an accurate representation of how shot noise would behave under each condition. It should be noted that the reference irradiance was scaled to be equivalent to the irradiance expected from a biological cell recorded using a standard confocal Raman microspectrometer, based on thousands of measurements. [68]

6.3.3 Simulating the Effect of Dark Current Noise

Dark current and read noise contributions were simulated based on the camera parameters and their respective probability distributions. The results of the simulations presented here for the three read modes,(FVB, crop mode, and image mode) are based on average results from 100 simulated noisy spectra in all cases. Fig. 6.1 illustrates the effect of increasing dark current noise on the SNR of the spectrum for each of the three read modes.

In Fig. 6.1 it is evident that dark current production severely degrades the SNR of a spectrum, particularly for the case of FVB. In the case of image read out, the dark current contribution is reduced to that produced by only a small number of rows of pixels that the user can select. However, image read out is hampered by the increased read noise associated with every pixel, which may be relatively large depending

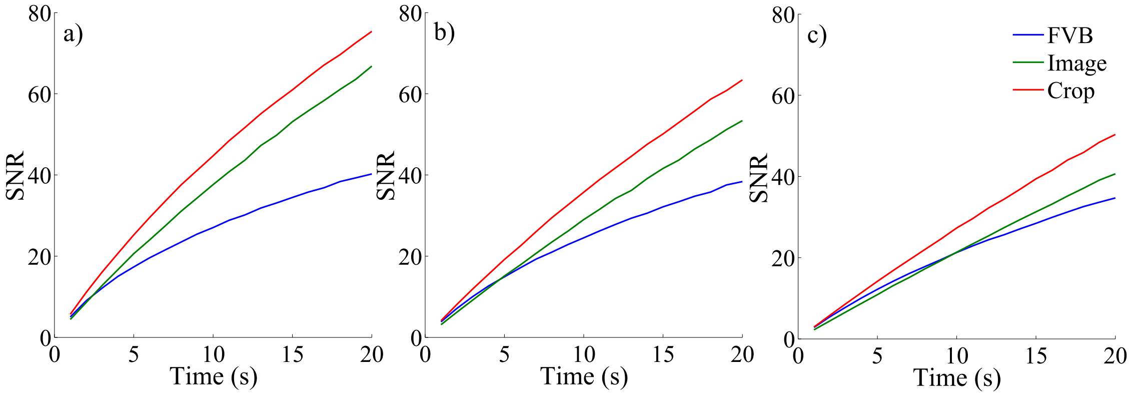


Figure 6.2: A diagram illustrating the effect on SNR of averaging multiple acquisitions together i.e. an average of $t/acqs$ compared to a single acquisition of time t. a) single acquisition, b) double acquisition, c) quadruple acquisition

on read out rates, and the number of rows that are extracted. Taking both read noise and dark current into account, crop mode provides the best results since there are a limited number of rows included in the read out, eliminating the majority of the dark current, yet only a single instance of read noise is contributed to each spectral component. In extremely high dark current conditions there is little difference between crop mode and image mode due to the relatively large dark current contribution dominating over the read noise contribution. However, considering that a cropped spectrum can be extracted from the CCD at a significantly faster rate than an entire image can be read out, crop mode is preferable.

6.3.4 Simulating the Effect of Multiple Acquisitions

Some applications may require multiple acquisitions of spectra to be averaged together, such as in the case of using the double acquisition method for cosmic ray removal. [10] The resulting spectra from this process has multiple instances of read noise associated with the averaged spectrum which will decrease the SNR of the spectrum when compared with a single acquisition of equivalent acquisition time, as illustrated in Fig. 6.2. In this simulation the negative impact of averaging spectra together is quantified. A comparison is presented of the SNR of a single acquisition of time, t, compared to a spectrum comprised of a number, $acqs = 1, 2, 4$, of spectra obtained over the time, $t/acqs$, and averaged together.

The simulations indicate that crop mode is again expected to yield the highest SNR except in very low dark current conditions where crop mode and FVB are approximately equivalent. A comparison of Fig. 6.2(a) with the other two figures, reveals that as expected the image read out mode is affected the most as the multiple instances of read noise associated with the read mode are compounded by the averaging process. It should be noted that the negative effect of multiple acquisitions on overall SNR, becomes significantly less apparent as the signal irradiance or dark current increases, or a long acquisition time is used in order to increase the overall signal intensity. This decrease of SNR due to multiple acquisitions need only be considered for spectra containing low signal intensities and low dark current contributions, and will be negligible for strong Raman scatterers; however, for weakly scattering samples, such as biological specimen, the effect of increased read noise may be an important consideration.

In Sections 6.4 a set of experimental results are presented that validate the simulations presented in this section. Although the simulated experiments include crop mode, the camera used to collect the datasets was unable to support cropping and so experimental results are not presented for that specific case. However, the experimental results presented below correlate closely with the trends that can be observed in the simulations and, therefore, it is reasonable to conclude that the simulated results provide a reasonably accurate representation of the performance of crop read out mode in terms of SNR.

6.4 Experimental Results

6.4.1 Experiment 1: FVB vs. Image, Single Acquisition

The aim of this experiment is to examine the SNR of spectra under various dark current conditions, recreating the effects of long acquisition times or applications where efficient cooling of CCDs may not be feasible. This was achieved by applying different levels of cooling to the CCD, which produce different base levels of dark current. A low numerical aperture microscope objective (Olympus UMplanFl 4x/0.1) and the system was defocused in order to produce spectra from the polymer reference material that had irradiance equivalent to that expected from an epithelial cell recorded under optimal conditions using a commercial Raman microspectrometer. An acquisition time of 2 seconds was used, resulting in a noticeable accumulation of dark current even at maximum cooling of the Andor iDus 420 BR-DD. The low magnification of the MO produces a beam with a wide depth of field, which prevents any major defocusing over the course of the experiment. The polymer reference material was acquired from Ibidi GmbH and chosen as the experimental sample due to it's thermal stability, resistance to photo-bleaching, and strong reliable signal that is not expected to change over time thus reducing the overall experimental variability. Ideally a biological sample would have been used in all of the experiments, but photo-bleaching and burning of such a sample would invalidate the results. It was, therefore, decided to use the polymer sample and reduce the recorded irradiance to match that of a biological sample. Low read out rates were applied to minimise the effect of read noise on the data.

A range of cooling temperatures were applied to the CCD, -80°C to -20°C in increments of 10°C, 100

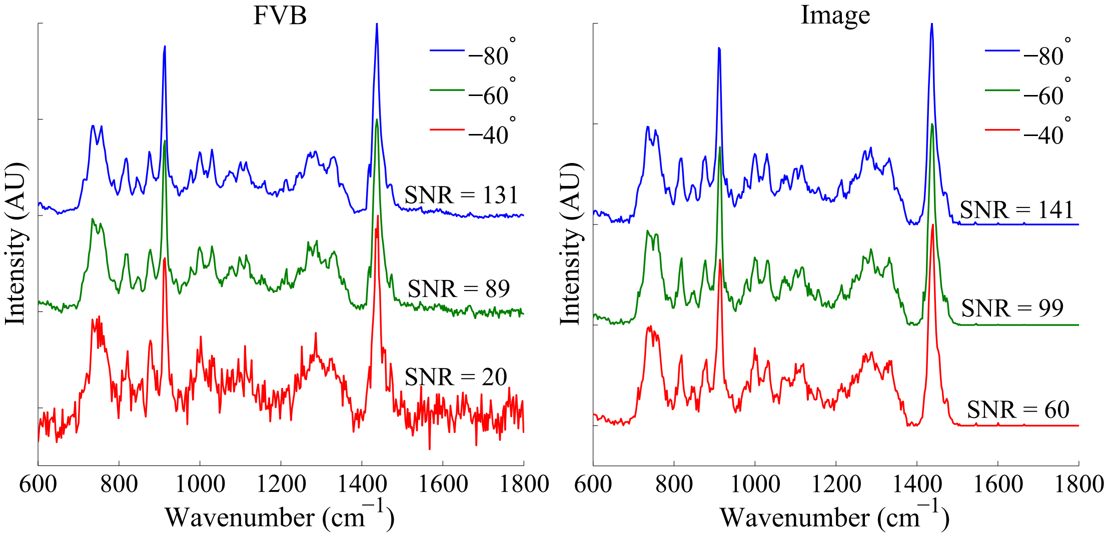


Figure 6.3: A side by side comparison of FVB vs image read modes under different dark current conditions.

spectra were recorded in both FVB and image mode. Fig. 6.3 illustrates the SNR that can be expected for both read modes under three distinct temperatures.

The spectra recorded using image read out mode are less affected by the dark current contribution than the FVB spectra. In terms of quantitative analysis, the mean SNR across the full dataset of 100 spectra for each read mode, as a function of temperature is displayed in Fig. 6.4. These figures are in agreement with the trends predicted by the simulations. The image extraction outperforms FVB mode to the extent that even at very low levels of cooling, where FVB mode is not feasible to use, i.e. -30°C and -20°C, due to saturation of the shift register through dark current alone, recognisable spectra may still be extracted using image read out mode. The usefulness of these spectra for the purpose of classification is not within the scope of this chapter but it can be expected that image read out would be more resilient to higher levels of dark current for this application.

It is clear from Fig. 6.4 that FVB mode outperforms image read out mode at -80°C and -70°C, which is also consistent with the results of the simulation; even at low read out rates, the additional read noise introduced by the multiple pixel read outweighs the dark current production. In applications where a longer exposure time is required, image mode will most likely yield spectra of higher SNR.

6.4.2 Experiment 2: Multiple Acquisitions - are two better than one?

One of the most popular methods for removing cosmic ray artefacts from spectra is the double acquisition method [10] whereby two spectra of equivalent acquisition time are recorded consecutively and then compared and/or averaged together to remove the cosmic ray artefacts. In this instance, due to the the elevated

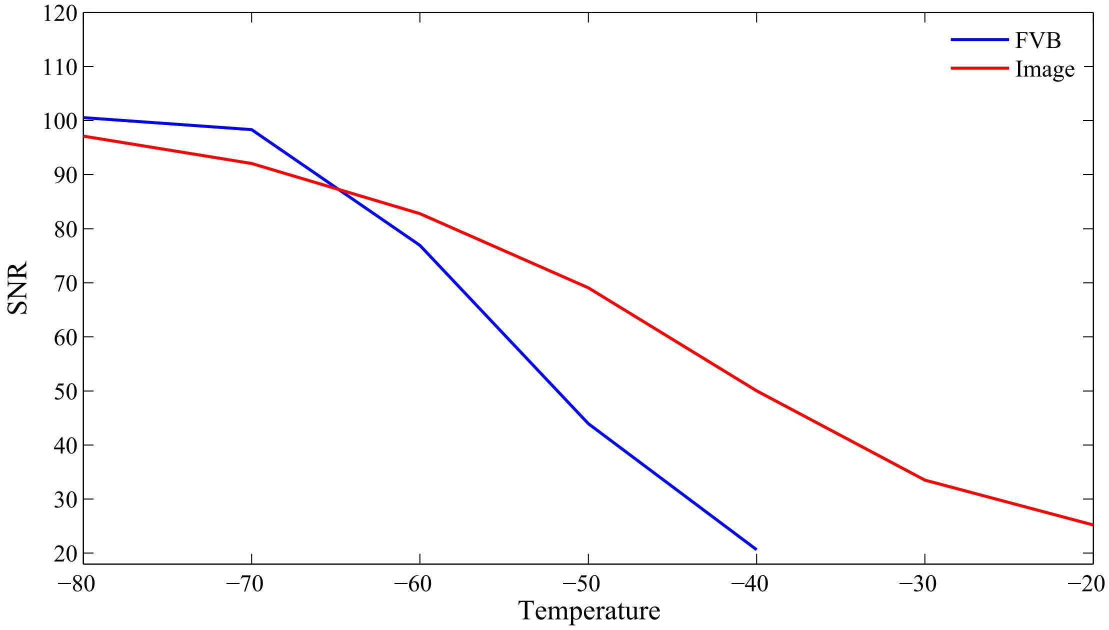


Figure 6.4: Quantitative analysis of the SNR trends of experimental data.

levels of read noise in the extraction of a spectrum from an image recorded using image read out mode, it may be possible that the addition of multiple instances of read noise will negatively impact on the SNR of the averaged spectrum.

Maximum cooling, -80°C, was applied to the CCD to reduce the presence of dark current in the experiments; this was further reduced by short acquisition times. Acquisition times were set to 2, 1 and 0.5 seconds for single, double, and quadruple acquisitions respectively. A read-out rate of 100 kHz was used in all cases, which provides the highest levels of read noise and would be most commonly used in order to provide the quickest results in terms of retrieving spectral data from the CCD. All other parameters and system configurations remain as described in the previous experiment.

Spectra were processed as before with the additional step of averaging together two spectra for double acquisition and four spectra for quadruple acquisition to produce a dataset of 100 spectra for each experimental condition examined. Fig. 6.5 illustrates the impact of the additional read noise on the spectra for both FVB and image read out by showing the statistical distribution of the SNR across the full dataset using a box and whisker plot (boxplot).

A box plot is based on a Gaussian distribution and is divided into 3 distinct parts, the red line, blue rectangle, and black 'whiskers'. The red line shows the median of the dataset. The blue box encompasses 50% of the data, covering approximately 0.6745 σ of the Gaussian distribution, where σ denotes the standard deviation of the SNR data. The top 'whisker' section represents the upper 25% of data, equivalent to a range of +2.0235 σ. The lower 'whisker' represents the equivalent for the lower SNR values. Any data falling

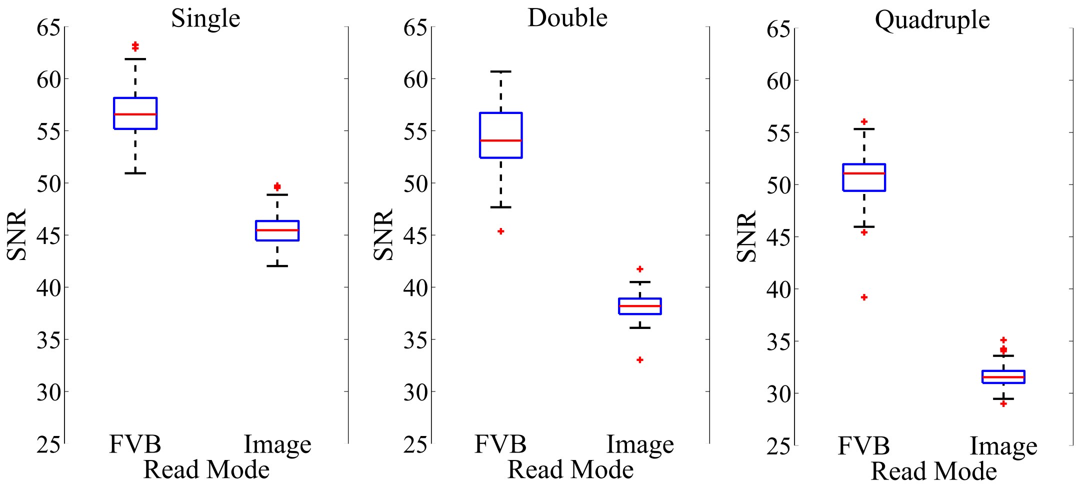


Figure 6.5: Boxplot comparison of multiple acquisition spectra for both read modes.

outside these σ boundaries is considered an outlier and is signified by a red star.

In keeping with the results from the previous experiment, FVB mode outperforms image read out mode under all of the the temperatures and acquisition settings of this experiment; the disparity between the FVB and image read modalities is increased as the number of acquisitions increases. It can be seen that the averaging process affects the SNR of both read modes negatively though the inclusion of multiple instances of read noise; however, the impact on the FVB spectra is lower. It can be surmised that, unless the benefit of using image read out mode in terms of reduced dark current outweighs the increase in read noise for multi-acquisition, it is more beneficial to use FVB mode for spectra that require multiple acquisitions to be averaged together.

6.5 Summary

The primary conclusion from both the simulations and experimental results is that crop mode, under all circumstances, will yield spectra with the highest SNR as opposed to the alternative read modes due to the exclusion of the majority of dark pixels and it's minimal contribution of read noise. Crop mode has the added benefit of an increased frame rate, which is useful in high throughput applications. While experimental results have not been provided for crop mode due to the CCD used in the experiment not being unable to support it, the similarity between the experimental results and the simulations concerning the other read modes indicate that it is reasonable to conclude that the simulation results for crop mode are valid. In situations where crop mode is unavailable and there are high levels of dark current, reading out an image will produce a spectrum with a higher SNR than one recorded using FVB mode. While the cycle time required to read out an image can be significantly longer than that for a FVB, the image can still produce usable spectra under high dark current conditions where a FVB may saturate the shift register, as demonstrated in Section 6.4. The second conclusion to be noted is that in all cases a single acquisition will produce a higher SNR than multiple acquisitions averaged together. In applications requiring multiple acquisitions of weak signals, it is recommended to use image read out mode, the lowest read out rates, and apply gain settings that produce the lowest levels of read noise in order to mitigate the negative impact on the SNR of the resulting spectra. However despite applying these precautions, this conclusion would present an issue when collecting spectra with weak irradiances and applying cosmic ray removal to the data using the double acquisition method. [10] Therefore an alternative algorithm for the removal of cosmic rays is proposed in Chapter 8.

In the following chapter the noise model presented in this chapter will be further expanded to include the QE of the CCD in order to investigate how replacing this the CCD with a CCD with different noise characteristics will affect the SNR of the system.

Chapter 7

Predicting the effect of changing optical elements in a given system

The work in this chapter is related to the following manuscript in preparation for submission to a journal:

Sinéad Barton and Bryan Hennelly. "Predicting the effect of changing an optical element in a given Raman micro-spectrometer". In Preparation. Analytical Methods.

7.1 Introduction

This chapter focuses on expanding the theory that was the developed in the previous chapter for the purposes of making Raman systems more efficient or, in some cases, more cost effective. In this chapter, the methodology for creating datasets of FVB spectra is expanded to account for the quantum efficiency of the CCD present in a given system. This allows the user to evaluate the impact of changing this optical element will have on the SNR of the recorded spectra. Optical elements for Raman systems can be very expensive; however, there is no rigorous method to compare the effect of a replacement element in a given system configuration, without purchasing and testing the new element. In the following sections, the principles of modelling a given system that makes use of an alternative optical element is developed and this is used to generate simulated datasets that might be expected from the newly adapted system. The methodology presented here may be used to evaluate the effect of a range of different optical elements for a given purpose, enabling the user to perform a cost benefit analysis of investing in new equipment without having to rely on qualitative estimates.

With the exception of changing the camera, changing optical elements in the spectrometer primarily affects the irradiance on the detector and, therefore, affects only shot noise. Source power, source wavelength, MO, confocal aperture, and the diffraction grating all affect the shot noise in some way. However, the CCD affects all of the primary noise sources described in Section 4.2 (including the shot noise since the quantum efficiencies of the two cameras in question may differ) and is arguably the most expensive. This chapter explores how to predict the effect of replacing a CCD in particular in a given Raman system for a given sample or set of samples. We also briefly discuss how this approach may be extended to replacing other components including the microscope objective.

7.2 Predicting the Performance of a Changed Optical Element

When predicting the effect of changing an optical element, it is important to first establish an accurate system specific irradiance for a given system for a given sample. In order for the resulting system specific irradiance to be used for predictive purposes it is also necessary that the Raman system be regularly maintained and calibrated such that this irradiance is not expected to change over time. System drift can result in subtle differences in collected irradiance including intensity variations and baseline contributions that may affect the SNR predicted by the methods described in this chapter.

The methodology in the following sections is based on creating the system specific irradiance as described in Section 4.4.1 and the model used to create simulated datasets of full vertical bin (FVB) spectra as described in Section 6.2.1. The FVB read mode is used throughout the remainder of this chapter; however, we note that any read mode could be used in generating the datasets.

7.2.1 CCD

In addition to considering the different read noise and dark current values, the primary difference between predicting the effect of a replacement CCD and modelling a given CCD, as described in Chapter 6, is the Quantum Efficiency (QE) of the CCDs. CCDs can be optimised for different wavelength regions and this will affect the overall percentage of photons that the CCD will successfully convert to electrons. This can have a serious impact on the shot noise detected by the system. The specifics of the theory behind modelling the QE are described in Section 4.2.2 in terms of the probabilistic QE of a single sample at a particular wavelength. However, in an experimental context the QE will fluctuate across wavenumber range of interest and, therefore, it is necessary to model the QE as a polynomial across the wavelength range that the CCD can detect, as illustrated in Fig. 7.1, if we are to fully account for the effect of QE in the noise model

7.2.2 Other Elements

All optical elements discussed in this subsection affect the collected irradiance only. Two important components of a confocal Raman microspectroscopy system are the confocal aperture and the MO. The interaction between the confocal aperture and the microscope objective affects of collection efficiency of a Raman system and thus the shot noise present in the system. [77] A confocal microscope achieves point illumination of the sample and rejects out of focus light. Thus the collected light is dependent on the interplay between confocal aperture and MO. [78] The NA of the MO and the size of the pinhole aperture define the optical sectioning properties of the confocal microscope. [79] This relationship can be mathematically defined in terms of the Full-Width Half-Maximum (FWHM). The FWHM is the width between the axial points where the intensity of an image defocuses to half of it's peak value in the image plane. Equation 7.1 represents this in mathematical terms.

$$FWHM = 0.67\frac{\lambda}{(n-\sqrt{n^2-NA^2})} \times \sqrt{1+AU^2} \tag{7.1}$$

Where λ is the source laser wavelength, n is the refractive index of the immersion medium, and AU is the size of the pinhole in Airy units. Airy units are defined as follows:

$$AU = \frac{(D \times NA)}{1.22\lambda \times M} \tag{7.2}$$

Where D is the aperture size in metres and M is the MO's magnification. The quantity of Raman scattered photons from a point source is assumed to be isotropic. Consequently, the collected irradiance is proportional to the square of the NA: $I_r \propto NA^2$ as well as the transmittance of the MO since the MO delivers the source laser and collects the Raman scatter this can be represented mathematically as: $I_r \propto T_{MO}^2$, where T_{MO} is the transmittance. The attenuation of the laser, due to absorption and scattering, as it propagates through the FWHM must be accounted for. In simple terms, a sample containing a high concentration of molecules will absorb and diffuse laser light as it propagates. This can be modelled by: $T = exp(-\alpha L)$, where T represents the sample transmittance of thickness, L, with attenuation coefficient, α. The attenuation coefficient is given by the sum of the absorption and the reduced scattering coefficient. [80] Taking into account this attenuation, the collected irradiance can be mathematically modelled as follows [81]:

$$I_r = (T_{MO}NA)^2 \int_0^{FWHM} exp(-2\alpha L)dL \tag{7.3}$$

Where the factor 2 appears in the exponential function in order to account for the loss of the back-scattered photons from the sample, as well as the laser.

Equations for modelling source laser wavelength, source laser power. and diffraction grating are given in Section 4.2.6. These elements primarily provide a scaling factor that acts on the the intensity. In the case of source power, the collected intensity should scale linearly with the an increase or decrease in source power. Equation 4.10 shows that the average number of photons emitted by source lasers of equivalent power is dependent on their wavelength i.e. the collected intensity is related to the fourth power of the frequency. Similarly, a change in source laser should scale linearly relative to the change in emitted photons. Diffraction gratings will provide a similar scaling factor based on the grating efficiency. In the event that the grating efficiency fluctuates across the grating then it is more appropriate to model it in a similar to that of the QE of the CCD.

Although modelling the scaling factor for intensity that results from a change in MO and/or confocal aperture is non-trivial, it remains a factor that will solely scale the collected intensity. Similarly, changes in source laser specifications and diffraction gratings will result in a change in collected intensity. However, the CCD affects all sources of noise described in Section 4.2. Therefore, it is advantageous to model this optical element to prove that system specific irradiances can be scaled appropriately and used to simulate experimental conditions.

7.3 Experimental Validation

7.3.1 CCD Specifications

The experimental validation presented in this section is designed to verify how the system specific irradiance created using data recorded from one CCD may be used to simulate the system specific irradiance recorded from an alternative CCD. Two cameras were used to record data for this experiment, an Andor iDus 420 BR-DD and an Andor Newton 920 BVF. A comparison of their specifications can be found in Table 7.1. Where the cooling represents the maximum cooling that may be applied to both cameras and DC represents the mean rate of dark current generation in electrons per pixel per second at the applied temperature. The read rate is given in MHz and the read noise is the standard deviation of the read noise per sample in

	Newton	iDus
Cooling (°C)	-80	-80
DC (e/p/s)	0.002	0.2
Read Rate (MHz)	3	1
Read Noise (e)	28.5	12.4
ADC (e/count)	9	12.6

Table 7.1: A comparison of experimental specifications for the Newton and iDus CCDs.

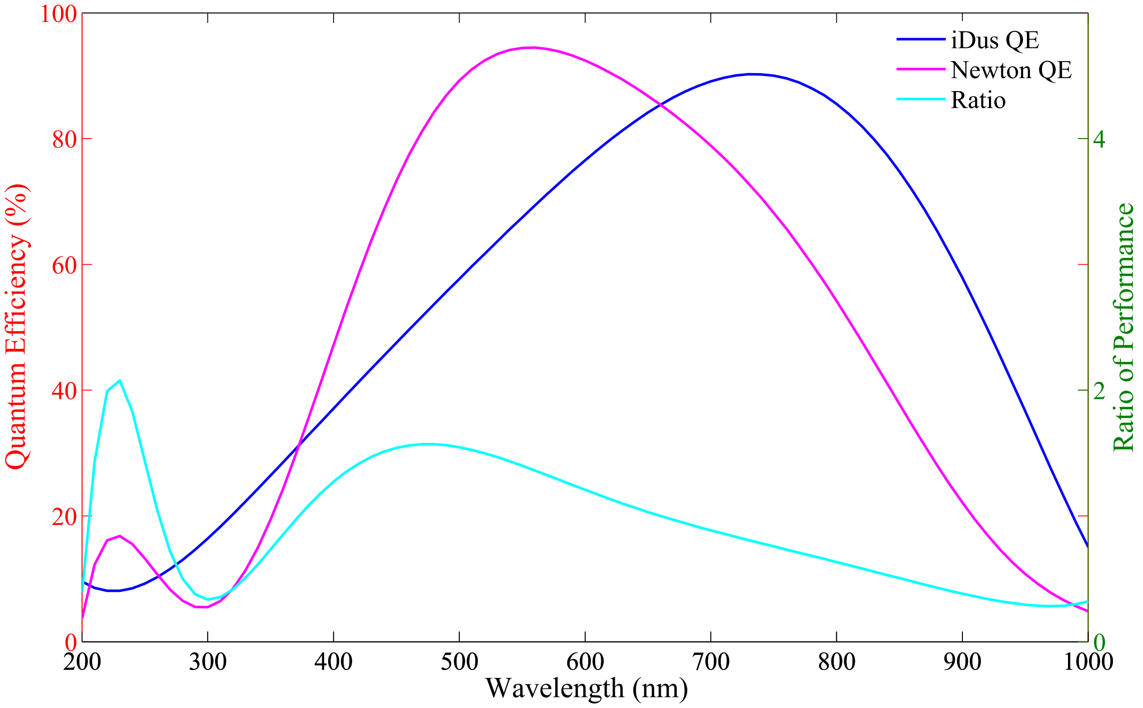


Figure 7.1: Illustration of the QE of a Newton 920 BVF, and iDus 420 BR-DD on the left hand axis and the ratio between the two QEs across the wavelength range on the right hand axis.

electrons. ADC represents the sensitivity of the ADC at the given read rate. When a CCD digitises the collected electrons it will convert them into a unit commonly referred to as counts, hence the units for ADC sensitivity being electrons per count.

The only specification not shown in Table 7.1 is the QE. This specification is illustrated in Fig. 7.1. The iDus CCD is optimised for use in the NIR wavelength as opposed to the Newton, which is optimised for the visible. Therefore, there is a significant difference in QE for the wavenumber region of interest.

In the above figure, the QE was derived from the online specifications sheets for both CCDs. This was done by measuring the percentage of converted photons relevant to the wavelength axis in increments of 25 Hz. These points were then used to fit a polynomial across the wavelength range.

Converting the system specific irradiance using this polynomial fit involves isolating the wavelength region across which the Raman spectrum was recorded using Equation 7.4 [82] to convert the wavenumber axis in cm^{-1} to wavelength in nm.

$$\lambda = \left(\frac{1}{\lambda_{source}} - \frac{\widetilde{\nu}}{10^7} \right)^{-1} \tag{7.4}$$

Where λ is the wavelength at a particular sample point, λ_{source} is the wavelength of the source laser, and $\widetilde{\nu}$ is the wavenumber at a particular sample point. These wavelength values can be related to the ratio between the QEs depicted in Fig. 7.1 in order to adjust the intensity values recorded at specific points.

7.3.2 Recording Data

Three sets of data were recorded using a standard Raman micro-spectrometer. The first two experiments are recorded to create system specific irradiances for both CCDs. The third experiment was recorded under non-ideal conditions to illustrate how an experimental dataset may be simulated. All experimental and simulated datasets are recorded and created using the modelling described for an FVB read mode.

The datasets illustrated in Fig. 7.2 were used for creating the system specific irradiances. Both datasets were recorded using the specifications detailed in Table 7.1 for the relevant CCDs. An Olympus 4x 0.1 NA mPlan FL MO was used to record the spectra due to it's large depth of focus that minimises the possibility of significant defocusing over the course of an experiment. Acquisition time was set at 5 s for each dataset depicted. 100 spectra as well as a dark current background were recorded for each dataset illustrated.

The third experimental dataset was obtained using the iDus CCD, under the same conditions as detailed for the previous two datasets with the exception that the cooling was altered to -60°C. By altering the temperature, the mean rate of dark current production was increased to 2 e/p/s.

7.3.3 Results

A system specific irradiance was created from the mean spectra depicted in Fig. 7.2 iii). These irradiances were then scaled using the QE ratio illustrated in Fig. 7.1 to simulate an irradiance recorded from the other CCD. The results of this conversion is displayed in Fig. 7.3.

Both simulated spectra show significant correlation with their experimental counterparts. In some regions, converted peaks show slight inconsistencies with experimental intensities. Inconsistencies such as these may be attributed to inaccuracies in fitting the polynomial when modelling the QE. However, the method for measuring SNR i.e. the maximum intensity of the spectrum to the standard deviation of the noise, should be only minimally affected. This is due to the fact that the intensities match well in the region of the highest peak, approximately 1450 cm^{-1}. The inconsistencies may also be attributed to a mi-

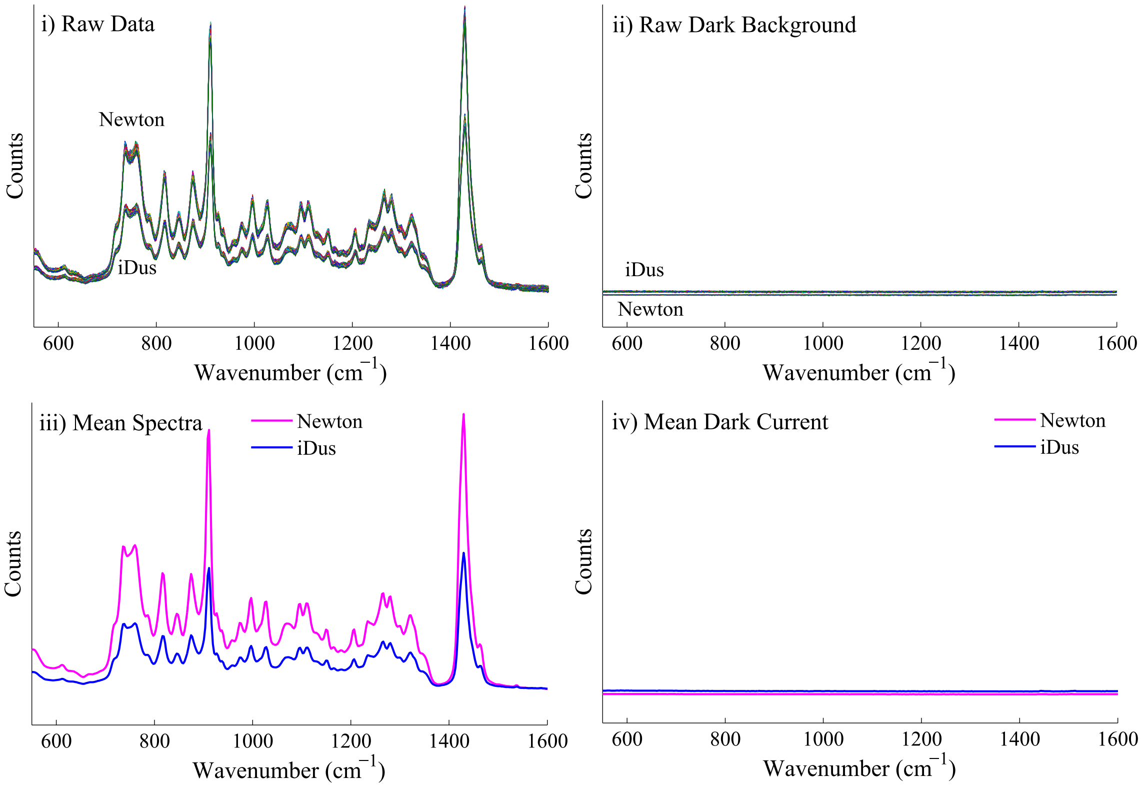


Figure 7.2: Illustration of the raw spectral datasets, i), and dark current backgrounds, ii), recorded from a Newton 920 BVF and an iDus 420 BR-DD. Also depicted are the mean spectra, iii) and iv), from which the system specific irradiances were created.

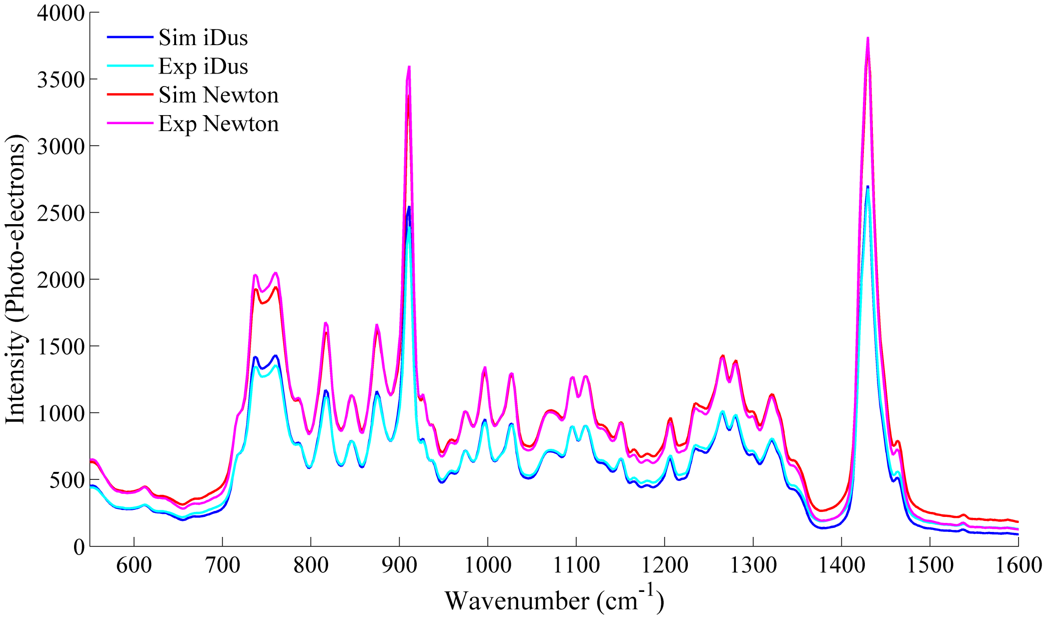


Figure 7.3: A comparison of the system specific irradiance extracted from both cameras and the predicted system specific irradiances created using the scaling ratio described in Fig. 7.1.

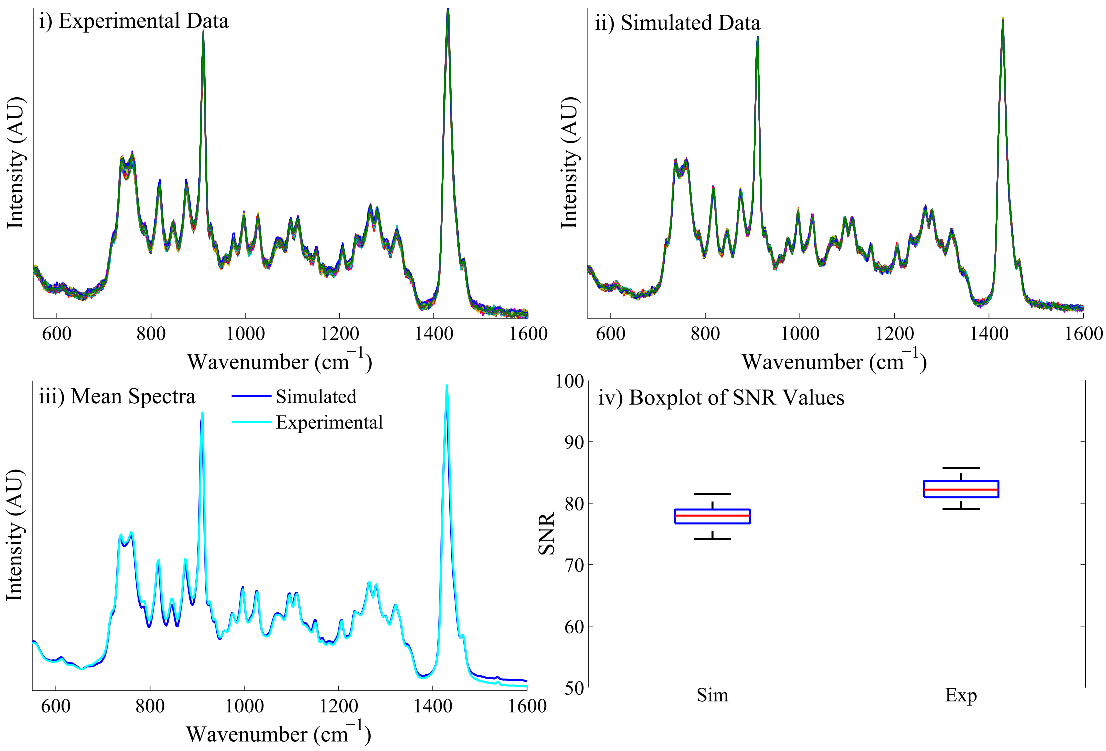


Figure 7.4: An illustration of i) experimental data recorded from the iDus at -60°C, ii) simulated data created from the system specific irradiance recorded from the Newton, iii) a comparison of the mean spectra from both experimental and simulated datasets, and iv) a boxplot comparison of the SNR values from both datasets.

nor misalignment caused by mounting the second CCD. While this is an unavoidable consequence of the experimental procedures required to experimentally validate the theory, the similarity between the experimental and simulated irradiances is a positive indicator of their validity for use of simulating experimental datasets. This hypothesis is tested by using the simulated iDus irradiance to simulate the experimental dataset recorded from the iDus CCD described in Section 7.3.2.

Fig. 7.4 illustrates accurately an experimental dataset recorded from the iDus at -60°C may be simulated using a system specific irradiance created from data recorded from the Newton at -80°C.

Visually, the experimental dataset displays a higher level of variance in the spectra. This can be expected in experimental datasets as minor vibrations in the environment will cause subtle shifts in the system and consequently alter the collected irradiance. However, both experimental and simulated datasets correlate well. They show similar spectral profiles and result in similar SNRs. The primary cause of the difference in SNR is the largest peak (1450 cm^{-1}) is slightly more intense in the experimental spectrum than in the simulated, illustrated in Fig. 7.4 iii). It is assumed that the increase in peak intensity is due to this increase in temperature. Increasing the CCD temperature can cause subtle changes in the QE of the CCD, particularly in the NIR region. [83] Increasing the temperature of the CCD allows the silicon face of the CCD to more readily create a photo-electron. Despite this, the difference in SNR is relatively low with simulated results being approximately 6% lower than experimental results.

While the results presented in this section do not precisely replicate experimental data, they show significant promise for the modelling of optical elements in the Raman spectrometer, particularly the CCD. By amending the experimental conditions to limit the experimental variation caused by the vibrational effects of the environment and the change in CCD temperature it is possible that the results presented here could be improved upon and provide reliable methodology for predicting the effect of replacing optical elements in a given Raman system.

7.4 Summary

In this chapter, it has been shown that it is possible to simulate the effect of replacing the CCD in a given Raman system. While there are a number of factors to consider, such as temperature and system drift, the spectral profile of an alternative CCD can be simulated using data recorded from a given CCD, provided the relevant *a priori* knowledge of the cameras cooling levels and ADC sensitivity is known.

Methodology that can simulate the effect of alternative optical elements are advantageous for the spectroscopist as it can provide them with an alternative avenue of predicting the improvement in SNR before investing in an expensive piece of equipment. It may also allow the user to reduce the cost of the equipment for example if a CCD with higher mean levels of dark current produces a similar SNR in crop mode when compared to a more expensive CCD with lower dark current levels in FVB mode.

The observation that the QE increases with temperature is notable and could provide an avenue for exploitation, especially in terms of portable Raman systems that cannot rely on bulky nitrogen or water cooling systems. The caveat of increasing the temperature is that the dark current contribution will increase significantly. In order to justify the increase in temperature it would be necessary to provide methodology to shorten the acquisition time and optimise the denoising software applied in post-processing to compensate for the reduction in SNR caused by the increase in dark current levels. While the relationship between temperature and increased quantum efficiency is not explored in this thesis, the following suggested avenues that may be used to exploit this phenomenon are investigated in terms of increasing SNR to increase system throughput in the aforementioned chapters.

Optimal software denoising techniques can also reduce the acquisition time by increasing the SNR from the resulting denoised signals. A CRA removal technique is proposed and is shown to produce spectra with superior SNR to that of the double acquisition method in Chapter 8. An enhanced version of Savitzky-

Golay smoothing is presented in Chapter 9 that has been shown to produce spectra with higher SNR and increased peak fidelity than traditional Savitzky-Golay smoothing.

Chapter 8

An Algorithm for the Removal of Cosmic Ray Artefacts in Spectral Datasets

This chapter is related to the following manuscript that has been submitted for journal publication:

Sinéad Barton and Bryan Hennelly. "An algorithm for the removal of cosmic ray artefacts in spectral datasets". Accepted for Publication. Applied Spectroscopy.

I would like to acknowledge Claire Molony's contribution to this work for creating the biological samples as well as recording the datasets used to test the algorithm in Section 8.6.2.

8.1 Introduction

Unlike most of the previous chapters, which are based on evaluating and predicting system performance for the purpose of designing an optimal experimental configuration in terms of SNR, Chapters 8 and 9 both deal with software solutions that can be applied to any recorded Raman spectrum for the purpose of increasing the SNR. The focus of Chapter 8 is the removal of cosmic ray artefacts (CRAS). CRAs are a source of noise that is system independent, as described in Section 4.2.5. The distortion of spectra caused by the presence of CRAs can pose problems for various applications that involve the identification of specific peaks and will affect the methods used to calculate the SNR of a spectrum, as outlined in Chapter 4. CRAs can also impact on the results of the post processing algorithms described in Chapter 3, due to biasing of the loading vectors towards large outliers, which in turn leads to the misclassification of spectra. [84] The misclassification of spectra can be of critical importance, particularly in the growing area of chemometrics. [35, 36].

As previously stated in the conclusion to Chapter 6, single acquisitions of spectra provide superior SNR to multiple acquisitions averaged together. Therefore, a CRA removal technique that does not require this averaging process would be advantageous. The primary goal of this chapter is to provide an algorithm that will reliably remove CRAs from spectra without the requirement of recording multiple acquisitions that the commonly used double acquisition method relies on. This proposed algorithm is shown to produce spectra with higher SNR than that achieved by the double acquisition method.

Existing CRA removal techniques fall into four distinct categories, that will be described in the following section. Here a novel CRA removal algorithm is proposed that combines aspects from the first two categories and has the advantages of both; the method requires only a single capture but works on the same principle as the double acquisition method and visually provides comparable results, i.e. it removes only cosmic rays and makes no other changes to the spectrum. However, the spectra processed using the proposed method requires the availability of a dataset of spectra that can be used for comparison, the most similar of which is identified using normalised covariance. The spectrum of interest is then directly compared with the matching spectrum and differences exceeding a specified threshold are identified as cosmic rays. The contaminated pixels are replaced with the corresponding spectral value from the matching spectrum. The optimal value of the threshold is estimated based on the standard deviation of the spectrum, which is indicative of the level of noise present in the spectrum. The algorithm can be applied to an entire dataset of recorded spectra without intervention from the user.

In addition to being a single acquisition method, the proposed algorithm has a second advantage over the double acquisition method, in that it may offer a significant improvement in the SNR of the denoised spectrum under certain conditions, due to the reduced instances of camera read noise that are included, which is discussed in more detail in later sections. The requirement for an available dataset of spectra is naturally met for a large number of applications that involve the repeated capture of data such as Raman based chemometrics for the detection of bladder cancer [85], cervical neoplasia [86], and breast cancer detection [87] etc. Applications such as these require repeated measurement from cell or tissue samples and, therefore, a dataset of related spectra will often be readily available.

8.2 Existing Methods

The first category of CRA removal methods comprises single scan methods that rely on the assumption that CRAs will have an appreciably narrower width than the expected peaks in the spectrum. This requires that the spectral resolution of the system is less than the width of the spectral peaks, which may not always be the case and depends on the properties of both the source laser and spectrograph in the recording system as well as the chemical composition of the sample under investigation. Methods in this category include the ‘missing point polynomial filter’ [88, 89], the wavelet transform method [90], filtering based on fuzzy logic [91], weighted moving filters [92], and median and low pass filtering. [93] In many cases, the methods in this category are unsuitable because they are either insensitive to CRAs that have comparable width to the features of the underlying spectrum, or they rely on empirically chosen thresholds that may vary between datasets. As a result, in some cases the denoised spectra must be subjected to robust error checking and this can limit the inclusion of these algorithms in fully automated applications.

The second category of methods for the removal of CRAs is based on the low probability of CRAs contaminating the same pixel in sequential or spatially adjacent spectra. This relates to the approach whereby spectra are sequentially recorded in a 2D or 3D grid for the purposes of hyperspectral mapping. The algorithms in this category include the upper bound spectrum (UBS) method and it’s improved variations [94–96], mapping techniques [97] and the double acquisition method used by manufacturers of commercial optical spectroscopy systems such as Horiba. [10] The mapping technique [97] requires a map of spatially adjacent spectra. A nearest neighbour comparison is performed and the most closely correlated spectrum is selected. An offset is selected based on the expected noise and if the intensity value of a spectral component in the original spectrum differs from the corresponding value in the offset spectrum by a value exceeding said offset then the lower value is taken. The algorithm presented in this paper is similar to this approach, except that comparison is performed across an entire dataset of spectra rather than over a set of spatially adjacent neighbours.

The third category is based on continuous analysis in the time domain [98, 99] and is applied to online process monitoring. This approach searches for sharp, unsustained alterations to the spectral intensity values in the time domain at a rate that is faster than the expected rate of change of the material being monitored. Optimisation of optical systems in order to avoid detection of CRAs, such as image curvature correction, [100] is a fourth option. In this case CRAs are detected by comparing spectra recorded along different rows of pixels on the detector. Aberration caused by the imaging system may necessitate numerical correction prior to comparison.

Although all of these methods have been shown to be effective CRA removal methods, in some cases they are computationally intensive or rely on expensive equipment, which may not be feasible. The double acquisition method [10] is arguably the most commonly used approach due to its simplicity and accuracy. The proposed algorithm aims to simulate the robust nature of this method while providing the advantages of single acquisition. In addition to being a single acquisition method, the proposed algorithm has a second advantage over the double acquisition method, in that it offers a significant improvement in the SNR of the denoised spectrum under certain conditions, due to the reduced instances of camera read noise that are included, which is discussed in more detail in later sections.

8.3 Noise in a spectrum: single vs. double acquisition

Multiple acquisitions, whereby a number of spectra are averaged together for the purpose of CRA removal, can have a negative impact on the SNR of the resulting denoised spectrum. Shot noise and dark current noise are both modelled by time-dependent Poisson distributions. Therefore, if only these two noise sources are considered, a spectrum collected with a 5 second acquisition time will have the same SNR to two 2.5 second spectra collected under the same conditions and averaged together. However, read noise is time-independent and will be included in each individual recorded spectrum and, therefore, averaging a number of acquisitions together will introduce multiple instances of read noise. The SNR in a single sample of the spectrum is defined as follows: [53]

$$SNR = \frac{iq(\lambda)tp_i}{\sqrt{[iq(\lambda)p_i + cp_{dc}]t + p_r n_r}} \tag{8.1}$$

where i represents the spectral irradiance that is incident on each individual pixel in a column of p_i pixels, which depends on the spatial distribution of the light arriving at the spectrograph slit; $q(\lambda)$ is the quantum efficiency of a pixel for the incident wavelength λ, and t is the total camera integration time; c is the mean rate of dark current production in electrons per pixel per second; p_{dc} is the number of pixels contributing dark current noise and p_r is the number of pixels contributing read noise to the spectral component. A detailed discussion on noise contributions for different camera read modes is given in Chapter 6. For this chapter we are focused on spectra recorded in FVB mode.

Using Equation 8.1, it possible to compare the SNR of a single acquisition of time T to X acquisitions, each of time T/X duration, which are subsequently averaged. Assuming the camera mode is consistent, both cases will result in a spectrum that has the same spectral intensity (i.e. the numerator in Equation 8.1 will be $iq(\lambda)p_iT$ for both cases). Similarly, the dark current contribution will also be the same. However, the read noise contribution will differ for both cases; for the single acquisition $p_r = 1$ and for the multi-acquisition $p_r = X$. Therefore, it can be expected that the multi-acquisition will have a reduced SNR when compared with a single acquisition of equivalent duration. The difference between these two SNRs will be determined by the values of i, c, n_r, and X.

8.4 Proposed Algorithm

The first step in the proposed algorithm is to assign a best matching pair to each spectrum in a given dataset, thereby, removing the need to record multiple spectra. These pairs of matching spectra are then denoised in a similar manner to that of the commonly used double acquisition method, by identifying corresponding samples in the spectrum for which there exists a difference in intensity that is greater than a threshold that relates to the expected noise level. The final step in the algorithm is to apply a smaller threshold to the immediate neighbours of a sample that has been contaminated with a cosmic ray in order to ensure that even broad CRAs are effectively removed from the spectrum.

Step 1: In order to pair spectra together, an approach similar to nearest neighbour comparison (NNC) [97] is employed, which identifies spectra in a given dataset that share a high normalised covariance. The normalised covariance is calculated as follows:

$$C_{nm} = \frac{(S_n \cdot S_m)^2}{(S_n \cdot S_n)(S_m \cdot S_m)} \tag{8.2}$$

where C_{nm} denotes the normalised covariance of spectra n and m in the dataset and '$\cdot$' represents the dot product. For each spectrum in the dataset, i.e.$n = \{0, 1, 2...N-1\}$, the value of C_{nm} is calculated for all values of $m = 0, 1, 2...N-1$ where $m \neq n$. For a given spectrum S_n, the spectrum S_m that corresponds to the maximum value of C_{nm} is taken to be the most similar and is paired with S_n for the next stage of the algorithm. In this way, each spectrum in the dataset, S_n, is given a pair denoted by $S_{n'}$.

Step 2: *A priori* knowledge can be used to calculate the standard deviation of the noise in a spectrum; CRAs are then identified as spikes that exceed some threshold that is proportional to this value. A similar approach has been proposed in the double acquisition method. [10] However, this method requires knowledge of the specifications of the spectrometer as well as the expected irradiance, which may not be available. For this reason, we propose a method to estimate the standard deviation of the noise in a given spectrum, n, without any *a priori* knowledge of the recording system, as defined by Equation 8.3.

$$\sigma_n = \frac{1}{N}\sqrt{\sum_{k=1}^{M}[S_n(k) - \bar{S}_n(k)]^2} \tag{8.3}$$

Where k is the k^{th} sample of the spectrum, the value of k ranges from 1 to M, and $\bar{S}_n(k)$ is the Savitzky-Golay filtered version of the raw spectrum. If the intensity of the residuals resulting from $S_n(k) - S_{n'}(k)$, exceeds the threshold given by $5\sigma_n$ the pixel is deemed to be corrupted and is replaced with $S_{n'}(k)$. This process is formally defined in Equation 8.4 and is repeated for all values of k from 1 to M.

$$S_n(k) = \begin{cases} S_n(k) & S_n(k) - S_{n'}(k) < 5\sigma_n \\ S_{n'}(k) & S_n(k) - S_{n'}(k) > 5\sigma_n \end{cases} \tag{8.4}$$

The $5\sigma_n$ threshold ensures that >99 of the noise inherent in the recorded signal i.e. shot noise, dark current, and read noise will fall within this boundary. The likelihood of a CRA being detected where there is none is < 1. We note that the algorithm described above is similar to the double acquisition method. Corresponding samples that have a disparity greater than the defined threshold are not averaged and the lesser sample value is taken.

Step 3: It is possible to further amend the algorithm described above in order to deal with the case in which a CRA has a larger width than a single pixel and extends into neighbouring pixels although possibly falling under the specified threshold. A reduced threshold can be applied to the pixels immediately around a detected CRA; this process is formally defined in Equation 8.5 and is repeated for each value of k corresponding to the sample location of a detected CRA in Step 2.

$$S_n(k \pm 1) = \begin{cases} S_n(k \pm 1) & S_n(k \pm 1) - S_{n'}(k \pm 1) < 2\sigma_n \\ S_{n'}(k \pm 1) & S_n(k \pm 1) - S_{n'}(k \pm 1) > 2\sigma_n \end{cases} \tag{8.5}$$

This addition improves the overall sensitivity of the algorithm to include broader CRAs.

In the case of biological spectra, varying baselines and sample heterogeneity can produce significant inconsistency across the spectra in the dataset, which can reduce the capability of the proposed method to find an accurate match within the dataset for a given spectrum in Step 1. In this case, it is recommended to perform a pre-processing step in the form of a background subtraction algorithm. [38,41,42] on the dataset in order to reduce variability and ensure a high correlation between matched spectra. This step can easily be reversed following the CRA removal algorithm by reintroducing the subtracted baseline back to each

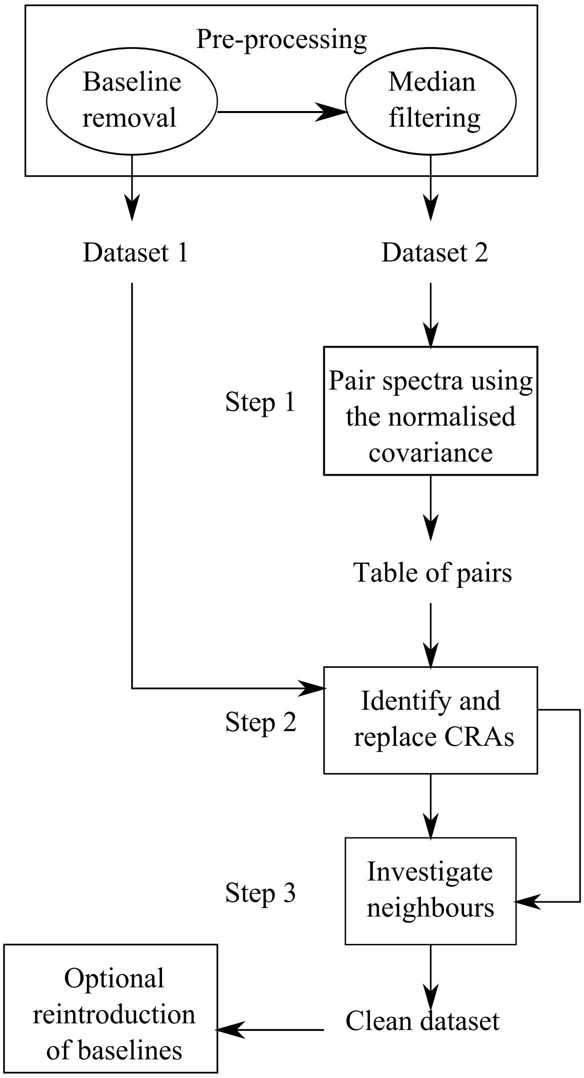


Figure 8.1: A flow chart of the proposed algorithm with additional pre-processing steps to deal with varying baselines.

respective spectrum, if desired.

An additional pre-processing step that may be required for large datasets is to apply median filtering in advance of matching the pairs of spectra. Large datasets that are obtained using long acquisition times, will contain a large number of CRAs and, therefore, the likelihood of a CRA appearing across multiple spectra in the dataset increases. Due to the intensity of these spikes, it is likely that Step 1 of the algorithm will match these spectra together due to their high covariance. In order to avoid this, a median filter can be applied to the dataset and the normalised covariance in Step 1 may be calculated based on this filtered dataset. A flow chart of the overall algorithm including these pre-processing steps is illustrated in Fig. 8.1.

In the sections that follow, the proposed algorithm is applied to datasets of Raman spectra, and the performance is compared to that of the double acquisition method.

8.5 Materials and Methods

8.5.1 Materials

A polymer reference material, acquired from Ibidi GmbH, [101] was chosen as the first sample for investigation due to it's thermal stability, resistance to photo-bleaching, and strong reliable signal, which reduces the overall experimental variability. The consistency of this sample and its insignificant baseline ensures an accurate assessment of the proposed CRA algorithm in terms of the SNR. The benefits of a single instance of read noise in terms of SNR will be more significant for weak spectral irradiances such as for the case of a Raman spectrum recorded from a biological sample. Ideally, a biological sample would have been used to demonstrate the improvement in SNR afforded by the proposed algorithm when compared with the double acquisition method. However, photo-bleaching and the heterogeneity of biological samples may complicate an accurate measurement of SNR. It was, therefore, decided to use the polymer sample for the evaluation of the proposed algorithm in terms of SNR and to reduce the recorded irradiance to match that of an epithelial cell such that the acquisition times and SNR values would relate to biomedical applications. Following this, the algorithm was applied to spectra recorded from three different cell groups; mesenchymal stem cells and their vascular and osteogenic progeny. For further details on cultivation and preparation of these cells please refer to Ref. [102].

8.5.2 Recording Spectra

A custom built confocal Raman micro-spectrometer was used to record spectra from the polymer material. This system uses a 150mW 532nm laser and a diffraction grating with 600 lines/mm. More details on the specific system can be found in Ref. [12]. A sufficiently defocused (i.e. focal plane above the sample), low numerical aperture MO (Olympus UMplanFl 4x/0.1) was used in order to produce spectra from the polymer material that had an SNR equivalent to that expected from an epithelial cell using a commercial Raman microspectrometer over a 60 second acquisition time. Maximum cooling of the CCD (Andor Newton 920 BVF, -80$^{\circ}C$) was used in order to minimise dark current noise. The low magnification of the MO provides for a large depth of field, which prevents any major change in focus over the course of the experiment, further reducing experimental variability across the acquired datasets. A single acquisition dataset of 100 spectra was acquired with a 60 second integration time and a double acquisition dataset of 2×100 spectra was acquired each with a 30 second integration time so that a comparison of the proposed algorithm to the double acquisition method could be made in the context of SNR. For the cell spectra, a commercial Raman

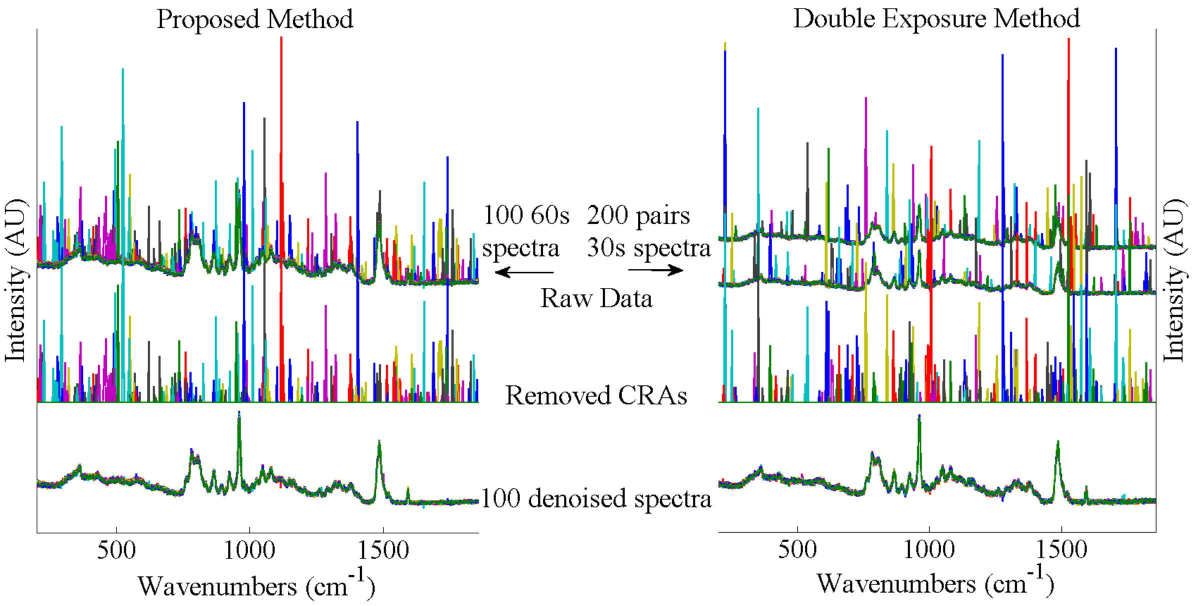


Figure 8.2: Illustration of the datasets used to evaluate the performance of the proposed algorithm in terms of SNR. On the left side a single dataset of 100 raw spectra is shown with an acquisition time of 60s. On the right hand side are two datasets of consecutively collected spectral pairs with an acquisition time of 30s. In both cases the raw data, removed CRAs, and denoised dataset are shown.

micro-spectrometer was employed also using a 532nm laser source. More information on this system is found in Ref. [102].

8.6 Results

8.6.1 Application to polymer data

The spectral datasets from the polymer material were processed using both the proposed algorithm and the double exposure method. The resulting CRA removed spectra were examined and compared in terms of SNR using the second approach described in Section 4.3.1. Fig. 8.2 illustrates the raw data, the removed cosmic rays and the denoised dataset following processing with both methods. Both algorithms make negligible changes to the underlying spectrum, aside from the areas contaminated with CRAs while retaining a high sensitivity for low intensity and broad CRAs.

Fig. 8.3 shows a magnified region ($825cm^{-1}$ to $975cm^{-1}$) of the spectra to further illustrate the effectiveness of the method. This region was chosen in order to illustrate the algorithms ability to discriminate between spectral features and CRAs as it contains a number of peaks that vary in width and height.

While both methods perform similarly in terms of CRA removal, there is, however, a difference in the SNR of the denoised spectra obtained using the two methods. It should also be noted that in the denoised dataset of the double acquisition method illustrated in Fig. 8.2, there are the remnants of two CRAs evident

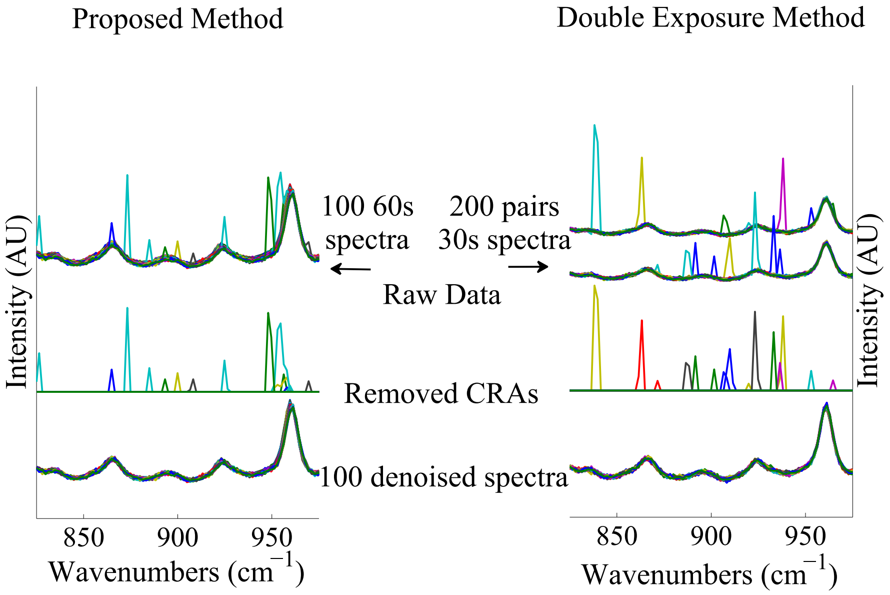


Figure 8.3: A magnified region of Fig. 8.3 showing the before to further illustrate the operation of the proposed algorithm.

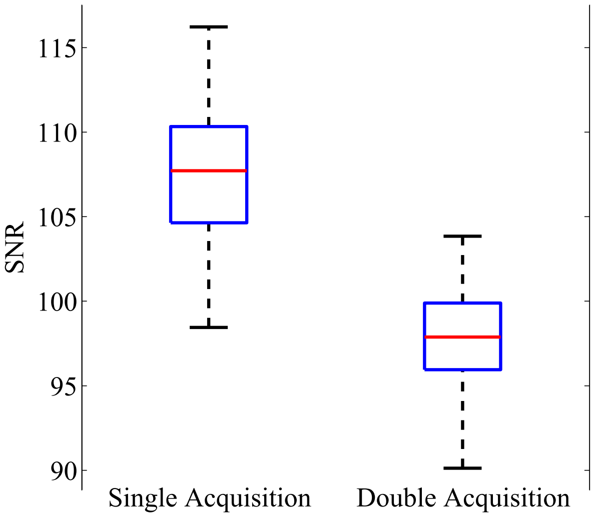


Figure 8.4: A boxplot of the resulting SNRs of the CRA removed spectra of both the proposed algorithm and the double exposure method.

at approximately $1800cm^{-1}$. These are the remnants of two intense cosmic rays that were spread over multiple pixels. The outer edges of these CRAs were small enough to fall under the designated threshold. In cases such as this, investigating the neighbouring pixels of identified CRAs with a lower threshold is necessary.

The mean spectrum of all 300 spectra collected in the experiment was used as the reference spectrum for measuring the SNR as described in the previous section. Fig. 8.4 illustrates the SNR calculated over the dataset of 100 denoised spectra for both the proposed algorithm and the double acquisition method.

Of the dataset that is denoised by the proposed algorithm, the the range of SNR values is 98 to 117 with the central 50 of SNR values in the range of 104 to 111. For the dataset that is denoised by the double acquisition method, the range of SNR values is 90 to 104 with the central 50% of SNR values in the range of 96 to 99. More than 75% of the denoised spectra processed using the proposed method exhibit higher SNR values than those denoised using the double acquisition method.

8.6.2 Application to biological data

In order to evaluate the proposed algorithms performance on biological spectra, three datasets recorded from three different cell groups were amalgamated into a single dataset to which the algorithm was applied. Three datasets were recorded from i) mesenchymal stem cells (MSC), ii) the vascular progeny of MSC samples, and iii) the osteogenic progeny of MSC samples. It should be noted that the osteogenic cells contain a noticeable difference to the other samples, specifically the peak at $960cm^{-1}$ that indicates the presence of phosphates. Both pre-processing steps were applied as illustrated in Fig. 8.1. A mean spectrum taken from the entire dataset was used in the background subtraction algorithm and a polynomial of order 5 was also used to remove varying baselines in the dataset. This fitted dataset was then filtered using a median filter of size 11 before applying Step 1 of the proposed algorithm. Fig. 8.5 illustrates the datasets at different stages of the CRA removal algorithm from raw spectra in Fig. 8.5 (i) to the final dataset without CRAs in Fig. 8.5 (iv).

Fig. 8.5 (ii) shows the raw data following background subtraction. The resulting dataset is then CRA removed using the proposed algorithm to produce the denoised dataset shown in Fig. 8.5 (iii). Finally, the background components are reintroduced to each individual spectrum.

8.7 Summary

CRAs can be removed from spectra using a number of different methods. These include algorithms that can be directly applied to a single recorded spectrum using some form of digital filtering; although, such algorithms have the advantage of being applicable to dynamically changing samples, the goal of removing the cosmic ray while making no other change to the spectrum is challenging. A second group of algorithms for the removal of cosmic rays involves the capture of successive spectra from a sample that is not expected to change between captures. A direct comparison of subsequent spectra allows for the accurate removal of cosmic rays while making little or no other alteration to the underlying spectrum. The proposed algorithm relates to both of these approaches; the algorithm requires only a single recorded spectrum so long as a dataset of similar spectra is available, a requirement that is naturally met for a large number of applications

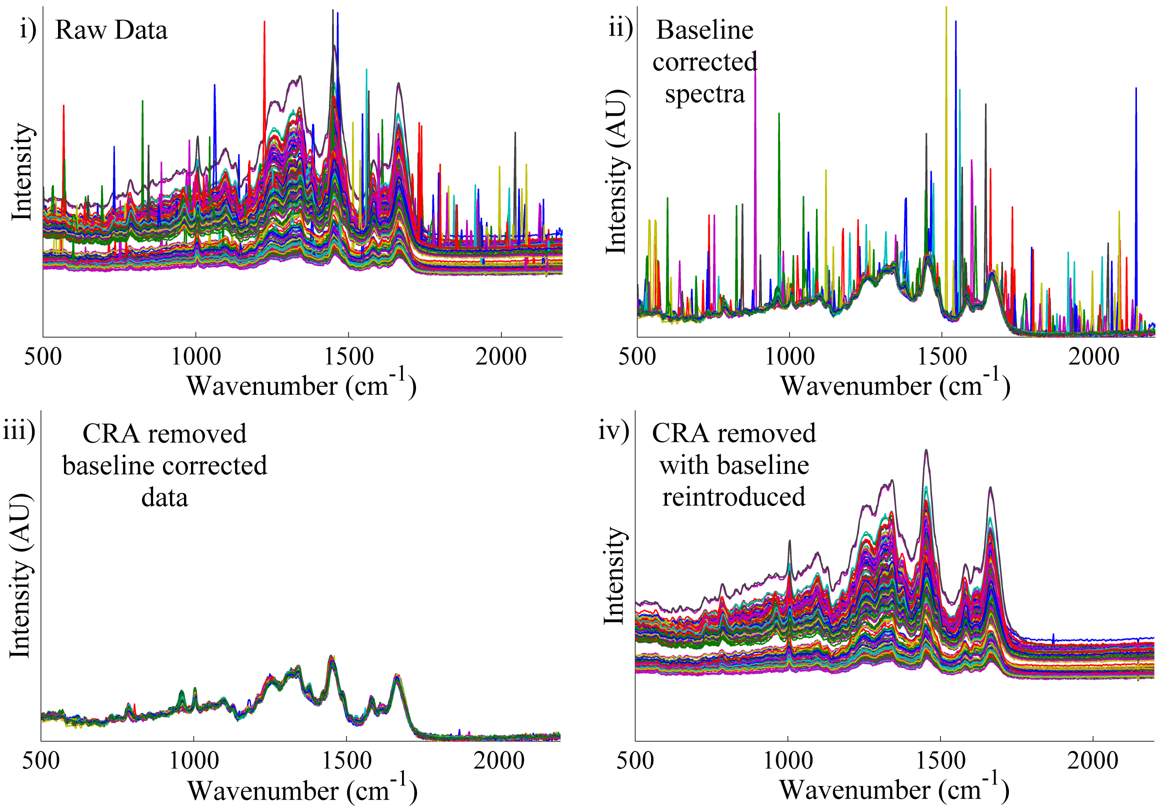


Figure 8.5: An illustration of the the CRA removal of the dataset. The raw data is background subtracted, CRA removed, and, finally, the background is then reintroduced to the data. The y axes are fixed to the same height for all figures.

that involve the repeated capture of data.

In this paper, it has been demonstrated that the proposed algorithm does not require any *a priori* knowledge of system and camera parameters that were used to record the spectrum. In terms of effectiveness at CRA removal, this method performs similarly to the double acquisition method [10] which is widely applied in the field of Raman spectroscopy. For those cases where the amplitude of the shot noise and camera dark current dominates, this difference in SNR between a single and a double acquisition may be negligible; however, in applications where low intensity spectra are collected or high read-out rate are required, this additional noise may become a significant factor and negatively affect the SNR. The proposed algorithm has the advantage that it does not require the repeated capture of spectra and has shown that an overall improvement of SNR of 10% can be expected for the recording conditions associated with biological samples. A second advantage of the proposed algorithm over the double acquisition method is that databases of previously recorded spectra can also be processed. It is notable that the algorithm is able to successfully pair the spectra within the dataset despite the presence of spectra from three distinct cell groups. It can be expected that this feature may be extended to datasets containing a large number of spectra originating from disparate sources.

It must be acknowledged that the proposed algorithm will fail if a recorded spectrum contains legitimate spectral peaks that are unique to the dataset that is employed for cosmic ray removal; in such a case, such peaks would be deemed to be cosmic rays and removed. However, for many applications of spectroscopy, and for a sufficiently large dataset, the probability of such an occurrence can be expected to be low.

CRA removal is typically only the first denoising technique applied to experimental datasets. After CRA removal, SG smoothing is routinely applied to spectra to reduce the the experimental noise. However, SG smoothing can degrade the spectral peaks present in the spectrum. Unfortunately, this provides a trade-off between degrading peaks and effectively removing the noise. Therefore the following chapter illustrates how SG smoothing may be enhanced with maximum likelihood estimation to yield spectra with increased SNR when compared to traditional SG smoothing.

Chapter 9

Algorithm for optimal denoising of Raman spectra

The work presented in this chapter is related to the following journal publication:

Sinéad J. Barton, Tomas E. Ward, and Bryan M. Hennelly. "Algorithm for optimal denoising of Raman spectra." Analytical Methods 10.30 (2018): 3759-3769.

9.1 Introduction

This chapter aims to improve the SNR of Raman spectra by using post-processing software to enhance the SNR of the recorded Raman spectrum. Effectively removing experimental noise from experimental spectra without affecting the underlying spectral profile is a challenge. Savitzky-Golay (SG) filtering [11] is commonly used to smooth spectra in order to reduce the impact of noise on statistical classification. [37,103] However, it can in some cases be counter-productive as the smoothing technique may degrade the spectral features that the multivariate statistical analysis, described in Chapter 3, relies on to accurately discriminate between spectra recorded from different biological samples. The goal of this chapter is to improve the SNR of experimental spectra by enhancing SG filtering using the knowledge of the noise present in the system described in Chapter 4. This chapter also uses the methodology described in Chapter 6 to create large datasets of artificial spectra to test the proposed algorithms performance across a range of different SNRs.

The SG filtering technique works by dynamically fitting a polynomial to consecutive windows of data points (local least-squares polynomial approximation) in order to follow the shape of the spectrum thereby mitigating the impact of a randomly varying noise signal. Under certain conditions, this can have a negative impact on spectral features; in particular high noise applications that require high levels of smoothing, which may severely affect sharp local features. Maximum Likelihood Estimation (MLE) is a statistical process that enables signal denoising [104] by searching for the most likely value of the signal based on a sequence of measured values and *a priori* knowledge of the noise distribution associated with the collected signal. The proposed algorithm merges the robust smoothing of the SG filter with the restriction that the denoised data must be constrained the noise distribution provided by MLE. In this chapter, we demonstrate that this algorithm consistently returns a spectrum with a higher SNR than SG filtering alone, as well as effectively preserving the fidelity of sharp peaks.

The proposed denoising algorithm is constructed on the combination of Maximum Likelihood Estimation (MLE), based on estimating the noise in a spectrum, with Savitzky-Golay (SG) smoothing. The algorithm attempts to overcome the classical problem of providing a smooth noise free spectrum, without affecting the fidelity of sharp spectral features in the process. The algorithm is comprised of two competing constraints that are applied to a given spectrum: the first condition, which makes use of SG filtering, assumes that the spectrum is smooth, i.e. that a given sample will not differ significantly from its neighbours; and the second condition, which is based on MLE, requires that the sample does not deviate significantly from the raw value that was recorded, taking into account the noise distribution that exists for that raw sample value.

The chapter is split into four main sections. Section 9.2 an overview of how the properties of the noise sources are modelled within the context of Maximum Likelihood Estimation, and the integration of Savitzky-Golay smoothing with MLE. Section 9.3 defines the metrics that are used to evaluate the performance of the algorithm, the creation of artificial data on which to evaluate and compare the data, and the steps taken to optimise the algorithm's input parameters. Section 9.4 provides the result and illustrates the SNR improvement provided by the proposed algorithm over competing denoising algorithms for experimental and simulated data. Finally, in Section 9.5 we offer a brief conclusion and propose a number of possible avenues for further improvements.

9.2 Theory

9.2.1 Modelling Noise

The distributions associated with the main forms of noise have already been defined and discussed in Chapter 4. However, it is also possible to approximate a Poisson distribution as a Gaussian distribution provided the mean photo-electron count registered in the camera pixel is high enough. Therefore, if the spectral irradiance is sufficiently high, the total noise, described in Chapter 4, can be estimated with a single additive Gaussian distribution, and this enables the denoising process to be modelled as a decomposition problem, $y = x + d$, where y is a vector of discrete samples that is the recorded spectrum, x is the true spectral intensity in units of photons collected in each pixel area over the full acquisition time, t, and d is the noise signal, which is defined in terms of the following Gaussian probably distribution:

$$p(d_i) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left[\frac{-(d_i - \mu_i)^2}{2\sigma_i^2}\right] \tag{9.1}$$

where:

- $i$is an integer index that denotes the i^{th} discrete sample in a spectrum
- $\mu_i = r + tc_i$, i.e. the mean value of the distribution in the i^{th} sample (in electrons per second) is given by the sum of the mean read noise, r (in electrons), and the product of mean dark current, c_i, and time, t.
- $\sigma_i^2 = x_i + tc_i + \sigma_r^2$, i.e. the variance of the noise distribution is given by the sum of the variances of the individual noise terms. [53]
- The spectral intensity can be defined in terms of the spectral irradiance as follows: $x_i = tl_i$, where l_i denotes the irradiance in photons per pixel per second.

It is notable in the above description of the noise term, d, that the dark current noise can vary from sample to sample, which is due to the variable properties of the semiconductor pixels of modern CCD detectors, while the read noise is assumed to have a constant mean value and standard deviation across all pixels in the detector.

9.2.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a statistical method whereby the parameters of a known statistical model can be estimated based on a number of observations; this is achieved by calculating the parameter values for that model, which maximise the likelihood of making the set of observations. In order to use MLE to reduce the noise in a signal, the statistical model for the noise must be known; take for example the above decomposition problem for the noise in a single sample of the spectrum, i.e. the i^{th} sample. Since the values for the dark current and read noise parameters within the Gaussian are known (these can be measured in advance of recording a spectrum), then the only unknown is x_i. If a number of, k, different spectra are recorded, y_{i1}, y_{i2}, ... y_k, then MLE can be applied to determine the most likely value of x_i that would have resulted in this set of observations. However, this approach requires a number of different recordings and the outcome of MLE would be the trivial result that the most likely value is the mean of all the observations minus the mean noise. Here, we set ourselves the problem of applying MLE based only on a single observation. A similar approach has recently been proposed for removing noise from astronomical images, [105] which is of particular relevance to the current discussion due to the similarity between astronomical images and Raman spectra, i.e. areas of dark (flat regions) interspersed with stars (peaks).

We begin the derivation of the algorithm by formally defining the probability of recording an intensity value in the i^{th} sample, y_i, given the true intensity, x_i, as follows:

$$p(y_i;x_i) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left[\frac{-(y_i-\mu_i-x_i)^2}{2\sigma_i^2}\right] \tag{9.2}$$

The mean noise, μ_i, may be subtracted from y_i by recording a dark frame of sufficiently long acquisition time. The standard deviation of the noise, σ_i, varies across the samples due to the varying dark current contributions, c_i, which are often pixel dependent, and the dependence of shot noise on the varying signal intensity, x_i. However, for simplicity and ease of computation, the algorithm assumes a constant standard deviation, denoted $\bar{\sigma}$, for all samples and is calculated as follows:

$$z = y - \mu - SG(y-\mu, v, q) \tag{9.3a}$$

$$\bar{\sigma} = \frac{1}{N}\sum_{i=1}^{N} z_i^2 \tag{9.3b}$$

where N is the total number of samples in the spectrum. The value of $\bar{\sigma}$ that is used in the algorithm is calculated by estimating the mean standard deviation of the global noise term. This is achieved by applying an appropriate SG smoothing filter ($v = 3$, $q = 9$, where v represents the polynomial order and q represents the window size of the filter) to the spectrum, subtracting the smooth from the raw, and finally taking the standard deviation of the remaining signal. Following from this, the negative log likelihood of observing a signal intensity at sample i is:

$$-log(p(y_i;x_i)) = \frac{(y_i - \mu_i - x_i)^2}{2\bar{\sigma}^2} \tag{9.4}$$

9.2.3 Maximising an 'a posteriori' estimator

Denoising in this context requires the use of an image prior, x', i.e. a reference signal to allow the user to deduce *a priori* knowledge of a given spectral sample on the basis that the spectrum should not deviate significantly from the image prior. The probability of the true intensity at sample i, can be defined in terms of the intensity values of the samples in the image prior in the neighbourhood around i as follows:

$$p(x_i) = \prod_{j=i-n}^{i+n} exp[-\lambda |x_i - x'_j|^p] \tag{9.5}$$

where $2n+1$ is the size of the neighbourhood and the λ and p parameters are used to define how closely a sample in x is expected to match the surrounding samples in the image prior; selecting $p > 0$ will impose a constraint that a smooth transition must exist from one sample to the next. In the limiting case, if only one observation is available, we can set $x' = y - \mu$, whereby the image itself serves as it's own image prior, eliminating the need for multiple acquisitions. [106] The basis of the first MLE model described here makes use of a similar approach, whereby the neighbouring samples in the signal can provide a reference for that sample. In this case the values of λ and p determine how smooth the transition should be from one sample to the next. [105–107] Following from this discussion, the negative log likelihood of Equation9.5 can be determined:

$$-log(p(x_i)) = \lambda \sum_{j=i-n}^{i+n} |x_i - x'_j|^p \tag{9.6}$$

Using Bayes' theorem, the negative log likelihood of $p(x_i;y_i)$ can be expressed as follows:

$$-log(p(x_i;y_i)) = -log(p(y_i;x_i)) - log(p(x_i)) \tag{9.7}$$

Explicitly:

$$MLE(x_i) = \frac{(y_i - \mu_i - x_i)^2}{2\bar{\sigma}^2} + \lambda \sum_{j=i-n}^{i+n} |x_i - x'_j|^p \tag{9.8}$$

Therefore, the most likely value of x_i is the one that minimises Equation 9.8 and this equation is the basis of the first MLE based algorithm that we propose here. The algorithm begins by setting $x' = y - \mu$; this involves subtracting a dark frame from the raw spectrum. The second step is to calculate the most likely estimate of x, which we denote as x^e; this is achieved by performing a brute force search to find the sample values that minimise Equation 9.8, which we denote as x^e_i. This process is repeated for each sample, i, until the entire spectrum is estimated. The third step is to set $x' = x^e$, and then to repeat the second and third step iteratively until the conditions for stopping are met; an early stopping strategy is important in order to prevent over-smoothing of key signal features. Although the denoising algorithm described by Equation 9.8 provides meaningful results, we do not investigate it any further in this chapter. A superior algorithm is proposed in the section that follows, which employs a similar approach; the detailed development of the first algorithm above is a necessary first step before introducing the algorithm below.

9.2.4 Improving the ‘a posteriori’ estimator by employing SG smoothing

The MLE algorithm defined in the previous section is similar to a method previously applied to astronomical images [105] and employs a two dimensional neighbourhood of nine pixels around the sample of interest. A Raman spectrum is inherently one dimensional and, therefore, only the samples immediately to the left and right of the sample i can be used in the MLE algorithm. In this section, we propose an improved MLE algorithm that makes use of Savitzky-Golay (SG) filtering. The algorithm is similar to that described in the previous section; however, in this case the first step is to set $x' = SG(y - \mu, v, q)$, where SG denotes the application of an SG filter to the raw spectrum with a dark frame subtracted, $y - \mu$. The second step involves finding the values of x_i that minimise Equation 9.8 as described for the previous algorithm, which results in the estimate x^e. The third step involves setting $x' = SG(x^e, v, q)$; the second and third steps are repeated iteratively and once again an early stopping strategy is employed to avoid over smoothing. The algorithm investigated in this chapter uses a neighbourhood of only 1, i.e. $n = 0$. Therefore, Equation 9.8 reduces to:

$$MLE(x_i) = \frac{(y_i - \mu_i - x_i)^2}{2\bar{\sigma}^2} + \lambda |x_i - x'_i|^p \tag{9.9}$$

The algorithm described above, which will be referred to by the acronym MLE-SG going forward, is essentially a pixel by pixel estimator, that is constrained in two opposing directions. The left term in Equation 9.9 will increase as x_i deviates from the raw value. Conversely, the right term will increase as the estimated value deviates from the SG smoothed version of the spectrum. It can be expected that the algorithm will perform at least as well as traditional SG filtering, and with the additional constraint that the smoothed spectrum is not permitted to deviate far from the recorded value within the bounds of the noise distribution. We can therefore expect superior results in terms of recovering a truer estimate of the underlying Raman spectrum.

9.3 Tuning the Algorithm

The algorithm outlined in the previous section requires five input variables, namely; λ, p, v, q, and the number of iterations, m. An investigation into the optimal values for these parameters was performed in order to minimise the number of input variables and maximise the denoising capability of the algorithm. The results of this investigation are detailed in this section in terms of SNR and a metric that is proposed for the first time here, which we refer to as the SNR product.

9.3.1 Noise metrics for optimisation of parameters

As discussed in Chapter 4.3, mean SNR of the signal is a suitable metric for evaluating the quality of the spectrum as a whole, whether this is calculated based on a reference spectrum or in an experimental context. Here we redefine the SNR equation in terms of the Root Mean Square Error (RMSE) and the notation used in this chapter as well as follows:

$$\begin{aligned} SNR(x^e) &= \frac{max(x^e)}{RMSE(x^e, x^{ref})} \\ RMSE(x^e, x^{ref}) &= sqrt\left\{ \frac{1}{N} \sum_{i=1}^{N} (x_i^e - x_i^{ref})^2 \right\} \end{aligned} \tag{9.10}$$

where $max()$ is a function that returns the maximum value in the input vector and $RMSE$ calculates the root mean square error of the input vector with respect to the reference signal intensity,x^{ref}, both of length N. However, while smoothing may increase the SNR of a spectrum as a whole it can also negatively affect sharp local features, which may be of importance. In order to monitor the effect of the algorithm, specifically on sharp spectral features, the SNR in the neighbourhood of a peak, $x^e[pk-n:pk+n]$, is calculated. This calculation is based on an $2n+1$ sample subset of x^e centred on a feature located at index pk. The SNR for this peak region is defined as follows:

$$SNR(x^e[pk-n:pk+n]) = \frac{max(x^e)}{RMSE(x^e[pk-n:p+n], x^{ref}[pk-n:pk+n])} \tag{9.11}$$

This definition uses the same maximum value as for the global spectrum definition given in Equation 9.10 but *RMSE* is calculated only over the peak region. This ensures a meaningful comparison with SNR values of the global spectrum. An estimate for the SNR of the raw spectrum can be determined by calculating $SNR(y-\mu)$ for the global case and for the peak area $SNR(y[pk-n:pk+n]-\mu[pk-n:pk+n])$, using Equation 9.10 and 9.11 respectively. Finally, in order to reflect the overall SNR improvement that is provided by the denoising algorithm, we propose a novel metric called the SNR product which takes into account the enhanced SNR globally as well as in the region of a sharp peak:

$$SNR_{prod} = \frac{SNR(x^e)}{SNR(y-\mu)} \times \frac{SNR(x^e[pk-n:pk+n])}{SNR(y[pk-n:pk+n]-\mu[pk-n:pk+n])} \tag{9.12}$$

Focusing on the left hand side of the above equation, this term concerns the global SNR and is used to evaluate the mean improvement in signal quality across the entire wavenumber range being examined. This term is primarily influenced by large low frequency regions. The right term focuses on a sharp local feature and is used to monitor whether the algorithm is negatively impacting peaks. If there is no SNR enhancement in the denoised spectrum compared to the raw, then the SNR product will return a value of 1 or lower. The typical range of results for the SNR product is $0 < SNR_{prod} < 4$.

9.3.2 Data Driven Parameter Optimisation

In order to robustly examine the recovery potential of the algorithm, large datasets with varying SNR were required, as well as *a priori* knowledge of x^{ref}. This requirement meant that artificial datasets were best suited for the initial testing and optimisation phase since large amounts of data can be created with known noise parameters and with knowledge of the underlying signal. Datasets with various SNRs were generated based on a signal in the form of a high quality low noise Raman spectrum recorded from a polymer slide (Ibidi Gmbh) [101] due to its resilience to photo-bleaching, thermal stability, intense and reproducible Raman spectrum. In total, 100 spectra were averaged together following subtraction of the mean dark current and mean read noise, which enabled an accurate estimate of the true irradiance in terms of the mean

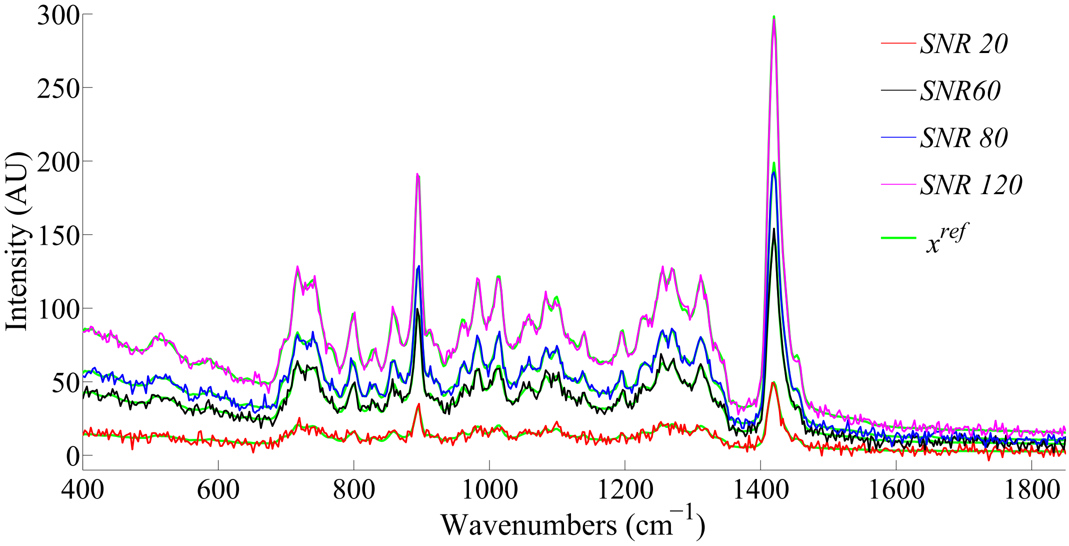


Figure 9.1: Illustration of the artificial dataset noise levels and the corresponding x^{ref}

photons collected per pixel per second. This then enabled the signal intensity, x^{ref} (calculated by scaling the irradiance), and the noise to be simulated based on any acquisition time using Equation9.1. In this way, six datasets were generated with the SNR values of $20, 40, 60, 80, 100, 120$; each dataset contained 100 spectra. Fig. 9.1 illustrates four sample spectra of SNR values $120, 80, 60, 20$ that are approximately indicative of low, medium, high, and extreme noise cases when recording Raman spectra from biological samples.

9.3.3 Optimal Parameters

The MLE-SG algorithm described in Section 9.2 is dependent on five parameters; namely λ, p, SG parameters (v and q), and finally the number of iterations, m. This section describes the steps taken in order to find the best set of parameters to use for a noisy signal with a given SNR. Using all six artificial datasets described in Section 9.3.2 a brute force search over a wide range of λ and p was performed and the results were found to be approximately similar for all six datasets. The λ and p values were fixed at 1.8 and 0.4 respectively, which were found to work well for all cases, and the other parameters were varied in subsequent investigations. Initial testing revealed that SG input parameter combinations made up of $v = \{3, 5\}$ and $q = \{5, 7\}$ showed the most promise for use as spectral priors. All of the parameters were examined in terms of the improvement in the global SNR, peak SNR, and the SNR product of the denoised spectra over a range of iterations ($m = 1, 2, 3...100$). An example of this analysis for an initial SNR of 60 is illustrated in Fig. 9.2, in which the results that are shown are an average across a dataset of 100 spectra.

Results beyond 50 iterations are not displayed in Fig. 9.2 since the SNR recovery has stabilised or is already in decline. These results were reproduced for all datasets previously mentioned. Two important

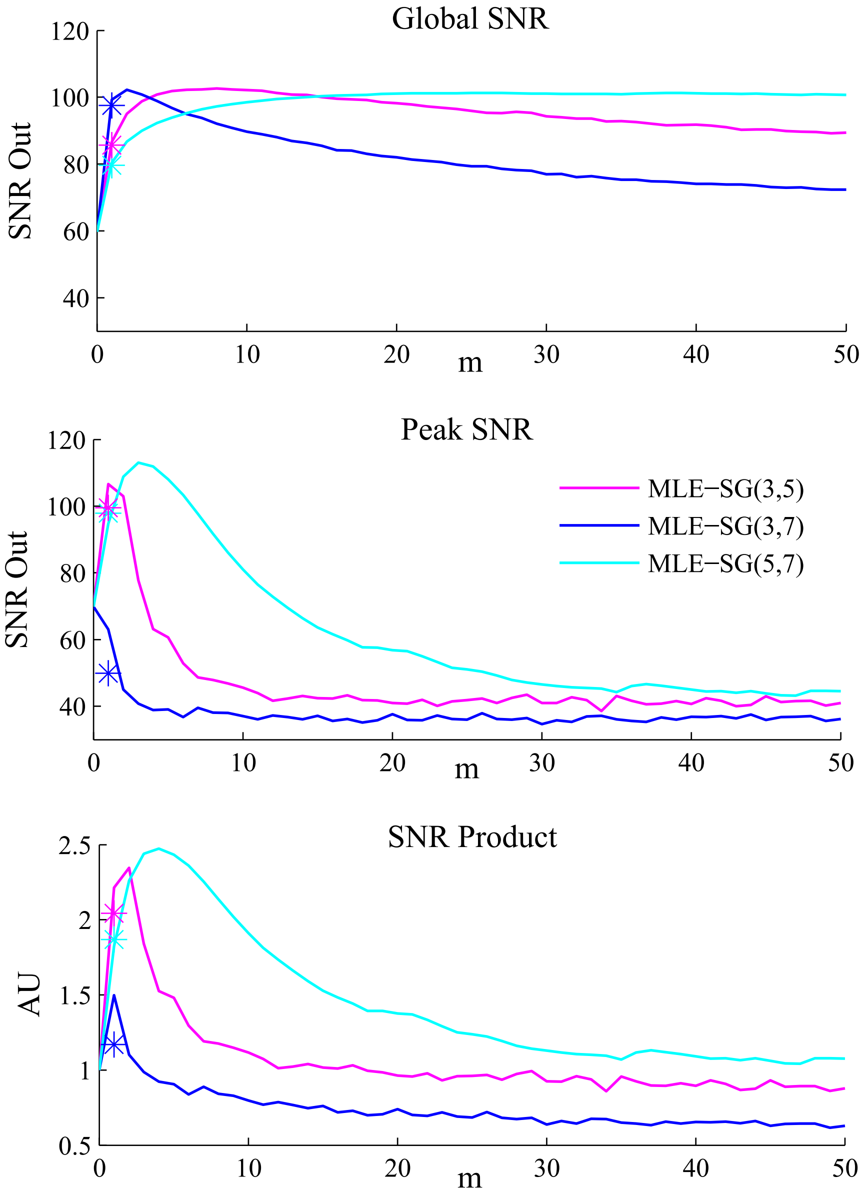


Figure 9.2: An illustration of mean SNR recovery for the datasets with an initial global SNR of 60 for three sets of SG input parameters over 50 iterations (The asterisks aligned with the first iteration are representative of the SNR achieved by SG filtering alone)

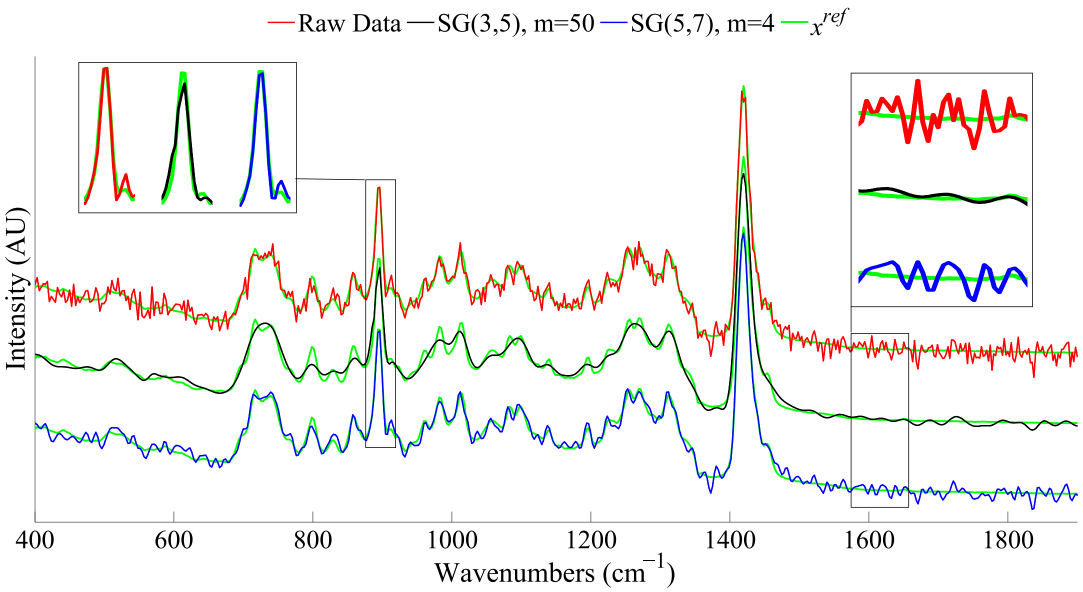
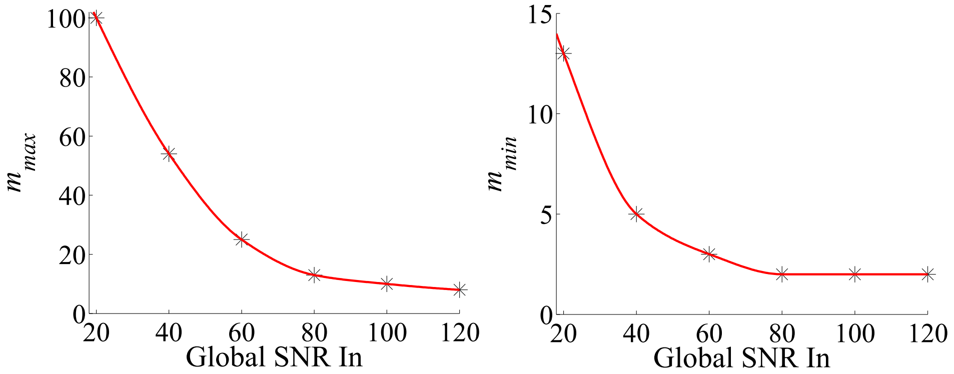


Figure 9.3: An illustration of MLE-SG denoising for different numbers of iterations; denoising of peak regions is optimal at low numbers of iterations, while smooth areas require a significantly larger number of iterations.

results become clear, (i) noisier signals require a greater number of iterations in order to achieve optimal denoising for both peak regions and globally; this is discussed further below, and (ii) in general optimal improvement in peak SNR occurs much earlier than for the global spectrum, in terms of the number of iterations. It was determined that all other parameters, other than m, can be fixed to constant values regardless of input SNR, with approximately similar results. This significantly simplifies tuning of the algorithm for a given input SNR to selecting the most appropriate value of m. From examination of the figures it was determined that the most reliable SG input parameters for preserving peaks was SG$(5,7)$; however, in terms of global SNR, SG$(3,5)$ produces a slightly higher result. The aesthetic difference of these two conflicting requirements, i.e. smoothing vs. peak preservation, is illustrated in Fig. 9.3.

Optimal numbers of iterations for global and peak denoising were derived from maxima in the SNR graphs created for each dataset; from this set of results the optimal number of iterations for both global (m_{max}) and peak (m_{min}) denoising, as a function of input SNR, were found and are illustrated for SG(5,7) in Fig. 9.4.

Ideally, the algorithm should provide high levels of smoothing while effectively preserving the integrity of sharp peak features. Thus, it was decided to further develop the algorithm to implement an early stopping procedure (setting $m = m_{min}$) in peak regions while also applying a late stopping procedure ($m = m_{max}$) in smoother regions. In order to avoid sharp discontinuities between regions of early and late stopping, an approach was developed to ensure a gradual change in the number of iterations from one sample to the next. The development of this procedure is discussed in the following section.



(a) (b)

Figure 9.4: (a) Number of iterations ($m = m_{max}$) required for optimal denoising of the global spectrum, as a function of input SNR; (b) Number of iterations ($m = m_{min}$) required for optimal denoising of a peak region, as a function of input SNR.

9.3.4 Early Stopping at Peaks

Equation 9.9 is comprised of two opposing constraints; the first constraint penalises deviation from the smoothed version of the spectrum, while the second constraint penalises deviation from the raw values. With the first application of Equation 9.9, an initial denoised estimate of the spectrum is obtained that is more accurate than that produced by SG filtering alone, in terms of SNR. A second application of Equation 9.9 is likely to produce a second estimate of the denoised spectrum with a further enhanced SNR. This is due to the fact that the smoothed version of the first estimate, which is used in this second iteration, is a more accurate representation of the spectrum than the smoothed version of the raw spectrum that was used in the first iteration. This argument can be applied to each subsequent iteration up to some point for which the spectrum has become over smoothed, and the SNR of the estimate will begin to reduce. In areas where sharp features are present, it is better to apply an "early stopping" strategy, i.e. to use only a few iterations of the algorithm in order to avoid over smoothing, while in areas of the spectrum that contain smooth features, "late stopping", i.e. application of a large number of iterations, will provide higher SNR values.

The number of iterations, m_i, associated with each sample index i, which is imposed by the presence of a peak at a wavenumber given by pk_j is determined using a Gaussian distribution as follows:

$$\begin{aligned} G(j,i) &= G(j,i)\left[\frac{-|wavenumber_i - pk_j|^2}{2\sigma_g^2}\right] \\ m_i &= min_j[G(j,i)] \times (m_{max} - m_{min}) + m_{min} \end{aligned} \tag{9.13}$$

where $wavenumber_i$ is the spectrum wavenumber axis as a function of sample index i, and σ_g is the

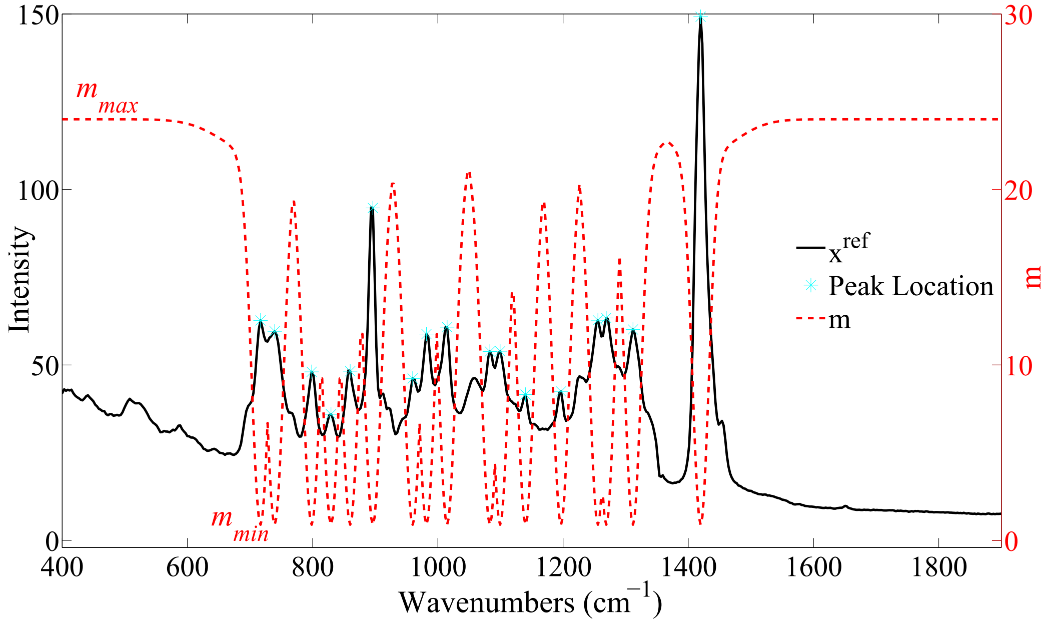


Figure 9.5: The number of iterations is determined using Equation 9.13 and provides an early stopping strategy for the MLE-SG algorithm in peak regions as well as a smooth transition in iteration numbers from one sample to the next.

standard deviation of the the Gaussian, all in units of cm^{-1}. The vector pk_j contains a series of k wavenumber peak locations that is input by the user and, therefore, j takes values of 1 to k. m_i denotes the number of iterations that will be applied to the i^{th} sample in the spectrum and m_{max} and m_{min} are as previously described. The $min_j[]$ operator returns the minimum value in the j dimension. The result of applying the algorithm defined by Equation 9.13 to the polymer spectrum is shown by the red line in Fig. 9.5, where $\sigma_g = 10$.

Samples that are located in large slowly varying regions are associated with high numbers of iterations (m_{max}), while peak locations are associated with a low number of iterations (m_{min}); an appropriate gradient of iterations from m_{max} to m_{min} is calculated by Equation 9.13 to prevent discontinuities in the denoised spectrum. However, disassociating the flat regions from the peaks in this way allows for greater smoothing in the low frequency regions. Rather than increase m_{max}, and therefore the run-time of the algorithm, the window size of the SG filter and λ are increased for the final 20% of the iterations to produce an improved spectrum by providing an increased rate of smoothing in areas that are relatively flat. A flowchart of the algorithm is illustrated in Fig. 9.6.

Automatic identification of peaks in a noisy spectrum is a challenging process and not within the scope of this chapter; therefore, it was decided that the user would input a number of distinct peak wavenumber locations. This is a reasonable approach since many applications involve a set of known peak locations in each recording, e.g. in the case of recording spectra from an epithelial cell, which is discussed in more detail

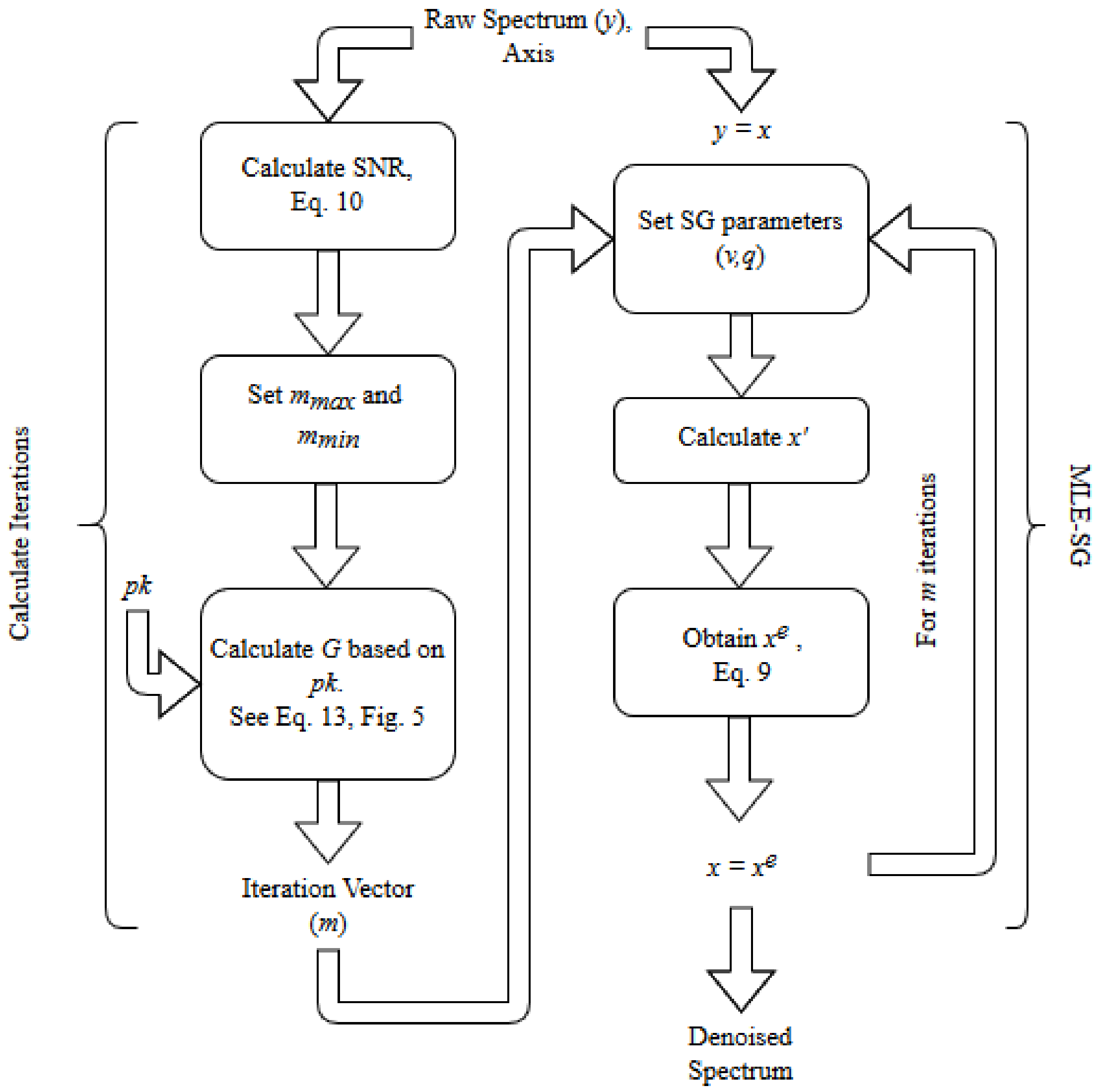


Figure 9.6: Algorithm flowchart

in the following section. This can be achieved by manually inputting a vector of wavenumber locations, which can be time consuming, or by defining a set number of locations and loading them automatically from a text file. In the case of known peak locations, this allows the algorithm to be applied as a single post-processing step for individual spectra or as part of a larger, automated process. If this is not the case, a peak for which the wavenumber location is not defined may be subjected to unnecessary smoothing. Two alternative approaches are discussed in Section 9.5.

9.4 Evaluation

The tuning of the algorithm described in Section 9.3 was performed with simulated datasets based on a spectrum recording from an Ibidi polymer slide. In order to test the performance of the algorithm on experimental datasets, spectra from the same slide were recorded using a confocal Raman microscopy system and the SNR of the recorded datasets was controlled by varying the acquisition time. A reference spectrum with low noise was collected using a long acquisition time and subtraction of a dark frame of equal acquisition time; the value of x^{ref} for a given dataset could then be calculated by scaling the reference spectrum appropriately in order to match the acquisition time used to record that dataset; this was done using an Extended Multiplicative Signal Correction (EMSC) algorithm. [38]

Both simulated and experimental data were processed using the MLE-SG algorithm with early stopping for peaks at the wavenumbers illustrated in Fig. 9.5 and the denoised spectrum was evaluated in terms of SNR, and compared with the other smoothing algorithms. The results of this analysis show that the SNR of the collected signals can potentially be doubled through applying the MLE-SG algorithm, which is advantageous in low light applications or in applications where cost or time constraints exist. Results from the experimental and simulated datasets were similar; however, experimental data had a slightly lower improvement in SNR results, which is to be expected due to experimental variability of the signal intensity and dark current both of which result in a variation in the SNR of the raw spectrum. Another possible cause of the slightly lower SNR improvement is the use of a Gaussian noise model, instead of the more accurate Poisson model, for the experimental noise. A qualitative comparison of signal recovery is illustrated in Fig. 9.7 where the quality of the denoised spectra by the relevant algorithms is illustrated.

In Fig. 9.7 it can be seen that the flat regions have a significantly lower standard deviation than that of the spectra that have been processed using the other techniques. However, this has had little to no effect on its capacity to preserve the characteristic features of the sharp peak, which has been highlighted in Fig.

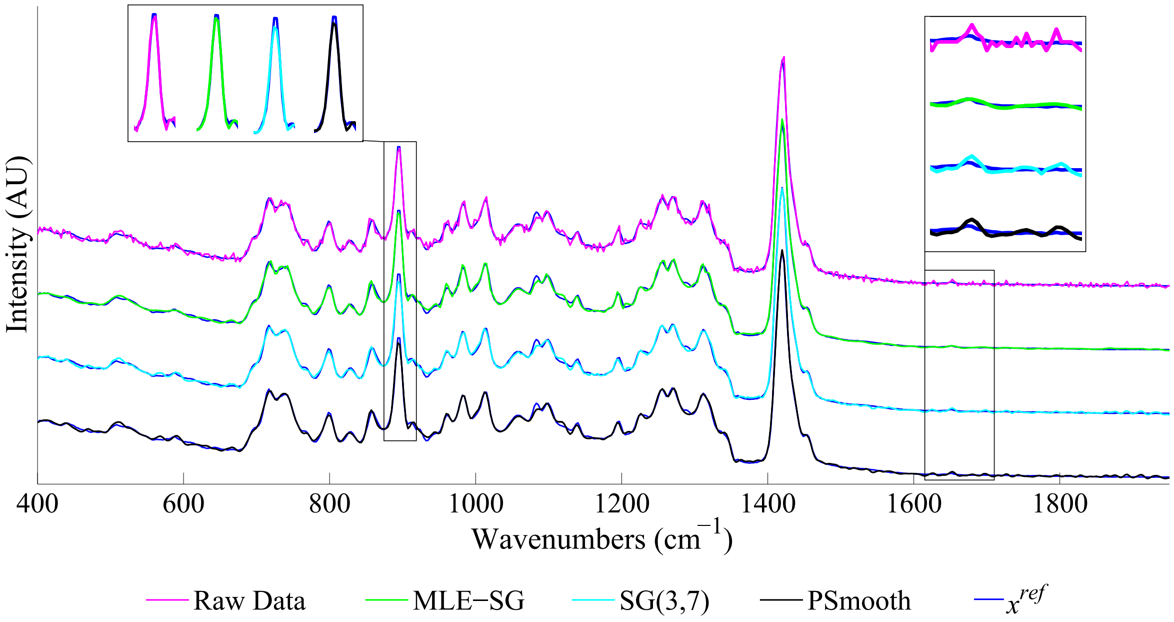


Figure 9.7: Qualitative comparison of signal recovery achieved by the three denoising algorithms on an experimental spectrum with an initial SNR of 97.

9.7. This is further demonstrated in Fig. 9.8, which compares the results in terms of the metrics previously discussed in Section 9.3.1. However, it is difficultly to visually appreciate a small improvement in the SNR of a signal; for example, a spectrum with a SNR of 90 may appear qualitatively similar to a spectrum with an SNR of 100. In order to provide a more rigorous quantitative evaluation, a set of tables that correspond to Fig. 9.8(a), Fig. 9.8(b), and Fig. 9.9 are given in Appendix B.

Although the experimental datasets were collected with the intention of having matching SNRs to the simulated datasets, this was not strictly possible and so the SNR range for the experimental data is from 33 – 141 approximately; however, the axes ranges have been kept the same for ease of comparison. While the results fluctuate more than the results for the experimental section, they show similar trends; in both cases pSmooth out-performs SG smoothing in terms of global SNR, which is to be expected, and MLE-SG shows the highest SNR improvement in all contexts. However, in the figures that describe the improvement in peak SNR it is clear that the MLE-SG algorithm is the only method that consistently and reliably improves the SNR within a peak region, and the SNR product further reflects this with significantly higher values for the MLE-SG algorithm than for the other methods, in all cases. The algorithms were also evaluated in terms of computational efficiency. The algorithms were implemented using MATLAB running on a Dell Inspiron 15 with an Intel Core i7 processor. The average time taken for MLE-SG, SG, and PSmooth was 195ms, 0.8ms, and 34.8ms, respectively. As expected, SG smoothing provides the fastest implementation.

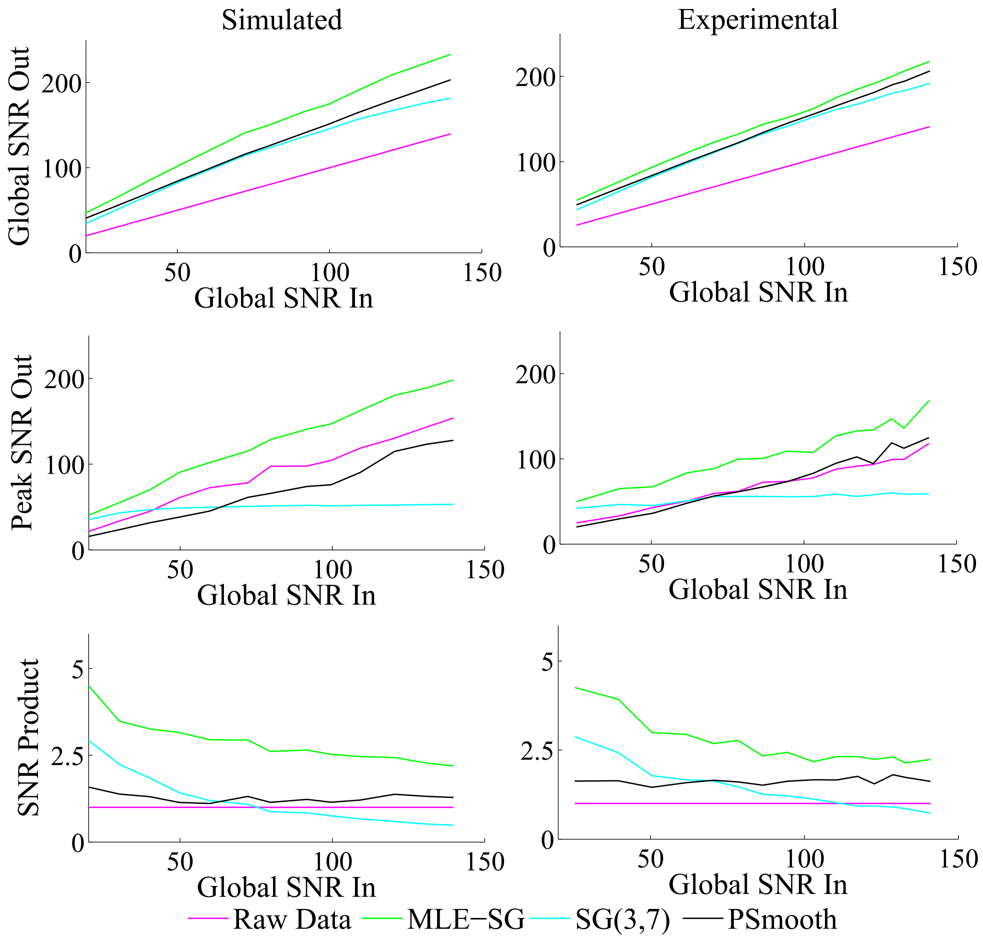


Figure 9.8: Comparison of SNR enhancement achieved by denoising algorithms for simulated and experimental datasets of similar SNRs. Corresponding table of values is available in Appendix B, see Tables B.1 and B.2.

9.4.1 Application to Biological Spectra

It may be difficult to record a reliable and low-noise reference Raman spectrum from a biological sample that could subsequently be used in an accurate quantitative evaluation of improvements in SNR for a denoised spectrum. An accurate representation of the irradiance would require a long exposure time on one sample point and would likely result in photo-bleaching/damage. In addition, biological cells are often biochemically heterogeneous at different locations in a single cell, as well as across a group of similar cells; such a heterogeneity presents an additional complexity in terms of finding an accurate reference spectrum that could be used in a quantitative assessment of SNR over a dataset. Therefore, it was decided to test the algorithm on a simulated dataset based on one high quality cell spectrum that is artificially noised; in this case the reference spectrum is the original cell spectrum before the addition of noise. Considering the similarity in the results between the experimental and simulated polymer datasets, it was inferred that the simulated cell spectra would provide a suitable representation of the algorithm's capabilities for this application. A low noise reference spectrum was generated by adding together more than fifty spectra recorded from a high grade bladder cancer cell line, following formalin fixation; cell preparation, recording, and appropriate processing methods. [75] The end result is a reference spectrum of 500s acquisition time recorded from the nucleus of 50 cells from this cell line using a $120mW$ $785nm$ laser. This low noise reference was then artificially noised as described in Section 9.3, and 17 datasets were generated with SNR values from 20 to 200 in steps of 10, each containing 100 spectra. Following this, the spectra were denoised using the MLE-SG algorithm with early stopping at appropriate wavenumber locations and the average improvement in SNR was measured for each dataset. Among the marked peak locations (in cm^{-1}) are: $785, 1004, 1090, 1127, 1262, 1319, 1341, 1451, 1585, 1619$, and 1662. These peak number locations correspond to well known biochemical assignments in epithelial cells, as shown in Table 9.1. [108, 109]

The standard deviation of the Gaussian modelling the peak regions was kept at 10, as in the previous section. The results, together with corresponding results for the PSmooth algorithm, and SG filtering (with polynomial and window sizes of 3 and 7) respectively, are shown in Fig. 9.9. In almost all cases the MLE-SG algorithm outperforms the other algorithms, with a minimum of 50% improvement in SNR compared to the raw data. For the low SNR case pSmooth shows a comparable result to MLE-SG for the global spectrum due to the higher amount of smoothing generated by that algorithm; however, the improvement in the SNR in the region of the phenylalanine peak is significantly higher for MLE-SG.

Similar results are observed to those in the previous section. The peaks have been effectively preserved

Wavenumber (cm^{-1})	Chemical Bond	Association
785 - 788	Stretching of DNA related bonds and DNA/RNA breathing modes	Nucleic Acid
1004	Phenylalanine	Protein
1090	Stretching of DNA related bonds Stretching of C-N backbone	Nucleic Acid Protein
1127	Stretching of C-N backbone Stretching of C-C	Protein Lipid
1262	DNA/RNA breathing modes Amide III	Nucleic Acid Lipid
1319	CH2, CH3 twisting DNA/RNA breathing modes CH deformation vibration	Lipid Nucleic Acid Protein
1341	DNA/RNA breathing modes CH deformation vibration	Nucleic Acid Protein
1451	CH_2 deformation vibration	Protein/Lipid
1585	DNA/RNA breathing modes	Nucleic Acid
1619	Tyrosine; tryptophan	Protein
1662	DNA/RNA breathing modes Amide I Fatty Acids	Nucleic Acid Protein Lipid

Table 9.1: A table of common biochemical assignments in epithelial cells.

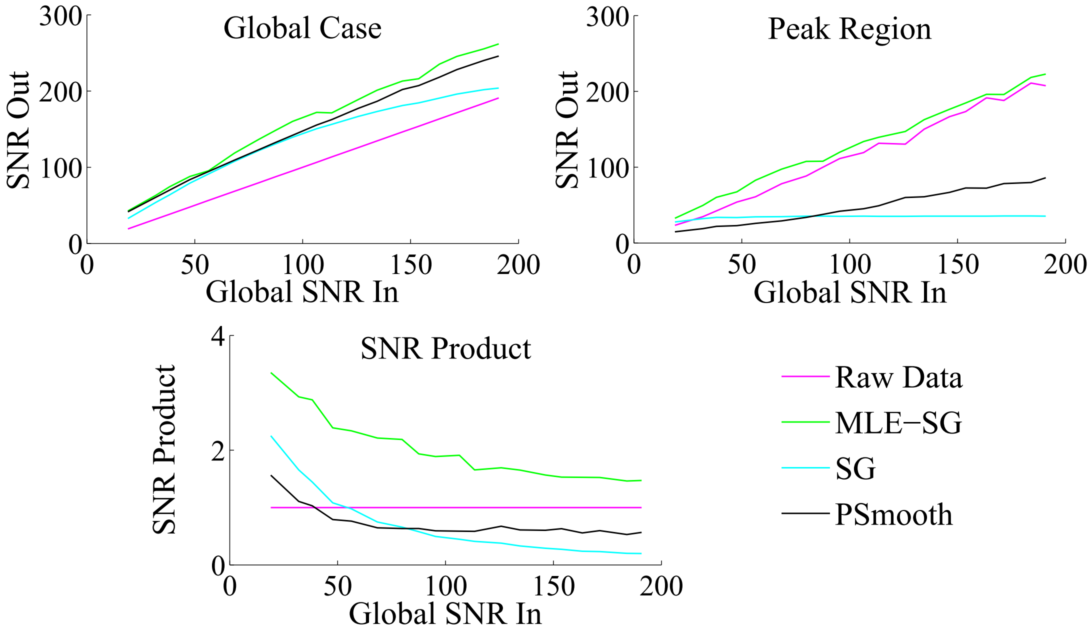


Figure 9.9: Comparison of SNR enhancement for simulated T24 datasets achieved by denoising algorithms. A corresponding table of values is available in Appendix B, see Table B.3.

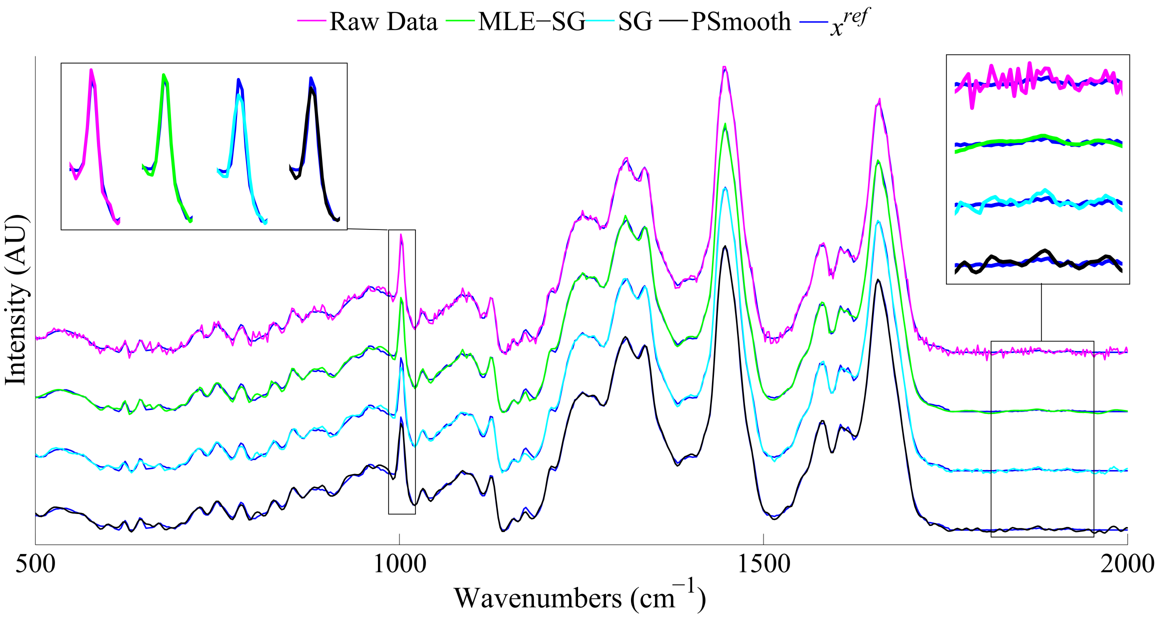


Figure 9.10: Comparison of SNR enhancement achieved by denoising algorithms for an initial SNR of 50.

and the large flat regions have a lower standard deviation than that provided by the other two denoising algorithms. In all cases the MLE-SG algorithm preserves the peaks better than the other two methods, while also out-performing the other methods in terms of global smoothing. The SNR product clearly demonstrates the superiority of the algorithm for all cases of input SNR by taking into account both global and peak SNR improvement in a single metric. Despite the tuning of the algorithm using the polymer spectrum, which has a significantly different spectral form the algorithm still provides a superior SNR enhancement over the other algorithms and produces high quality denoised spectra. This indicates that the algorithm performs robustly across different types of spectra. This does not preclude the possibility of improving performance through spectra-specific tuning.

9.5 Summary and Future Work

This chapter has demonstrated how Savitzky-Golay filtering may be enhanced with Maximum Likelihood Estimation to produce an algorithm that consistently out-performs competing algorithms. MLE provides bounding properties, based on the noise distribution associated with the signal, to prevent the SG smoother from significantly altering the underlying spectral features. The algorithm is iterative, with increased smoothing occurring with each iteration; inclusion of an early stopping procedure, based on a user input of peak locations, further ensures that sharp local features are effectively preserved while allowing further smoothing in low frequency regions. The resulting algorithm provides up to a 100% improvement in SNR when compared to the raw data. It also consistently out-performing competing algorithms (PSmooth and SG filtering) in terms of all metrics used to evaluate algorithm performance. While inputting all peak locations of interest may not always be possible, particularly if unexpected components exist, many applications are based on recording a known sample repeatedly, therefore, the features of interest are generally well known and will require the same wavenumber locations for subsequent experiments.

The proposed algorithm was optimised and rigorously tested on simulated datasets based on a polymer spectrum before being tested on experimentally collected datasets; a close correspondence was observed in the results for the simulated and experimental datasets. Finally the algorithm was tested on simulated datasets of epithelial cells and the results showed similar trends in SNR improvement despite there being no retuning of the algorithm.

Another contribution of this chapter is the development of a rigorous approach to evaluate Raman smoothing algorithms in general in terms of SNR and the proposed SNR product metric. This analysis is based on estimating the RMSE with respect to a known reference and is applied to both the global spectrum, as well as in a peak region; since these regions can be adversely affected by smoothing. The proposed metric known as the SNR product, i.e. the product of the improvement in global SNR multiplied by the improvement in peak SNR, is used to monitor the overall spectral quality provided by the denoising algorithms. This allows the user to evaluate how effectively the algorithm preserves peaks and smooths low frequency regions simultaneously. We believe that this chapter constitutes the first attempt to rigorously investigate the effects of smoothing algorithms on Raman spectra in terms of SNR.

Recently, a blind deconvolution algorithm has been proposed that appears to have some similarities to the denoising method presented here [110, 111]. Their method also makes use of the maximising *a posteriori* technique in an iterative manner, and uses a modified Tikhonov regularization model that appears to be similar to the constraint used in our approach that penalises deviation from neighbouring values. Their method also includes a deconvolution process during each iteration in order to take into account, and correct for, the system response function, which is also varied with each iteration. This approach is demonstrated to recover highly degraded and noisy Raman spectra, particularly for cases in which spectral structure is corrupted due to the instrument response. Although this blind deconvolution algorithm has some similarities to the proposed method, both are derived in fundamentally different ways, and each has its own unique characteristics. More work is needed to fully elucidate the relationship between the two algorithms and to compare their results.

Future work could also include varying the value of λ in Equation 9.9 for each sample in the spectrum in a similar manner to that of the adaptive regularisation model discussed in the previous paragraph. [110,111]

By varying this parameter rather than *m* it may be possible to produce a result in fewer iterations. An obvious area for future work is to develop an automatic peak identifier [112] in the operation of the algorithm to negate the need for user input; this would also have the benefit of including unexpected spectral peaks in the early stopping process.

Once the proposed software methodologies and algorithms, that have been discussed in the preceding chapters, have been applied to the system the principle option for improving SNR is the use of reflective substrates. Chapter 10 proposes the use of reflective substrates in Raman systems, particularly systems that use Near Infrared (NIR) lasers, in order to increase the SNR of the collected spectra and/or to improve the system throughput.

Chapter 10

Improving SNR using reflective substrates

This chapter is related to the following manuscript that is in preparation for submission to a journal:

Sinéad Barton and Bryan Hennelly. "Improving Signal to Noise Ratio Using Reflective Substrates". In Preparation. Analytical Methods.

I would like to provide an acknowledgement to Adam Dignam and Marion Butler of the Immunology Department at Maynooth University for the cultivation and deposition of the cell lines presented in this paper, Section 10.3.1.

10.1 Introduction

This chapter is the only chapter that proposes a solution to the issue of improving SNR that does not make use of software. Assuming the optical elements in the Raman spectrometer are well aligned, the aforementioned optimisation methodologies have been applied, and optimal denoising techniques are being utilised, the substrate remains the principal avenue of investigation for improving the SNR of the recorded spectra; the reason for this is not immediately clear, but by the end of this chapter it will be obvious that the substrate can generate noise that exceeds any other source of noise in the Raman spectrum of a biological cell. Investigating the use of reflective substrates is not only important for improving the SNR but also for reducing the cost of biological experimentation where NIR lasers are commonly used but require specialised high cost calcium fluoride substrates. Unlike previous chapters, no artificial data is used in this chapter;

however, the SNR of the spectra from various substrates is used as a metric to quantify the improvement in spectral quality. The reflective substrate is shown to significantly improve the SNR due to the increase in collected Raman scatter, when compared with more expensive calcium fluoride (CaF2) slides.

Excitement of Raman scattering using a laser with a wavelength in the NIR wavelength region is often preferred to laser in the visible region for applications that require high accuracy. This is due to a number of key advantages of NIR excitation including higher wavenumber resolution, reduced photo-damage, as well as reduced auto-fluorescence from the spectrum. However, the dominant spectrum from glass substrates under NIR excitation necessitates the use of highly expensive pure crystal substrates, such as Raman grade CaF2. These substrates produce little background with a caveat that they place a cost limitation on this approach. [12] Furthermore, reduced Raman scattering occurs for NIR excitation when compared with excitation with lower wavelengths resulting in lower SNRs, which necessitates longer acquisition times. [12]

In this chapter, the use of reflective substrates for Raman cytology using NIR excitation is investigated as an approach to overcome the two disadvantages mentioned in the previous paragraph. These substrates are made from traditional glass slides with a thin film of metal deposited on the surface, such as aluminium or gold, and are relatively inexpensive. The authors have previously demonstrated the potential for reflective substrates for NIR Raman cytology [12]; here, we expand on this work to rigorously investigate the advantage that these substrates afford in terms of SNR when compared with Raman grade CaF2. Multivariate classification of two prostate cancer cell lines is applied to cells recorded on both types of substrates and it is shown that the reflective substrates display enhanced performance.

Finally, despite the dominant background spectrum associated with a glass slide, it is shown in this paper that modern background subtraction methods [42], can still recover the Raman spectrum of a cell. Although the recovered spectrum has a very low SNR, it is demonstrated that classification is still possible using the PCA loadings generated from the previous multivariate analysis with the gold and CaF2 substrates. It is believed that it has been shown possible to record a reliable Raman spectrum from a cell deposited on glass using NIR excitation.

10.2 Confocal Raman Microspectroscopy Using NIR Excitation

Excitation of the Raman spectrum using a source laser with a wavelength in the near infrared (NIR) has a number of key advantages for analysing biological samples when compared with visible wavelength excitation: (i) Firstly, NIR wavelengths are less absorbent in tissue and cells and are, therefore, less likely to burn or photo-damage the sample, (ii) secondly, biological samples are known to produce little or no auto-fluorescence with NIR lasers [113], which can be a significant advantage [114], and thirdly (iii) the wavenumber resolution is superior in a Raman spectrum scattered using a NIR wavelength.

The baseline resulting from auto-fluorescence often varies randomly from one recording to the next. Although we have just now ascribed this baseline to an auto-fluorescence from the sample itself, as it is the most common explanation for the source of this disturbance [113], it must be noted that other authors [114] have suggested that it may originate from sample morphology and Mie scattering of the source laser wavelength. Regardless of its origins, it is well known that this baseline is significantly less pronounced for Raman spectroscopy with NIR wavelength excitation. Various algorithms have been developed to identify and remove the baseline signal from Raman cell spectra, with polynomial fitting techniques being the most common technique used today. [38, 115] Recently, least squares algorithms have been shown to provide superior results. [42, 43] In Raman spectroscopy, where units such as Raman shift or wavenumbers are used, the spectral resolution (in wavenumbers) of the system increases with the source wavelength. When the wavelength is shifted through Raman spectroscopy from the excitation wavelength λ_e to the scattered wavelength λ_s, the shift in wavelength is given by $\Delta\lambda = \lambda_e - \lambda_s$ but the corresponding wavenumber shift $(\Delta\widetilde{v})$ is given by:

$$\Delta\widetilde{v}(cm^{-1}) \quad = \quad \left[\frac{1}{\lambda_c(nm)} - \frac{1}{\lambda_s(nm)}\right] \times \frac{10^7(nm)}{(cm)} \tag{10.1}$$

Therefore, assuming that the wavelength bandwidth, $\Delta\lambda$, remains constant regardless of the center wavelength chosen, it follows that the ratio of spectral bandwidths for two different source wavelengths, λ_{e1} and λ_{e2}, will be given by:

$$\frac{\Delta\widetilde{v}_1}{\Delta\widetilde{v}_2} = \frac{\lambda_{c2}(\lambda_{c2} - \Delta\lambda)}{\lambda_{c1}(\lambda_{c1} - \Delta\lambda)} \tag{10.2}$$

Thus, for a sample case of $\Delta\lambda$ = 100 *nm*, we can conclude that 532 *nm* source wavelength results in approximately 2.3 times more spectral bandwidth than 785 *nm*. Conversely, the resolution at 785 *nm* is ≈2.3 times smaller than that at 532*nm*. The overall efficiency of a grating also depends on the blaze angle of the grating, which we do not consider here.

However, despite the advantages discussed above, there also exist a number of disadvantages associated with NIR excitation; (i) Firstly, and most significantly, the spectrum from a glass substrate has a large intensity that overwhelms the cell spectrum. This necessitates the use of expensive, highly pure Calcium Fluoride (CaF2) substrates (or similar), which are expensive (>$100 per slide), and are, therefore, not useful for high throughput applications. (ii) Secondly, the rate of Raman scattering is significantly reduced for NIR excitation, which results in lower SNR values.

The number of photons scattered is also related to the laser wavelength, with the intensity of Raman lines being proportional to the fourth power of the laser frequency [116]: Therefore, when a comparison is made between a 532 *nm* laser and a 785 *nm* laser, as in Equation 4.10, it can be seen that the 532 *nm* laser produces Raman lines that are approximately 4.74 times more intense than those produced by an 785 *nm* laser for the same laser powers, assuming non-resonant conditions. [115] It is clear that the scattering efficiency is higher at lower wavelengths, resulting in the use of shorter integration times and lower powered lasers in the blue/green regions, although the quantum efficiency of the grating and CCD detector being used must also be taken into consideration.

Despite the drawbacks listed above, NIR Raman spectroscopy is often preferred for accurate biochemical cellular analysis due the aforementioned advantages, provided that cost is not a limiting factor. In the following sections it is demonstrated that using a low cost reflective substrates can significantly enhance the intensity of a cell spectrum recorded using NIR excitation, thereby mitigating the effect of both disadvantages outlined in the previous paragraph.

10.3 Methods

10.3.1 Sample Preparation

High grade prostate cancer epithelium cells (PC3; Sigma-Aldrich) and androgen-sensitive human prostate adenocarcinoma (LNCaP, Sigma-Aldrich), were cultured in 1:1 mixture of DMEM and Hams-F12 medium supplemented with 5% fetal bovine serum and 2 ml Glutamine. Flasks were maintained in a humidified atmosphere with 5% CO2 at 37°C. When the cell lines reached 80% confluency, the culture medium was removed, and the cells were rinsed with sterile PBS. Trypsin-EDTA (0.5%) was added to the flask, which was incubated at 37°C until the cells had completely detached (not exceeding 15 min). An equal volume of 5% serum-containing medium was added to the flask to neutralise the trypsin enzyme. The entire contents of the flask was transferred into a sterile container, and centrifuged at 1200 rpm for 5 min. The supernatant was removed, and the cell pellet was resuspended in fresh medium. This solution was centrifuged at 1200 rpm for 5 min, the medium decanted; and resuspended in 1 ml PBS. This step was repeated and the cell pellets were resuspended into a vial containing 20 ml of a methanol based fixative (PreservCyt; Hologic, USA), and left at room temperature for 15 min. The vial was inserted into the ThinPrep 2000 (T2; Hologic, USA) machine, and the cells were transferred on to a gold coated glass slide (100 nm gold thin film on glass; Deposition Research Laboratory Inc., USA)), a CaF2 (Raman Grade; Crystran, UK) slide, or a glass slide (Thinprep slide; Hologic, USA). The Thinprep standard has previously been shown to be compatible with Raman microspectroscopy. [117, 118]

10.3.2 Acquisition of Data

All spectra were recorded using a HORIBA Jobin Yvon HR 800 (Villeneuve d'Ascq, France) Raman spectrometer, which was coupled to an Olympus BX41 upright microscope equipped with a 100x objective (MPlanN, Olympus, NA = 0.9) and a 785.16 nm diode laser source (300 mW). Raman scattering was collected through a 400 μm confocal hole onto a back illuminated air-cooled CCD detector with 13.5 μm pixel size (Synapse; Horiba, Villeneuve d'Ascq, France) for the range of 500–1800 cm−1 using a 300 lines mm−1 diffraction grating, yielding a dispersion of ~1.5 cm−1 per CCD pixel. The instrument was calibrated using the 520.7 cm−1 peak of silicon. Raman spectra were recorded using 20 s acquisitions. In total, spectra were recorded from 52 PC3 cells on CaF2, 52 LNCap cells on CaF2, 52 PC3 cells on gold, 52 LNCap cells on gold, and finally 52 LNCap cells on glass.

10.3.3 Processing of Data

Following recording of the raw spectra CRA removal was performed using the method described in Chapter 8. Cosmic ray removal is followed by application of an EMSC algorithm as described in detail in Chapter 3. This algorithm computes a background signal in the form of a baseline N-order polynomial (to remove the baseline signal that results from the cells auto-fluorescence) plus the background signal from the substrate. Briefly described, the EMSC algorithm applies a least squares fit to (i) a reference Raman spectrum from a cell; (ii) an N-order polynomial; and (iii) a reference spectrum taken from the substrate. The algorithm returns the weight of (i), which enables normalisation of the spectrum relative to the reference, as well as the total background made up of the appropriately weighted substrate spectrum plus the polynomial. It has been shown that the use of high values N does not result in over-fitting with the EMSC algorithm. [43] For this study, a 7th order polynomial was used in the EMSC subtraction algorithm for all datasets, and similar results were found for N=3.

The reference spectrum of the cell provides the basis for all of the spectra to be fitted; the reference spectrum chosen in this study is the mean spectrum of all of the raw spectra recorded from all cells on both the gold and CaF2 substrates. In order to remove any potential bias, the same reference spectrum was used for the EMSC algorithm applied to process the spectra of all cells that are investigated in this paper, including those recorded on glass. The background signal from both the gold and CaF2 substrates are flat in the fingerprint region [108] and, therefore, no reference spectrum from the substrate is provided to the EMSC algorithm when processing spectra recorded from cells on those substrates. However, for the spectra recorded from glass, a reference glass spectrum was also input to EMSC algorithm to be included in the least squares fit in addition to the cell reference and polynomial. This reference glass spectrum was calculated based on the mean of 6 acquisitions of 20 seconds recorded from the Thinprep slide. [45–48] Multivariate statistical analysis is often applied to Raman spectroscopic data for classification. This involves the application of pattern recognition techniques, such as PCA or LDA, in order to identify subtle changes across datasets that can be used to accurately differentiate between different pathological groups and subgroups. PCA was directly applied using the 'pca' function in MATLAB to the four datasets recorded from cells on gold and CaF2. In advance of PCA, all spectra were smoothed with a Savitzky-Golay filter [11] (k = 3; w = 11) in order to reduce noise.

10.4 Results

10.4.1 CaF2 vs. Gold

The raw spectra recorded from the PC3 and LNCap cell lines on both gold and CaF2 are shown in Fig. 10.1 (i) and (ii) following cosmic ray removal. It should be noted that no scaling of any kind has been applied; the spectra shown in Fig. 10.1 (i) and (ii) are shown on the same axis and with the same relative intensity as when they were recorded. No shifting of the spectra along the intensity axis has been performed. On average, the spectral intensities of the cell spectra recorded from both cell lines on the gold is similar and is approximately four times that of the spectral intensities recorded on the CaF2.

The corresponding datasets after EMSC correction are shown in Fig. 10.1 (iii) and (iv). The EMSC algorithm uses a 7^{th} order polynomial at the mean spectrum of all of the raw spectra shown in Fig. 10.1 (i) and (ii) as a reference cell spectrum. The resulting datasets have been corrected relative to this reference spectrum; a 7^{th} order polynomial has been subtracted and the result is normalised with respect to the refer-

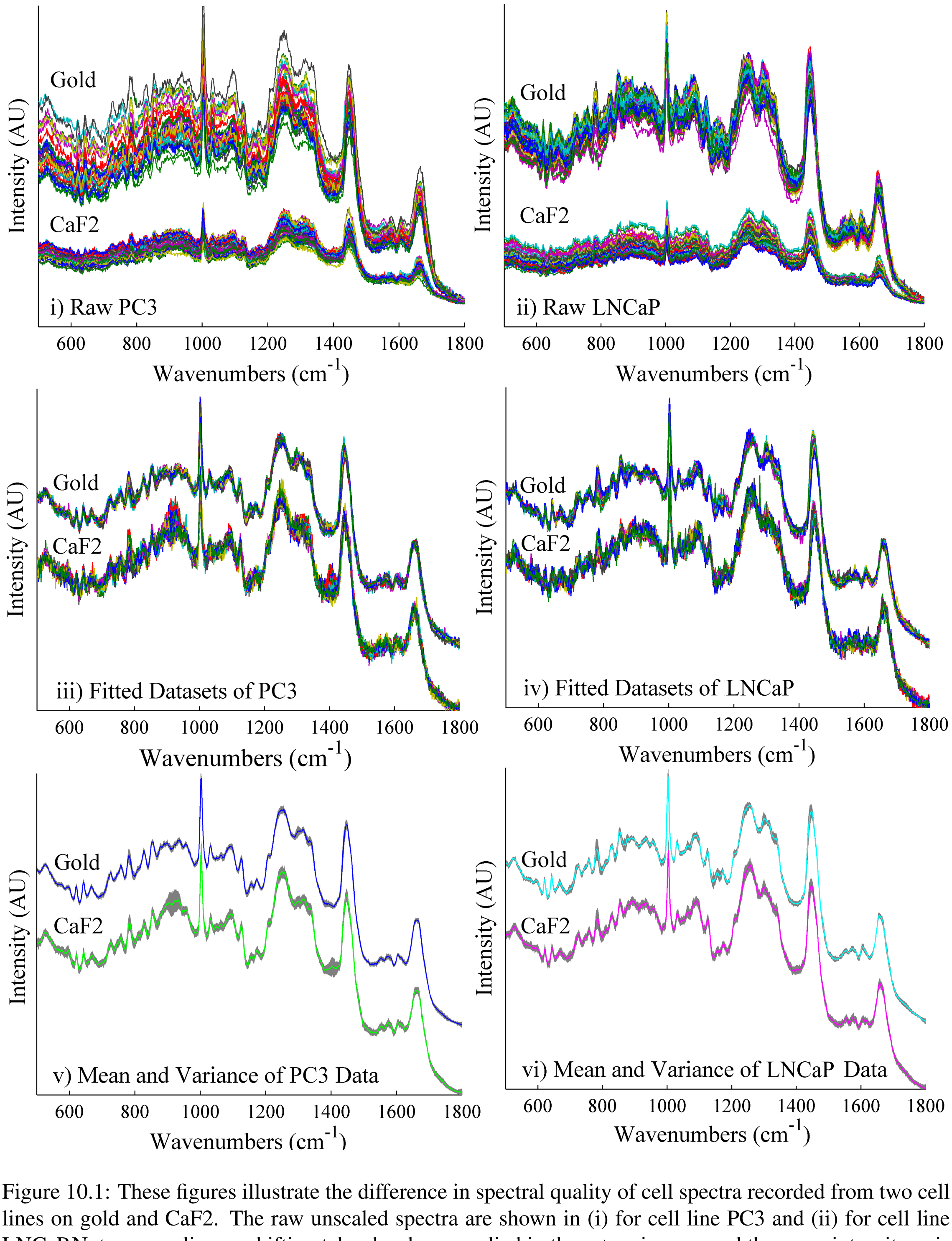


Figure 10.1: These figures illustrate the difference in spectral quality of cell spectra recorded from two cell lines on gold and CaF2. The raw unscaled spectra are shown in (i) for cell line PC3 and (ii) for cell line LNCaP.Note, no scaling or shifting takes has been applied in these two images and the same intensity axis is used for all data in (i) and (ii). In (iii) and (iv) the EMSC corrected datasets are shown and the two groups are shifted in the intensity axis for ease of comparison. In (v) and (vi) the mean spectrum and variance are shown for each of the four datasets shown in the previous two figures.

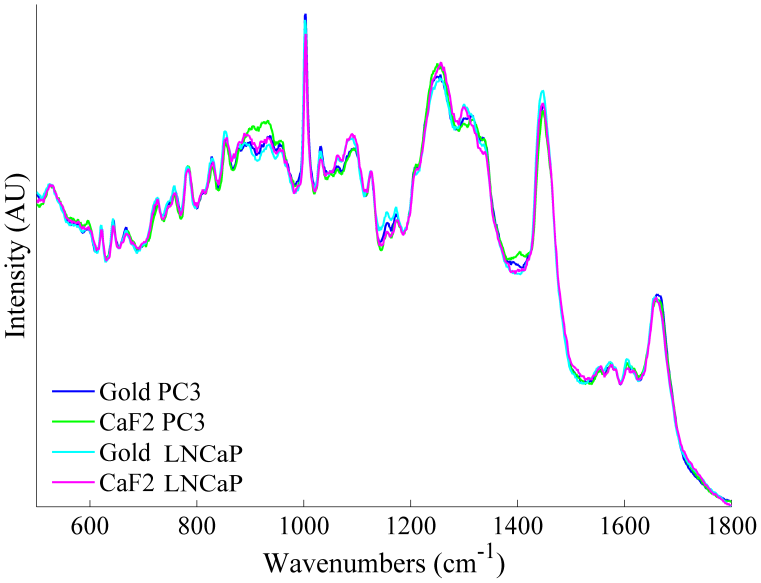


Figure 10.2: The four mean spectra from the EMSC corrected data shown in Fig. 10.1 (iii) and (iv).

ence spectrum. These figures show that there exists significantly less variance across the datasets recorded from the gold substrate, particular in the wavenumber band around 900 cm^{-1} and in the band 1200 cm^{-1} – 1400 cm^{-1}. For ease of comparison the four mean spectra are overlayed in Fig. 10.2, in which it can be seen that there is a strong similarity between the four mean spectra, with the greatest differences occurring in the two regions mentioned earlier in which high variance is observed for the cell spectra recorded from the CaF2 substrate.

Fig. 10.1 (i) and (ii) illustrates that the spectral intensities of the two cell lines varied on average by approximately a factor of three across the gold and CaF2 substrates. The possible reasons for this are discussed in in the next section. Here, we investigate further this difference in spectral intensity by analyzing the SNR of the spectra recorded from both substrates. In order to do this, the method for estimating SNR, described in Section 4.3.1, is employed. We recall that the basis for this method is simply to subtract a SG smoothed version (polynomial order = 3, window size = 9) of each spectrum from its raw counterpart in order to isolate the noise signal, from which the standard deviation is calculated. The signal value, S, is taken as the maximum point in the mean spectrum used to EMSC fit the data. The SNR is calculated from each of the raw spectra shown in Fig. 10.1 (i) and (ii), using this method. The results are shown using a box and whisker plot in Fig. 10.3.

It is clear that there is a significant difference across the two substrates in terms of SNR. The mean SNR for the PC3 and LNCaP cell spectra were 141.2 and 146.4 for the gold substrate, respectively. The corresponding SNR values for the CaF2 substrate were 88.4 and 86.2. The gold substrate shows a consistently higher SNR than that of their CaF2 counterparts as well as a more narrow grouping of SNR values, although

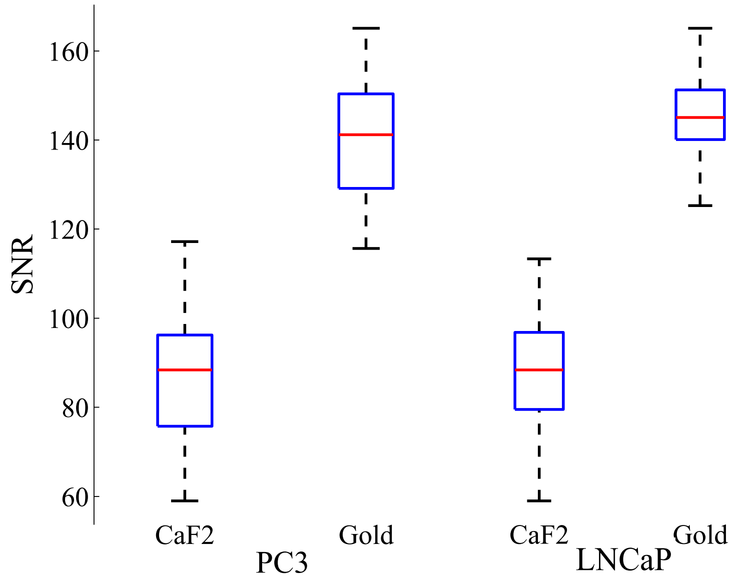


Figure 10.3: A comparative boxplot of SNR values for both cell lines across CaF2 and gold substrates.

in the case of PC3 the increase in SNR clustering is less obvious than that of LNCaP.

There are a number of interesting conclusions that can be made from this result. Most notably, the mean SNR value of all the spectra recorded from the gold is 143.8 and the mean SNR from CaF2 is 87.3. The ratio of these to values is 1.65, which is approximately the square root of the ratio of the mean intensities of the raw spectral data recorded from both substrates, shown in Fig. 10.1 (i) and (ii). This result is consistent with what would be expected given the definition of shot noise given in 4.2.1.

Here we apply PCA, as described in Section 10.3.3, only in order to investigate (i) the differences between the spectra recorded across both substrates as well as across both cell lines and (ii) any differences in the separation of cell groups across the two substrates. The results of this analysis are shown in Fig. 10.4 and Fig. 10.5 where we have used the same colour codes for the four cell groups that appeared in Fig.10.1 and elsewhere in this chapter. Illustrated are the principal components for which there was a clear separation of the corresponding score plots. In Fig. 10.4 i), the mean spectra of PC3 and LNCaP deposited on gold are shown as well as the first principal component, ii) shows the scatter plot that is related to that analysis. Similar results are shown for both cell lines when deposited on CaF2 in Fig. 10.4 iii) and iv). However, in this case the scores are separated principal components 1 and 3. Prominent peaks in the relevant loadings have been highlighted for ease of comparison. A close correlation is observed across these peaks for both cases. These encompass numerous spectral peaks that have previously been noted in the classification of prostate and bladder cells, with key biomolecular peak differences observed at 792 (DNA), 938 (proteins), 1002 (Phenylalanine; Protein), 1058 (DNA, lipids), 1089 (DNA), 1252 (Amide III), 1302 (CH2; lipids) 1340 (CH2/CH3 wagging of nucleic acids), 1459 (DNA), 1484—1574 (DNA), and 1676 cm−1 (Amide

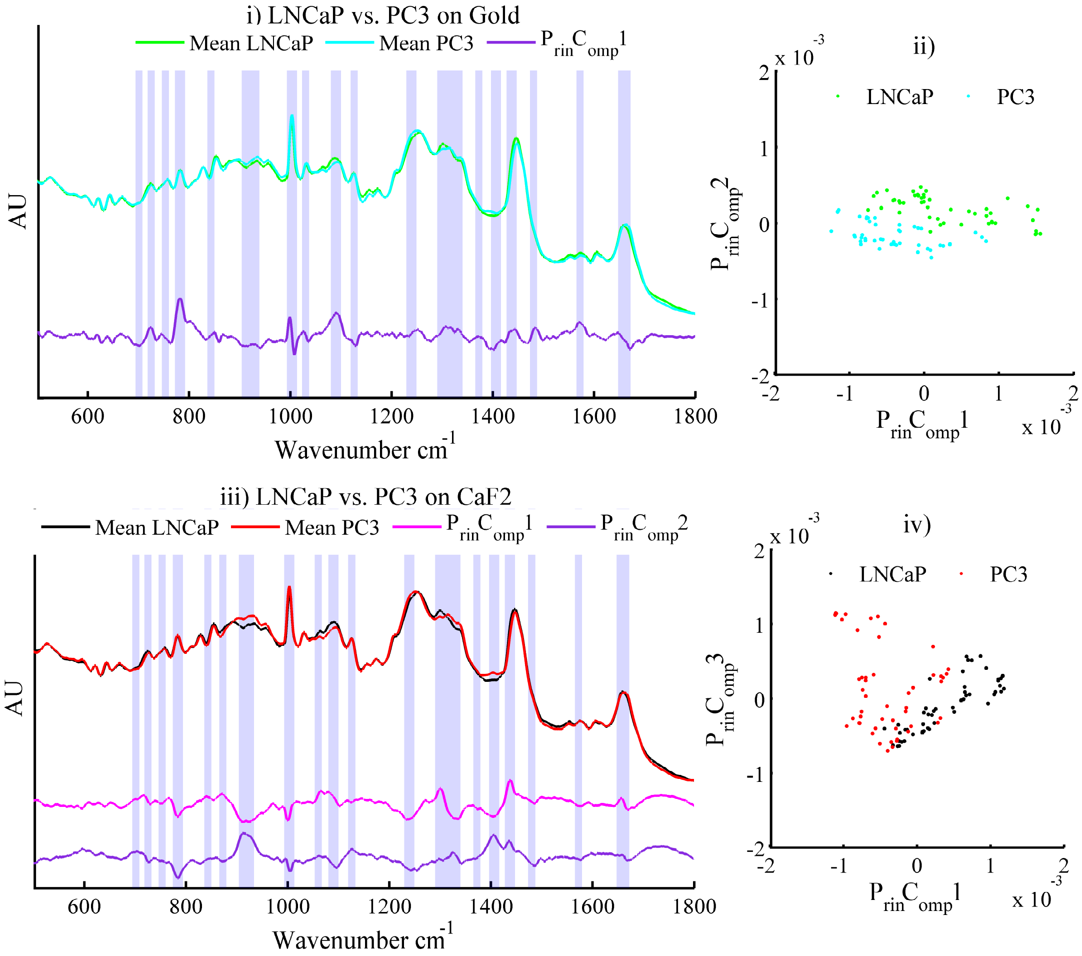


Figure 10.4: (i) Illustration of mean LNCaP cell spectra on gold and CaF2 compared to the PCA loading providing maximum separation (ii) Corresponding scatter plot iii) Mean spectra of PC3 deposited on gold and CaF2 when compared to the PCA loadings that provide maximum separation iv) Corresponding scatter plot

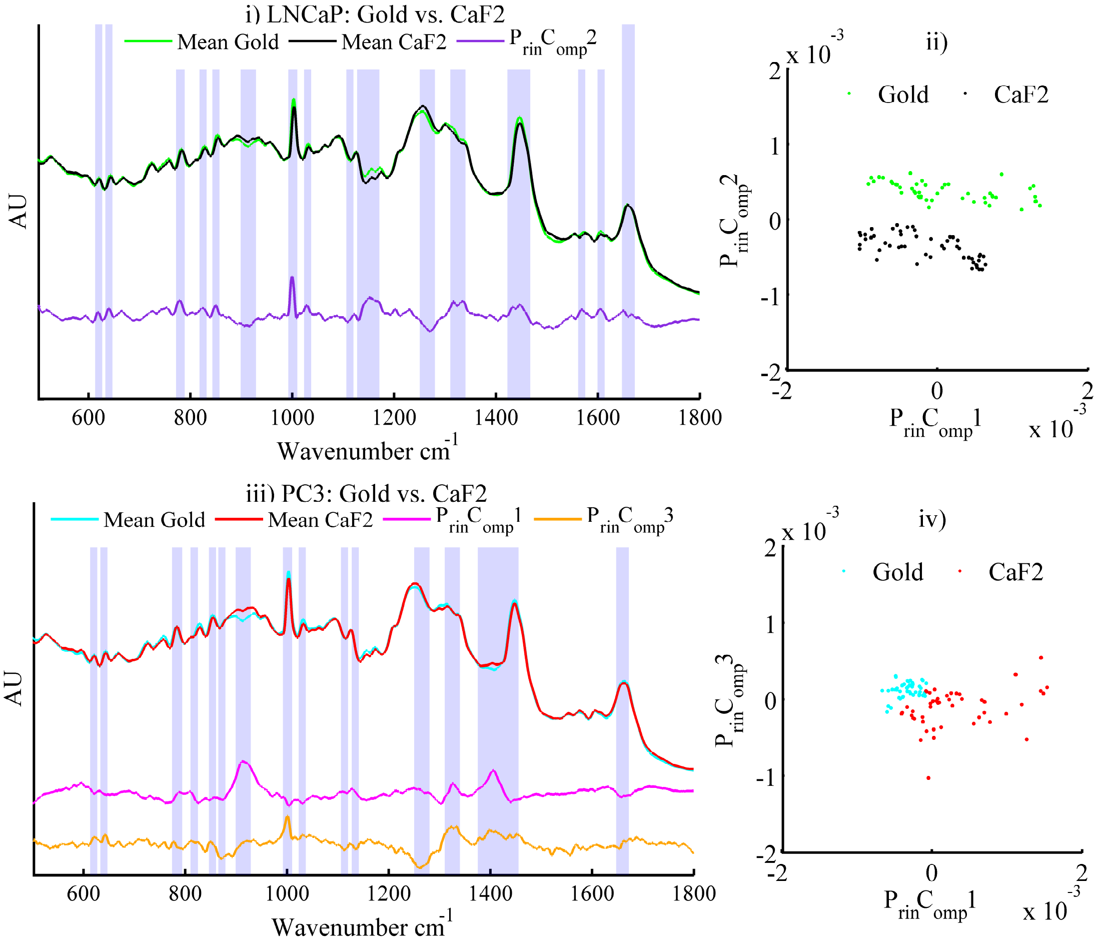


Figure 10.5: (i) Illustration of mean LNCaP and PC3 cell spectra on Gold compared to the PCA loading providing maximum separation (ii) Corresponding scatter plot iii) Mean spectra of LNCaP and PC3 cells deposited on CaF2 when compared to the PCA loadings that provide maximum separation iv) Corresponding scatter plot

I); similar peak differences have previously been observed in the separation of prostate and urothelial cell lines. [42, 68, 119, 120]

We infer from these results that the same biomolecular differences account for the separation of both cell lines regardless of substrate. However, the important question of why different results are observed across both substrates remains. In an effort to elucidate the reason(s) for this, PCA was applied to the same cell line across both substrates in isolation, as illustrated in Fig. 10.5.

There appears to be no distinct relationship that accounts for the separation of LNCaP across both substrates with the corresponding separation of PC3. It may be possible that the substrates affect the cells in different ways. More work is required to fully understand this phenomenon.

10.4.2 Glass Substrates

Raman spectra were recorded from the nuclei of 52 LNCaP cells deposited on a glass ThinPrep slide. The original purpose of this experiment was to compare the raw spectra with equivalent results from CaF2 and gold in order to validate the well-known assumption that the glass signal dominates the cell spectrum to such a degree that it renders these slides unusable for NIR Raman spectroscopy. The raw spectra of the LNCaP cells recorded from glass are shown in Fig. 10.6 (i), where it can be seen that the glass spectrum dominates the cell spectrum. The only obvious visible feature from the cell spectrum is the phenylalanine peak at 1000 cm^{-1}.

It must be noted that the cell spectra are not weak, they are approximately equal in intensity to the cell spectra recorded from the CaF2 substrate. Rather, the cell spectrum is weak relative to the glass spectrum. It should be noted that a number (12) of the glass LNCaP spectra, which are not shown here, were sufficiently intense as to saturate the CCD sensor. An attempt was made to recover the cell spectrum from the raw dataset in Fig. 10.6 i) by applying EMSC correction to remove the glass spectrum, as described in Section 10.3.3 and 3.4.

The corrected dataset is shown in Fig. 10.6 ii) where it can be seen that all of the principal peaks associated with the Raman spectrum of a cell are indeed present. However, it is clear that there is a significantly higher level of noise in this dataset than in the corresponding datasets for the gold and CaF2 case shown in Fig. 10.1 iv), which is confirmed by calculating the mean corrected spectrum and the variance of the dataset, shown in Fig. 10.6 iv). The reason for this is due to residual shot noise introduced by the glass spectrum that cannot be removed using the EMSC algorithm. This noise will have a standard deviation that is equal to the square root of the intensity of the glass spectrum and will, therefore, be appreciably large when compared with the intensity of the cell spectrum. If this reasoning is extended, it can be expected that that this residual noise term will be strongest in the region in which the glass spectrum is most intense. Two estimates of SNR were calculated independently, one corresponding to the left hand side (LHS) of the spectrum (everything before 1150 cm^{-1}) and the second corresponding to the right hand side (RHS) of the spectrum as indicated in Fig 10.6 iii). The resulting box plots are shown in Fig. 10.7 i) and show a clear disparity between the two bands of spectra. The median value for the SNR of the LHS is equal to 45 and the median value of the RHS is 26.

Despite the relatively low values for SNR for the cell spectra from glass, the spectral profile is clearly distinguishable and for some applications this spectral quality may be sufficient. In Fig. 10.7 it is demon-

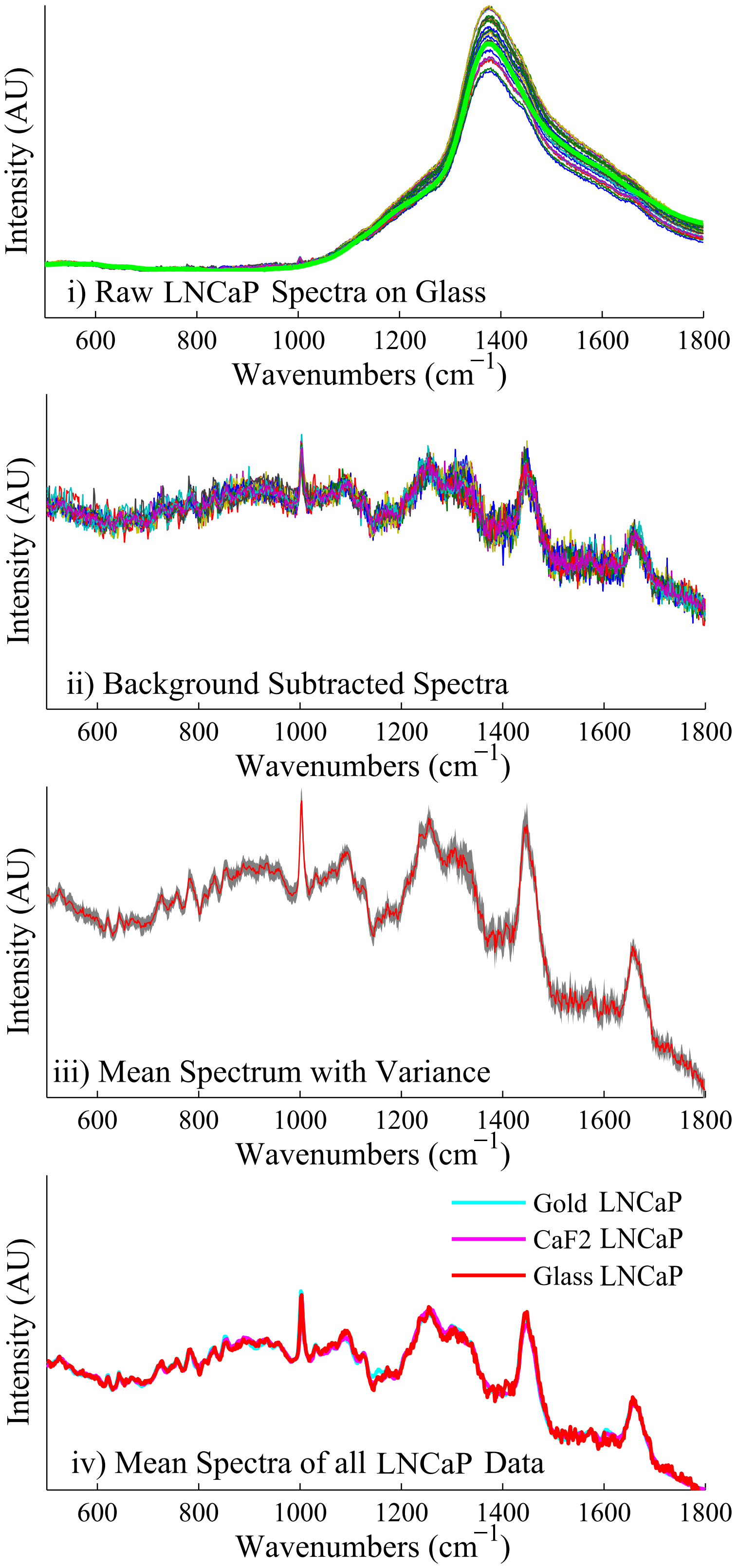


Figure 10.6: Results for LNCaP cells deposited on glass slides. The raw Raman spectra are shown in (i) in which it can be seen that the glass spectrum overwhelms the cell spectrum; (ii) The same data is shown in (ii) following EMSC correction using a glass reference spectrum. (iii) shows the mean spectrum of the EMSC corrected dataset as well as the variance and finally (iv) overlays the mean spectra taken from gold, CaF2 and glass.

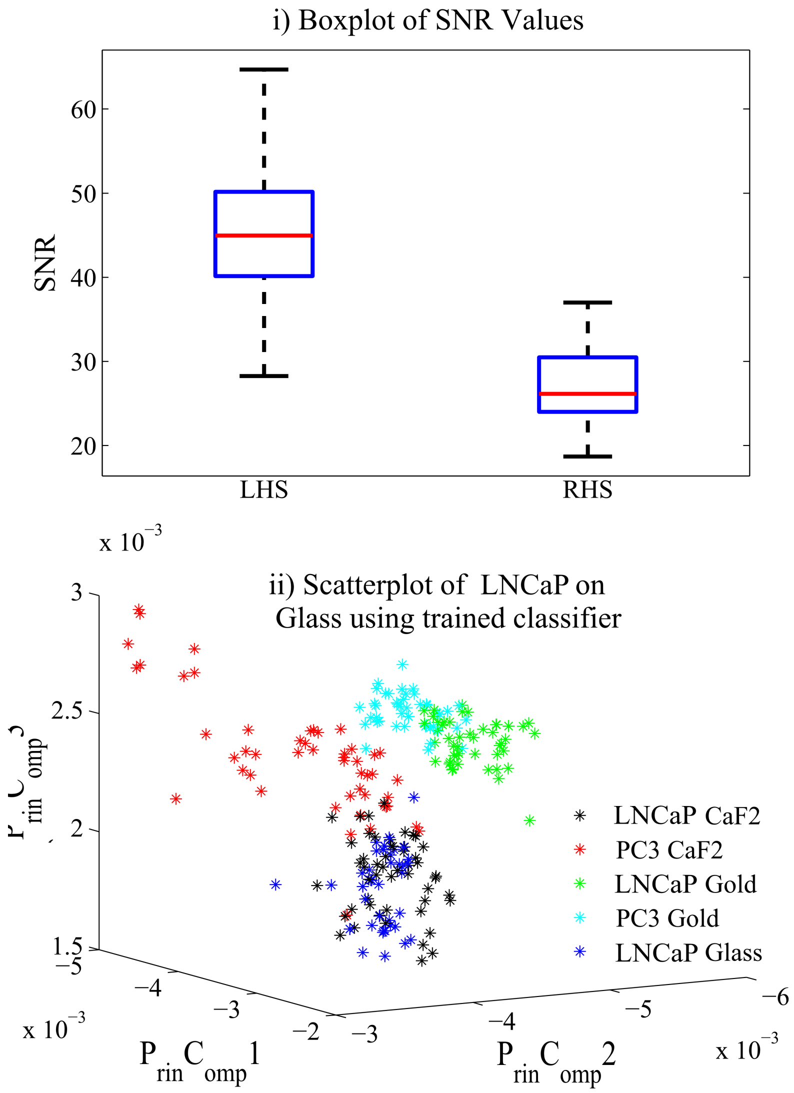


Figure 10.7: i) Boxplot of the SNR values of the LHS and RHS indicated in Fig. 10.6 iii). ii) shows the results of projecting the LNCaP cell spectra from glass onto the $P_{\mathrm{rin}}C_{\mathrm{omp}}$ loadings created by training a classifier on all spectra obtained from CaF2 and Gold. The resulting scores correlate well the LNCaP cell spectra on CaF2.

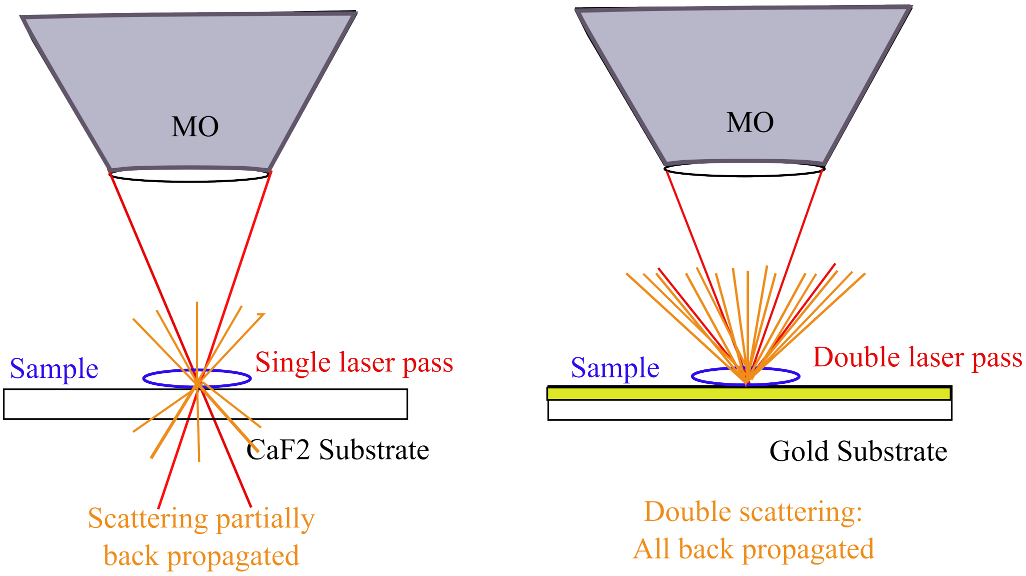


Figure 10.8: Illustration of the increased Raman intensity afforded by a reflective substrate.

strated that it is possible to classify these spectra with high accuracy using the PCA model that was trained on the gold and CaF2 data. The resulting scores from the first three $P_{rin}C_{omp}$ loadings are shown in Fig. 10.7 ii). This 3D figure represents similar information to that contained in Fig. 10.4 and Fig. 10.5, however, it shows the group and substrate separation simultaneously in 3D space. The variance explained by $P_{rin}C_{omp}$ 1, 2, and 3 is 34%, 14%, and 12% respectively. It can be seen that the scores overlap with the scores resulting from the most closely related dataset i.e. LNCaP on CaF2. We can, therefore, postulate that it may be possible to classify cells on pathological glass slides using Raman spectroscopy with 785 nm excitation.

10.5 Summary

The first conclusion that may be drawn from this chapter is that a reflective substrate may be optimal for cell analysis using Raman microspectroscopy with NIR excitation rather than the current gold standard of CaF2. In this chapter, it has been demonstrated that a reflective substrate in the form of a glass slide with a 100 *nm* deposition of gold, yields spectra with significantly higher SNR. In the context of the unwanted background signal from the substrate, both the CaF2 and gold substrates provide a negligible signal. The gold substrate is significantly less expensive than Raman grade CaF2, approximately 5 euro per substrate compared to 150. In addition, the intensity of a cell spectrum recorded with 785 *nm* is shown to be approximately 4 times stronger than that recorded from CaF2. The explanation for this interesting result is illustrated in Fig. 10.8 below.

The reflective substrate allows for forward scattered Raman photons, that would be lost through an opaque substrate, to be reflected back towards the microscope objective and collected. The result is an approximate doubling in Raman intensity. Secondly, the source laser is also reflected backwards, effectively doubling the excitation power. Therefore, reflective substrates will yield approximately quadruple the spectral intensity resulting in an increase in SNR as a direct consequence. Disregarding dark current and read noise from the camera, the standard deviation of the shot noise is equal to the square root of the intensity. Taking the definition of signal to noise ratio as the intensity divided by the standard deviation of the noise, it can be expected that an increase in intensity by a factor of four will increase the SNR by a factor of two, as evidenced by Fig. 10.3.

An investigation of the all four recorded datasets using principal components analysis revealed that the CaF2 cell spectra separated from the gold cell spectra due to a higher variance in the CaF2 data. The two cell lines separated over the same components for both substrates. The results clearly demonstrate that the lower noise less variant spectra recorded from the gold substrates provided for significantly better clustering of the data. It can be expected that this would result in superior classification. It should be noted that not all reflective substrates are as chemically stable as the gold. In the course of this work, aluminium slides (100 *nm* Al deposited on glass) were investigated but were found to be reactive with the fixing agents that are used to preserve the cells.

The second conclusion in this chapter is that glass slides, in conjunction with background removal algorithm based on least squares fitting, are usable substrates for Raman microscopy with NIR excitation. Glass is seldom applied in this area due to the strong spectrum from the glass itself. Although, the resulting cell spectra have low SNR, approximately 30-50% the SNR that of equivalent spectra recorded from CaF2, it has been demonstrated that these spectra are still of sufficiently high quality for classification using PCA. Given the prevalence of glass slides in pathology and their low cost, it may be preferable to use glass slides with NIR excitation for certain applications. However, the use of glass would require optimal software denoising techniques to increase the SNR of the cell spectra recovered from the glass substrate.

The following chapters in this thesis focus solely on software based denoising techniques to increase the SNR. These software based solutions are based on two common post-processing steps applied to experimental Raman spectra: cosmic ray removal and smoothing to remove system noise. Chapter 8 proposes an alternative algorithm for the removal of cosmic ray artefact removal that is demonstrated to result in spectra with a higher SNR than that of the double acquisition method. Chapter 9 proposes to enhance traditional SG smoothing with maximum likelihood estimation that yields spectra with higher SNR and greater peak fidelity than that of SG smoothing alone.

Chapter 11

Conclusion

Raman spectroscopy has widespread applications in several fields including clinical medicine, material science, industrial inspection, and pharmaceuticals. However, Raman scattering is a weak process and for a majority of samples, low photon counts can be expected in the Raman spectrum. This is particularly true in the field of medicine, in which low laser powers are required to avoid burning the tissue or cells. Most commonly, Raman spectroscopy is coupled with multivariate classification for the purpose of distinguishing or identifying different samples. One example is the diagnosis of different diseases based on the Raman spectra recorded from cell or tissue. The low photon count in a Raman spectrum, results in high noise levels and low values of signal to noise ratio; this can adversely affect the capability of Raman spectroscopy to be used for classification. The various contributions outlined in this thesis are related to the development of methodologies that enhance the SNR in a Raman spectrum. These contributions take several forms, including methods to evaluate and predict performance of a given recording system in terms of SNR or the output spectra, as well as algorithmic methods that can be applied post-capture for the purpose of increasing SNR. These increases in SNR, which are achieved through the various contributions proposed herein can lower acquisition times, increase throughput in clinical systems, increase classification accuracy, and may reduce cost. Specifically the contributions in this thesis are as follows:

- A key contribution in this thesis is a method to model the noise in a given Raman spectroscopy system. By doing this, the user can create artificial datasets without extensive experimental procedures and use them to optimise the performance of the system. This process enables rigorous evaluation of the effect of each recording parameter on the SNR of the collected data and the resulting classification accuracy that can be expected for a given application. The application (and, therefore, sample) of interest as well as the recording system will determine the system specific irradiance .

- In Chapter 5 it is shown how artificial datasets may be used to systematically evaluate the required minimum acquisition time of a particular sample on a given Raman spectrometer in order to provide reliable classification. In a similar manner, these datasets can be used to predict the optimal read-out parameters for the CCD, as discussed in Chapter 6. The subsequent improvement in SNR achieved through this optimisation may be translated into a reduction in acquisition time, which is significant when one considers that recording times of up to one minute are not uncommon for a single spectrum. It is this limitation that has prevented Raman spectroscopy from being applied to high throughput applications such as cervical screening. Any minor saving in acquisition time for a single spectrum may contribute to a major saving in time over the course of an extensive experiment.

- Optical elements present in a Raman spectrometer can be expensive, the CCD being arguably the most expensive as well as affecting all sources of noise in the recorded spectra. Investing in a new CCD can be an expensive procedure, for which cost and performance must be balanced. This new CCD may not necessarily guarantee a significant improvement in system performance. Following the procedures described in Chapter 7, it is possible for an experimentalist to establish a more reliable cost-benefit analysis of replacing the CCD in the system. This chapter also provides a brief overview of how this methodology may be extended to other optical elements in the system such as the microscope objective, diffraction grating, and source laser.

- While not strictly part of the system, the substrates upon which the samples are deposited can have a significant impact on the spectral intensity and the SNR. NIR lasers are commonly used in the inspection of biological samples; however, it is difficult to use NIR excitation in conjunction with glass substrates due to the overwhelming spectrum from the glass. A potential solution to this is to replace the glass substrates with reflective substrates, particularly a glass substrate with a 100 *nm* gold film deposited on the surface. Gold is chemically stable, produces a negligible background, and significantly enhances the Raman scatter produced by and collected from the sample. In Chapter 10, we demonstrate the potential of these substrates to enhance the SNR by 100% or, alternatively, reduce the acquisition time by 50%. A second contribution is also contained in this chapter, whereby we show that recently proposed background subtraction methods can reliably extract recognisable spectral features from data recorded from glass substrates. Although is is possible to subtract the glass spectrum, the fluctuation in SNR across the wavenumber range that is introduced by the shot noise from the intense glass background may provide challenges for reliable classification, especially if the glass background saturates the camera.

- Once the possibilities of hardware optimisation have, within reason, been exhausted. Post-processing algorithms designed to efficiently denoise spectra have the potential to further improve the SNR of the recorded spectra. These algorithms may be used to reduce the acquisition time or to sufficiently improve system performance so that reliable classification may be achieved. The two algorithms proposed in this paper are an algorithm for cosmic ray removal and a denoising algorithm. Chapter 8 illustrated how the averaging together of multiple acquisitions can degrade the SNR of recorded spectra. This is an undesirable but inevitable consequence of the most commonly used method for cosmic ray removal. The proposed algorithm alters the operating principles of the double acquisition method so that it may be applied to databases of spectra of single capture spectra. This results in an improvement in the SNR of 10% when compared to the double acquisition method. Another advantage of this method is that it may be used to process pre-recorded datasets for which double acquisition is no longer possible.

- The SNR naturally fluctuates across the wavenumber range being inspected due to the difference in intensity of the spectral peaks and, therefore, the different degrees of shot noise. Typically, smoothing algorithms are applied to all wavenumber regions of the spectrum in the same manner. This can negatively affect the underlying spectral features particularly for the case of a sharp peak. The proposed algorithm, discussed in Chapter 9, enhances traditional Savitzky-Golay smoothing with Maximum Likelihood Estimation, as well as implementing an early stopping procedure in user defined peak regions, resulting in up to 100% improvement in SNR when compared to the raw data. The algorithm consistently outperforms competing algorithms in terms of the SNR and the preservation of peak fidelity.

Appendix A

Supplementary information for Chapter 8

A.1 MATLAB Code

This section contains the MATLAB code to perform database CRA removal as detailed in Chapter 8.

```
% Dataset is an mxn matrix with m baseline corrected spectra of n sammples
function[dataset] = CRARreboot(dataset)

% Determine dimensions of matrix for reference
dim = size(dataset);

figure, plot(dataset.’)

% Median filtering to prevent CRA matching in high CRA data
for h = 1 : 1 : dim(1)
    refData(h,:) = medfilt1(dataset(h,:),11);
end

% Ascertain the correlation between each spectrum
% For each row of data
```

```
for i = 1 : dim(1)

    % Correlate that spectrum with every other spectrum

    for j = 1 : dim(1)

        % Breaking down the correlation calculation

        d1 = dot(refData(i,:), refData(j,:));

        d2 = dot(refData(i,:), refData(i,:));

        d3 = dot(refData(j,:), refData(j,:));

        corrVec(j) = (power(d1,2)/(d2*d3));

        % If a spectrum is going to be correlated with itself, record a null value

        if i == j

            corrVec(j) = 0;

        end

    end

    % Determine the optimum pairings of spectra (corrVec should have the same number of

    %elements as rows in the dataset so dim is reused)

    for k = 1 : dim(1)

        % If a particular element is equal to the max of the vector then

        % this is the most similar spectrum and the position of it is recorded in pairVec

        if corrVec(k) == max(corrVec)

            pairVec(i) = k;

        end

    end

end


% For every spectrum

for h = 1 : dim(1)

    % Take histogram of signal and exclude the the highest 5% of the signal

    % from the calculation to prevent stdDev skew

    counter = 1;

    remains = dataset(h,:) - sgolayfilt(dataset(h,:),3,9);

    [counts,centres] = hist(remains,6);
    % Discard potential cosmic rays from the standard deviation calculation
    for q = 1 : 1 : length(remains)
    if remains(q) < centres(4)
        cleanRem(counter) = remains(q);
        counter = counter + 1;
    end
    end
    % Define threshold
    sigma = std(cleanRem);

    % Subtract the correlated spectrum from the original
    tempSp = dataset(h,:) - dataset(pairVec(h),:);

    % If there is a significant difference between them then copy over the values
    for g = 1 : dim(2)
        if tempSp(g) > sigma*5
            dataset(h, g) = dataset(pairVec(h), g);

            % If a cosmic ray is detected then apply lower filters to the surrounding area
            for p = -1 : 1 : 1
                if g+p < 1 || g+p > dim(2)
                    continue
                else
                if tempSp(g+p) > sigma*2
                    dataset(h, g+p) = dataset(pairVec(h), g+p);
                end
                end
            end
        end
    end
end
```

```
figure, plot(dataset.’)


end
```

Appendix B

Supplementary Information for Chapter 9

B.1 MLE-SG Matlab Code

These two function files are written for MATLAB software and can be copied and pasted into .m files. In order for them to run correctly the two files must be saved into the same folder as each other and the data that they are required to process. The code for this is presented in the following sections, as well as the commands to call the code from the MATLAB workspace.

B.1.1 Workspace Function Call

```
% Command to call the MLESG algorithm from the MATLAB workspace
% Please note that this function call uses the algorithms inbuilt parameters to process the spectra
m = calculatem(y,wavNo,pkLoc,gSig);
xe = MLESGcore(y, m);
figure, plot(wavNo, y, ’r’)
hold
plot(wavNo, xe, ’g’)
```

B.1.2 Early Stopping Procedure

Please note that the last 4 input parameters are to enable users to perform additional optimisation of the algorithms performance. If these inputs are not specified, then the algorithm will estimate the SNR automatically and use the inbuilt parameters specified for that particular SNR.

```
% y is the raw data vector, which is a function of sample index i
% wavNo is the axis for y in cm^-1
% m is the iterations vector, which is a function of sample index i
% pkLoc is a vector containing a list of peak locations in cm^-1
% gSig is the standard deviation of the Gaussian used in Eq.14 in paper, 10 is recommended
% SNR is the SNR of the signal itself
% minm and maxm are respectively the min and max numbers of iterations
% mu is the background signal vector (mean dark current read noise)
function m = calculatem(y,wavNo,pkLoc,gSig,SNR,minm,maxm,mu)
% Auto-fit Gaussian curves of iterations to peak regions
% defined by the user
% Scale gSig to the wavenumber axis
gSig = gSig*(wavNo(2)-wavNo(1));
if nargin == 7
    mu=0;
end
if nargin == 6
    mu=0;
    SNR=0;
end
if nargin == 5
    mu=0;
    if SNR<=200
        p = [-0.0000000000002051163714896419 4179,
        0.00000000017456738606819003079,
        -0.00000006108370350381461555,
        0.000011326343084825261969,
        -0.0011967668932480550 99,
        0.072278932710701945807,
        -2.3738144356931862866,
        36.543962848162728108];
        minm = round(polyval(p,SNR));
        p2 = [0.000000049662053296461794872,
        -0.000045735657830153646274,
        0.013867582814278392803,
        -1.7581457324626106331,
        88.680082559339525284];
        maxm = round(polyval(p2,SNR));
    else
        minm = 2; maxm = 5;
    end
end
if nargin == 4
    % If SNR value is not given then estimate from raw
    mu=0;
    temp = sgolayfilt(y-mu,3,9); noise = y - mu - temp;
    clear temp;
    SNR = (max(y-mu)/(std(noise))); clear noise;
    if SNR<=200
        p = [ -0.00000000000020511637148964194179,
        0.00000000017456738606819003079,
        -0.000000061083703503814615 55,
        0.000011326343084825261969,
        -0.0011967668932480550 99,
        0.072278932710701945807,
        -2.3738144356931862866,
        36.543962848162728108];
        minm = round(polyval(p,SNR));
        p2 = [0.000000049662053296461794872,
        -0.000045735657830153646274,
        0.013867582814278392803,
        -1.7581457324626106331,
        88.680082559339525284];
        maxm = round(polyval(p2,SNR));
    else
        minm = 2; maxm = 5;
    end
end
x=y-mu; clear y; clear mu;
N = length(wavNo);
G = zeros(length(pkLoc),N);
for i=1:1:length(pkLoc)
    G(i,:) = 1 -
    exp(-(power(wavNo-pkLoc(i),2))/(2*power(gSig,2)));
end
% Calculating minimum number of iterations along vertical
m=min(G);
% Scaling with respect to minimum and maximum values m can have
m=m*(maxm - minm) + minm;
end
```

B.1.3 Core MLE-SG Algorithm

Please note that the input parameters lambda, p, v, q, and mu are to enable users to perform additional optimisation of the algorithms performance. If these inputs are not specified, then the algorithm will utilise the inbuilt parameters available to it.

```
    % y is the raw data vector, which is a function of sample index i
% m is the iterations vector, which is a function of sample index i
% lambda and p, default values of 1.8 and 0.4
% v and q are SG papramaters and have default values of 5 and 7
% mu is the background signal vector (mean dark current + read noise)
function xe = MLESGcore(y, m, v, q, lambda, p, mu)
if nargin == 2
    v=5;q=7;lambda=1.8;p=0.4;mu=0;
elseif nargin == 4
    lambda=1.8;p=0.4;mu=0;
elseif nargin == 6
    mu=0;
end
%Ensure m contains integer values
m=round(m);
% Determine first guess for the algorithm
x=y-mu; clear y; clear mu;
N = length(x);
% An estimate is required for the search window in which to calculate the MLE; this estimate is based on
% the approximate value for the standard deviation of the noise terms as calculated here.
% The range is equal to 6 times this value.
temp = sgolayfilt(x,3,9); noise = x - temp; sigma = std(noise);
clear temp; clear noise;
% Apply MLE
for j = 1 : 1 : max(m)
    if(j <= min(m))
        v =5;q =7;
    elseif (j>max(m)-round(max(m)/5))
        q = q+4; lambda=lambda*10;
    else
        v=3;q =5;
    end
    % First guess
    if j==1
        xe=x;
    end
    % Note this is denoted as x' in the paper and represents the smoothed signal used within the MLE estimation
    xdash = sgolayfilt(xe, v, q);
    % Apply MLE to each sample index
    for i = 1 : 1 : N
        % Only apply MLE if iteration number for this sample index is greater than the current iteration
        % number
        if m(i) >= j
            % This is the vector of possible xe values we apply MLE over
            MLErange = (xe(i)-3*sigma:6*sigma/100:xe(i)+3*sigma);
            % Minimise the likelihood function (lFunc)
            % This will be assigned the likelihood estimations for each value of x in the range of values
            % MLErange
            leMLErange = zeros(size(MLErange));
            for k = 1 : 1 : length(MLErange)
                limit1 = lambda*
                (power(abs(MLErange(k)-xdash(i)),p));
                limit2 = power(x(i) -
                MLErange(k),2)/(2*power(sigma,2));
                % Calculate likelihood estimation (le) for each value in range for sample i
                leMLErange(k) = limit1 + limit2;
            end
            [a, b] = min(leMLErange);
            % MLE for sample i in iteration j
            xe(i) = MLErange(b);
        end
    end
end
```

end

B.2 Early Stopping Optimisation

The following section provides details of the optimisation process of the proposed algorithm in terms of Signal to Noise Ratio (SNR), as defined in the main body of the paper. These optimisation processes take place after the optimisation of the input parameters λ and p. In this case the focus is placed on the Savitsky-Golay (SG) input parameters and the number of iterations required by the algorithm to achieve the maximum SNR for both the whole spectrum and a peak of interest. The peak was chosen as a region of interest due to it's narrow width and relatively low intensity which renders it a challenge to preserve effectively. Optimisation processes were performed on artificial plastic datasets with a range of different input SNRs. Results shown are a mean of 100 individual spectra which were denoised using the algorithm. These graphs are where the optimum values for m_{max} and m_{min} were derived for three sets of SG input parameters. For aesthetic purposes the number of iterations is limited to 50 in the graphs, since for the majority of datasets the optimal number of iterations is under 50. However, for the lower SNR datasets the number of iterations was increased until the increase in SNR for the de-noised spectra had stabilised or reached it's maximum, the final decision for m_{max} is provided in the caption for these images. Please note that the iteration curves provided in Fig. 9.4 were derived from the curves associated with SG(5,7), which are illustrated in the following figures.

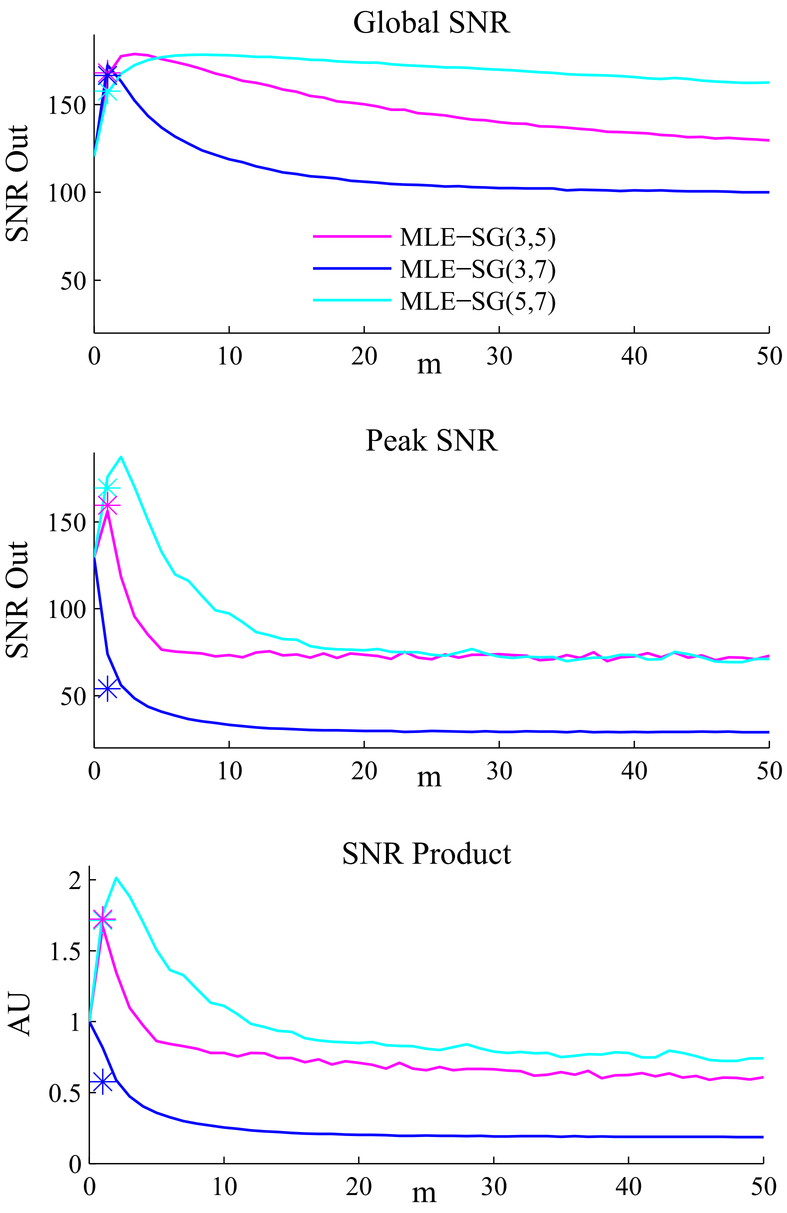


Figure B.1: Initial SNR: 120

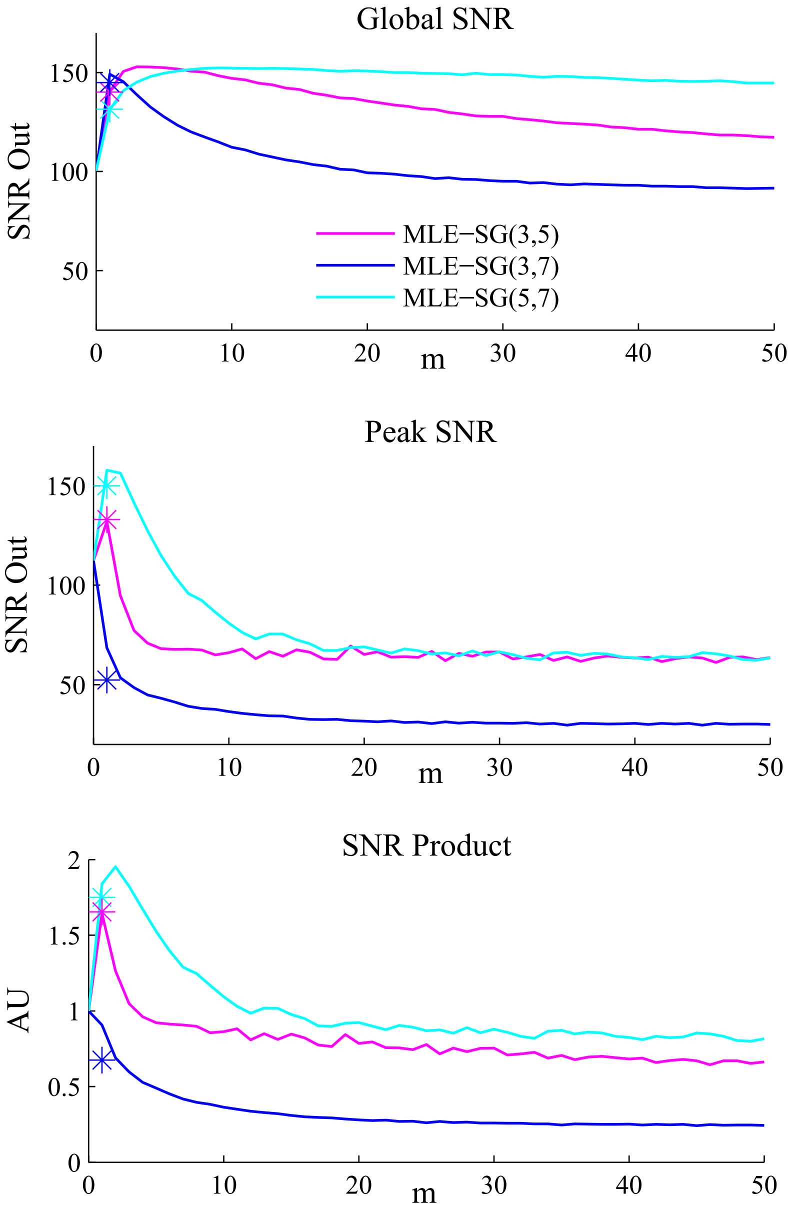


Figure B.2: Initial SNR: 100

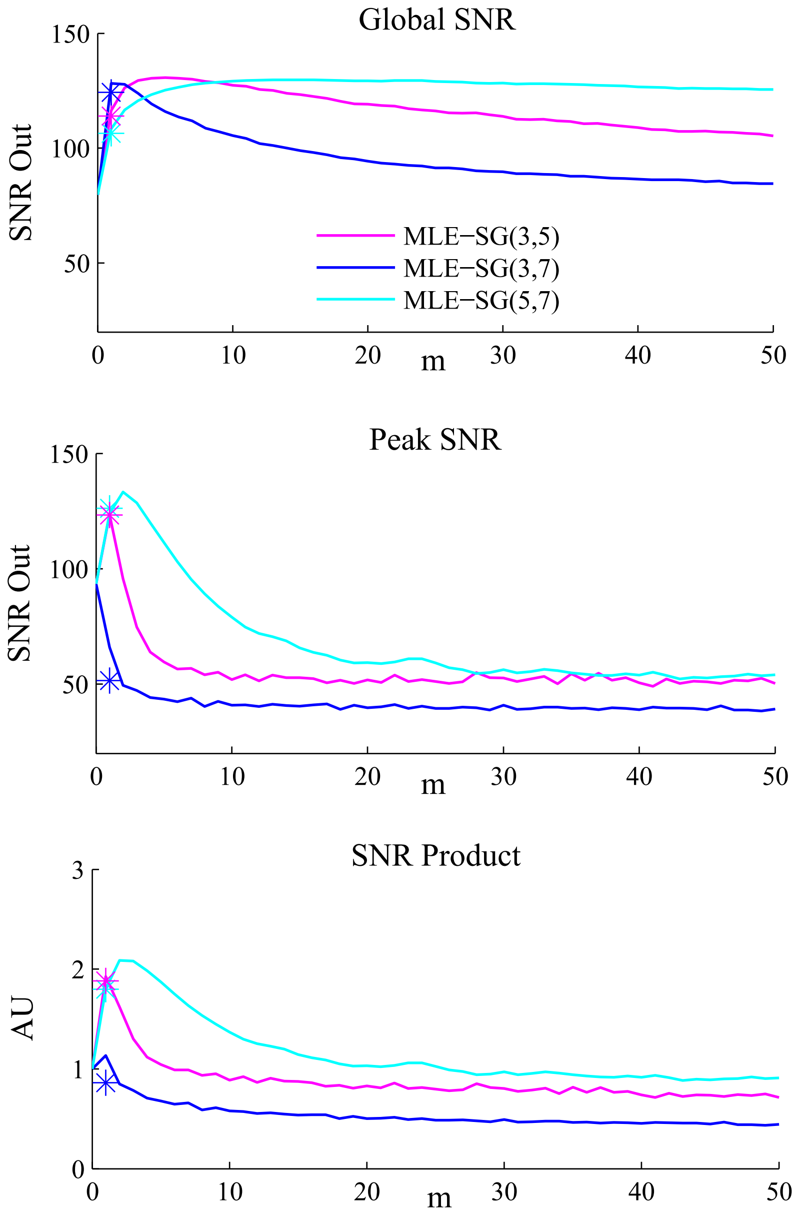


Figure B.3: Initial SNR: 80

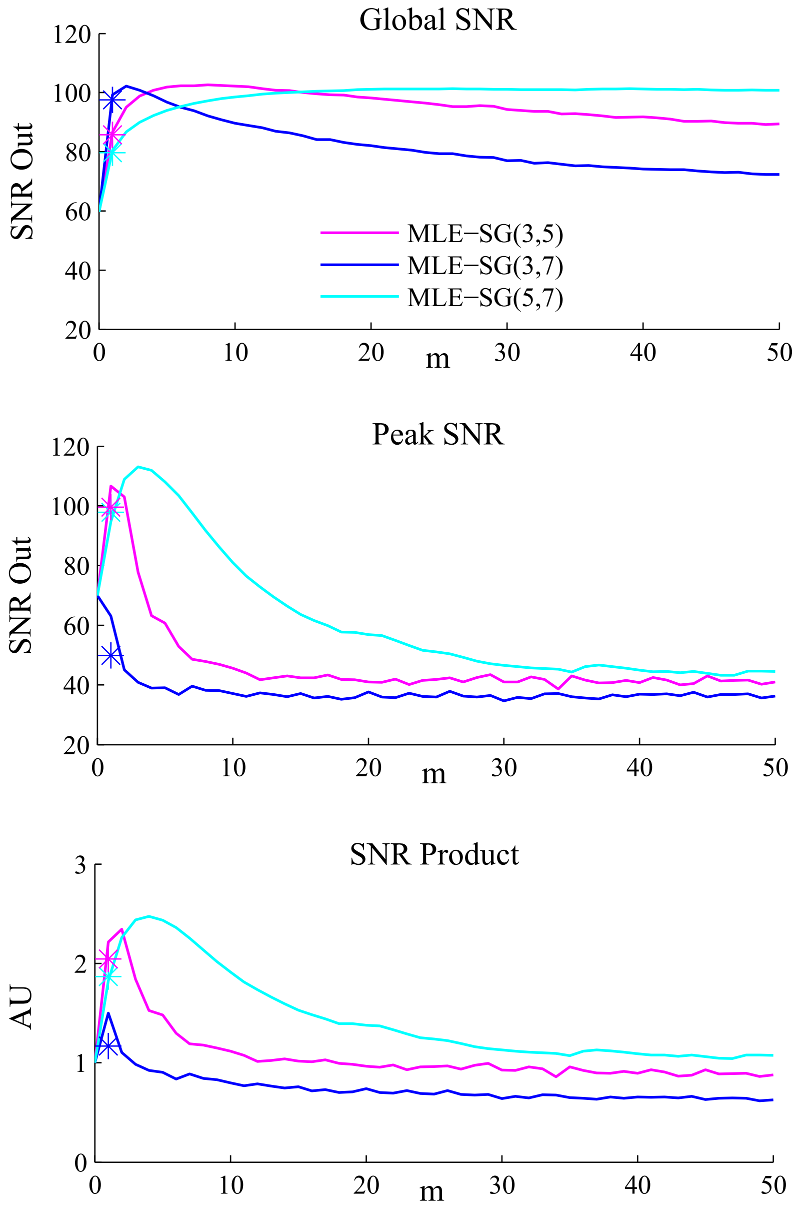


Figure B.4: Initial SNR: 60

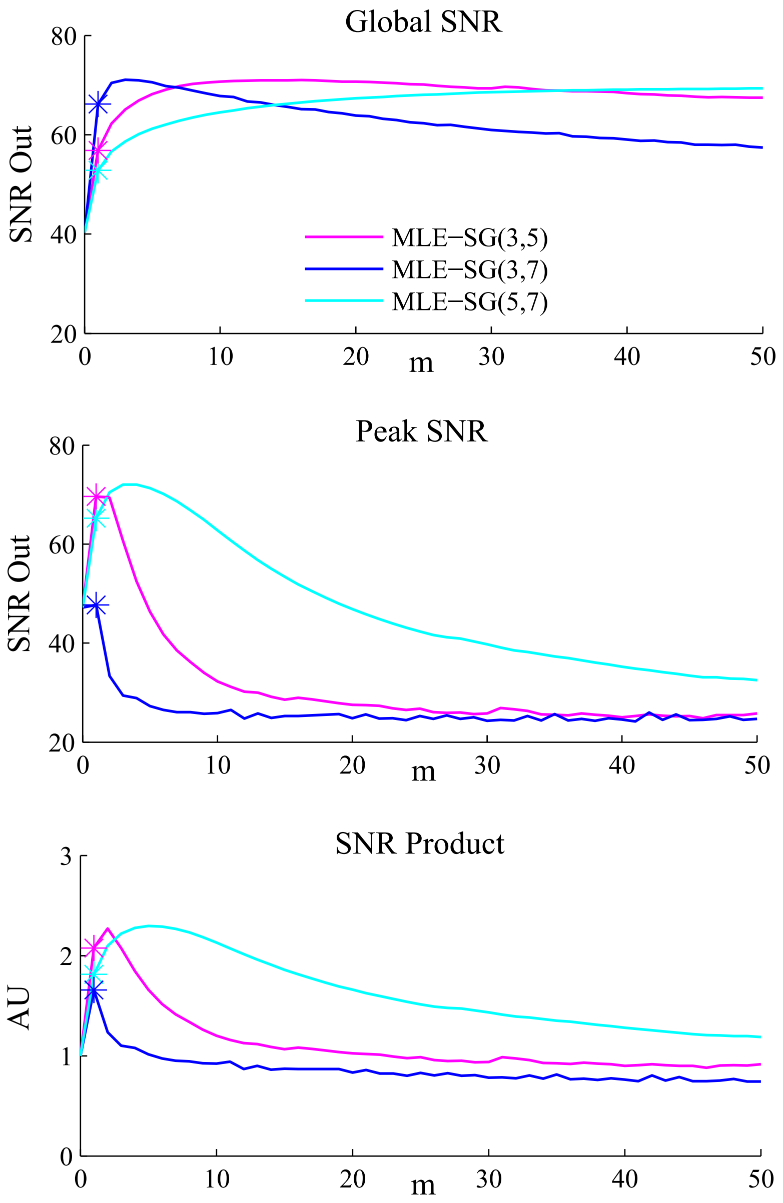


Figure B.5: Initial SNR: 40

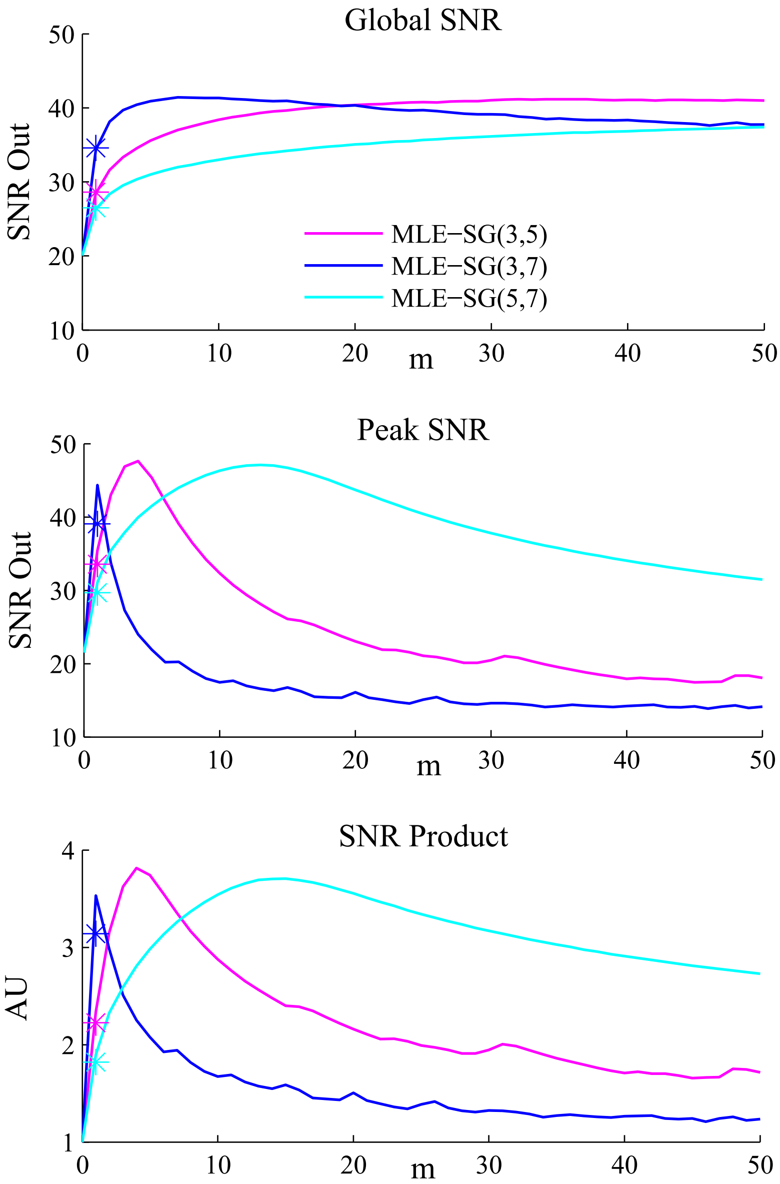


Figure B.6: Initial SNR: 20

B.3 Tables of Quantitative Values

In this section the quantitative results correspond to Fig. 9.8(a), Fig. 9.8(b), and Fig. 9.9 in the main body of the text are presented. These results clearly demonstrate enhancement oin SNR achieved by the proposed denoising algorithm compared to the two alternative methods examined in the paper. The value of the SNR is presented for each of the denoising methods. The metrics of Global SNR, Peak SNR, and the SNR product have all been calculated separately using the mathematical definitions provided in Equations 9.10, 9.11, and 9.12 respectively in the main chapter. Each value given in the tables is the average value calculated over a dataset of 100 similarly noised spectra.

	Raw Data	MLE-SG Data	SG(3,7) Data	Perfect Smoother
Global SNR	20.05	47.08	34.77	40.67
	30.07	64.90	50.58	55.10
	40.09	83.72	66.60	69.53
	49.79	101.34	81.83	83.73
	59.70	118.81	96.23	97.72
	72.25	140.88	114.46	115.93
	79.66	149.58	122.72	125.12
	91.67	165.90	136.53	140.61
	99.58	174.35	145.16	150.90
	109.11	190.49	156.62	164.38
	120.42	208.94	166.89	179.08
	130.82	221.94	175.39	191.92
	139.78	233.17	181.59	203.28
Peak SNR	21.65	40.47	35.41	15.84
	33.74	55.19	43.10	23.55
	44.99	70.12	47.04	31.64
	60.96	90.47	48.78	38.22
	72.50	101.86	49.73	45.31
	78.18	115.15	50.74	61.23
	97.53	128.54	51.24	65.94
	97.71	140.68	52.05	73.99
	104.36	146.95	51.40	75.90
	118.70	162.36	52.08	90.27
	130.30	180.20	52.35	114.81
	143.08	188.77	52.87	123.18
	153.78	198.30	53.22	127.84
SNR Product	1	4.50	2.93	1.58
	1	3.48	2.24	1.38
	1	3.26	1.84	1.31
	1	3.16	1.42	1.14
	1	2.95	1.20	1.11
	1	2.94	1.09	1.32
	1	2.61	0.88	1.15
	1	2.65	0.84	1.23
	1	2.53	0.76	1.15
	1	2.47	0.67	1.21
	1	2.44	0.59	1.38
	1	2.28	0.52	1.32
	1	2.19	0.48	1.29

Table B.1: A table of the values of the SNR calculated based on the simulated plastic datasets for the three denoising methods using the three SNR metrics described in the main body of the paper. Corresponds to Fig. 9.8(a).

	Raw Data	MLE-SG Data	SG(3,7) Data	Perfect Smoother
Global SNR	25.41	54.51	43.15	49.21
	39.55	76.73	65.46	69.03
	50.48	93.99	82.28	84.24
	61.56	110.40	98.04	99.69
	70.44	122.73	110.66	111.68
	78.38	132.27	121.47	122.19
	86.51	143.70	132.74	134.21
	94.42	151.43	141.50	144.86
	102.95	162.12	152.46	155.56
	110.52	175.34	161.24	165.56
	117.26	184.77	167.35	174.31
	122.74	191.63	173.49	181.21
	128.74	200.17	180.24	190.06
	132.74	206.34	183.28	194.39
	140.96	217.59	191.57	206.19
Peak SNR	24.84	49.73	41.92	20.12
	33.12	65.13	46.58	29.60
	42.79	67.25	45.22	36.08
	50.36	83.42	50.80	47.80
	59.41	88.45	55.51	56.23
	61.90	99.63	55.88	61.16
	72.48	100.52	55.92	67.09
	73.44	108.90	55.59	73.11
	77.63	107.43	55.73	82.97
	87.85	127.07	58.63	94.73
	91.44	132.82	55.92	102.23
	93.55	133.88	57.81	94.35
	99.21	146.91	60.09	118.80
	99.38	135.94	58.57	112.49
	118.04	168.48	58.74	124.92
SNR Product	1	4.26	2.88	1.63
	1	3.93	2.43	1.64
	1	2.99	1.78	1.46
	1	2.94	1.67	1.59
	1	2.68	1.63	1.65
	1	2.77	1.47	1.61
	1	2.33	1.26	1.52
	1	2.43	1.21	1.63
	1	2.17	1.12	1.67
	1	2.32	1.02	1.66
	1	2.32	0.93	1.76
	1	2.24	0.93	1.55
	1	2.31	0.90	1.81
	1	2.14	0.86	1.73
	1	2.23	0.73	1.62

Table B.2: A table of the SNR values calculated based on the experimental plastic datasets for the three denoising methods using the three SNR metrics described in the main body of the paper. Corresponds to Fig. 9.8(b).

	Raw Data	MLE-SG Data	SG(3,7) Data	Perfect Smoother
Global SNR	19.03	42.58	32.71	41.43
	31.98	63.07	53.91	60.69
	38.25	74.17	63.73	69.85
	47.75	87.80	79.01	83.78
	56.39	95.29	91.15	94.42
	68.42	118.58	107.84	109.28
	79.83	136.76	122.15	123.16
	87.58	148.70	130.91	132.74
	95.33	160.07	139.45	141.97
	106.43	171.97	150.45	155.34
	113.45	171.34	156.24	162.44
	125.73	188.91	166.63	177.33
	134.51	201.01	173.19	186.50
	146.21	213.07	180.95	202.04
	153.76	216.01	184.38	207.15
	163.31	235.19	190.65	218.16
	171.43	245.32	195.91	228.00
	183.99	255.30	201.73	240.17
	190.86	261.84	203.76	245.94
Peak SNR	23.37	32.65	27.79	14.79
	34.81	49.66	32.12	18.93
	42.35	60.27	33.80	21.92
	54.10	67.40	33.47	22.89
	61.01	82.75	34.55	25.94
	78.23	97.36	34.60	29.25
	88.43	107.57	35.53	33.77
	100.03	107.78	35.35	37.77
	111.43	119.77	35.20	41.96
	119.32	133.85	35.30	45.27
	131.53	139.37	35.24	49.09
	130.34	147.12	35.29	60.04
	150.08	162.68	35.51	60.96
	166.51	176.44	35.43	66.66
	173.27	184.73	35.41	72.62
	191.48	196.19	35.30	72.27
	187.99	195.95	35.56	78.38
	211.02	218.31	35.72	79.74
	207.38	222.73	35.47	86.14
SNR Product	1	3.35	2.25	1.56
	1	2.93	1.65	1.11
	1	2.87	1.44	1.03
	1	2.39	1.08	0.79
	1	2.34	0.98	0.76
	1	2.21	0.75	0.65
	1	2.18	0.66	0.63
	1	1.93	0.58	0.63
	1	1.89	0.50	0.60
	1	1.91	0.45	0.59
	1	1.66	0.41	0.58
	1	1.69	0.38	0.67
	1	1.65	0.33	0.61
	1	1.57	0.29	0.60
	1	1.53	0.27	0.63
	1	1.52	0.24	0.56
	1	1.52	0.23	0.60
	1	1.46	0.20	0.53
	1	1.47	0.20	0.57

Table B.3: A table of the SNR values calculated based on the simulated T24 datasets for the three denoising methods using the three SNR metrics described in the main body of the paper. Corresponds to Fig. 9.9

Bibliography

[1] Tineke Vankeirsbilck, Ann Vercauteren, Willy Baeyens, G Van der Weken, Francis Verpoort, Geert Vergote, and Jean Paul Remon. Applications of raman spectroscopy in pharmaceutical analysis. *TrAC trends in analytical chemistry*, 21(12):869–877, 2002.

[2] L-P Choo-Smith, HGM Edwards, H Ph Endtz, JM Kros, Freerk Heule, Hugh Barr, Joe Sam Robinson Jr, HA Bruining, and GJ Puppels. Medical applications of raman spectroscopy: from proof of principle to clinical implementation. *Biopolymers: Original Research on Biomolecules*, 67(1):1–9, 2002.

[3] Andrzej Kudelski. Analytical applications of raman spectroscopy. *Talanta*, 76(1):1–8, 2008.

[4] Qiang Tu and Chang Chang. Diagnostic applications of raman spectroscopy. *Nanomedicine: Nanotechnology, Biology and Medicine*, 8(5):545–558, 2012.

[5] Bernhard Schrader. *Infrared and Raman spectroscopy: methods and applications*. John Wiley & Sons, 2008.

[6] EE Lawson, BW Barry, AC Williams, and HGM Edwards. Biomedical applications of raman spectroscopy. *Journal of Raman Spectroscopy*, 28(2-3):111–117, 1997.

[7] B.E.A. Saleh and M.C. Teich. *Fundamentals of Photonics*. Wiley Series in Pure and Applied Optics, 2007.

[8] L Max Almond, Joanne Hutchings, Neil Shepherd, Hugh Barr, Nick Stone, and Catherine Kendall. Raman spectroscopy: a potential tool for early objective diagnosis of neoplasia in the oesophagus. *Journal of Biophotonics*, 4(10):685–695, 2011.

[9] MCM Grimbergen, CFP Van Swol, C Kendall, RM Verdaasdonk, N Stone, and JLHR Bosch. Signal-to-noise contribution of principal component loads in reconstructed near-infrared raman tissue spectra. *Applied spectroscopy*, 64(1):8–14, 2010.

[10] H. Takeuchi and I. Harada. Simple and efficient method to eliminate spike noise from spectra recorded on charge-coupled device detectors. *Applied Spectroscopy*, 47:129–131, 1993.

[11] A. Savitsky and M.J.E Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[12] L.T. Kerr, H.J. Byrne, and B.M. Hennelly. Optimal choice of sample substrate and laser wavelength for raman spectroscopic analysis of biological specimens. *Analytical Methods*, 7:5041 – 5052, 2015.

[13] Chandrasekhara V Raman. The colour of the sea. *Nature*, 108(2716):367, 1921.

[14] Chandrasekhara Venkata Raman and Kariamanikkam Srinivasa Krishnan. The optical analogue of the compton effect. *Nature*, 121(3053):711, 1928.

[15] NVVSS Raman, AVSS Prasad, and K Ratnakar Reddy. Strategies for the identification, control and determination of genotoxic impurities in drug substances: A pharmaceutical industry perspective. *Journal of pharmaceutical and biomedical analysis*, 55(4):662–667, 2011.

[16] An Tu and PD Persans. Raman scattering as a compositional probe of ii-vi ternary semiconductor nanocrystals. *Applied physics letters*, 58(14):1506–1508, 1991.

[17] Nikolaos Kourkoumelis, Georgios Gaitanis, Aristea Velegraki, and Ioannis D Bassukas. Nail raman spectroscopy: A promising method for the diagnosis of onychomycosis. an ex vivo pilot study. *Medical Mycology*, 56(5):551–558, 2018.

[18] Nikolaos Kourkoumelis, Ioannis Balatsoukas, Violetta Moulia, Aspasia Elka, Georgios Gaitanis, and Ioannis D Bassukas. Advances in the in vivo raman spectroscopy of malignant skin tumors using portable instrumentation. *International journal of molecular sciences*, 16(7):14554–14570, 2015.

[19] LS Ornstein, J Rekveld, et al. Intensity measurements in the raman effect and the distribution law of maxwell-boltzmann. *Physical Review*, 34(5):720, 1929.

[20] Ewen Smith and Geoffrey Dent. *Modern Raman spectroscopy: a practical approach*. John Wiley and Sons, 2013.

[21] Stephen T Thornton and Andrew Rex. *Modern physics for scientists and engineers*. Cengage Learning, 2012.

[22] James E Wollrab. *Rotational Spectra and Molecular Structure: Physical Chemistry: a Series of Monographs*, volume 13. Academic Press, 2016.

[23] Dixit N Sathyanarayana. *Vibrational spectroscopy: theory and applications*. New Age International, 2015.

[24] Michael Mazilu, Anna Chiara De Luca, Andrew Riches, C Simon Herrington, and Kishan Dholakia. Optimal algorithm for fluorescence suppression of modulated raman spectroscopy. *Optics express*, 18(11):11382–11395, 2010.

[25] Hirotaka Fujimori, Masato Kakihana, Koji Ioku, Seishi Goto, and Masahiro Yoshimura. Advantage of anti-stokes raman scattering for high-temperature measurements. *Applied Physics Letters*, 79(7):937–939, 2001.

[26] Peter Atkins, Julio De Paula, and James Keeler. *Atkins' physical chemistry*. Oxford university press, 2018.

[27] Peter W Atkins and Ronald S Friedman. *Molecular quantum mechanics*. Oxford university press, 2011.

[28] JR Ferraro, K Nakamoto, and Chris W Brown. Introductory raman spectroscopy, 2003.

[29] Robert L Brooks. *The fundamentals of atomic and molecular physics*. Springer, 2013.

[30] Ian R Lewis and Howell Edwards. *Handbook of Raman spectroscopy: from the research laboratory to the process line*. CRC Press, 2001.

[31] Didier Hutsebaut, Peter Vandenabeele, and Luc Moens. Evaluation of an accurate calibration and spectral standardization procedure for raman spectroscopy. *Analyst*, 130(8):1204–1214, 2005.

[32] Steven J Choquette, Edgar S Etz, Wilbur S Hurst, Douglas H Blackburn, and Stefan D Leigh. Relative intensity correction of raman spectrometers: Nist srms 2241 through 2243 for 785 nm, 532 nm, and 488 nm/514.5 nm excitation. *Applied spectroscopy*, 61(2):117–129, 2007.

[33] R Wolthuis, TC Bakker Schut, PJ Caspers, HPJ Buschman, TJ Römer, HA Bruining, and GJ Puppels. Raman spectroscopic methods for in vitro and in vivo tissue characterization. In *Fluorescent and Luminescent Probes for Biological Activity (Second Edition)*, pages 433–455. Elsevier, 1999.

[34] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2(1):1–38, 2015.

[35] N. Stone, C. Kendall, J. Smith, P. Crow, and H. Barr. Raman spectroscopy for identification of epithelial cancers. *Faraday Discussions*, 126:141–157, 2004.

[36] S. Dochow, et al. Classification of raman spectra of single cells with autofluorescence suppression by wavelength modulation. *Analytical Methods*, 5:4608–4614, 2013.

[37] N.K. Afseth, V.H. Segtnan, and J.P. Wold. Raman spectra of biological samples: A study of preprocessing methods. *Applied Spectroscopy*, 60:1358–1367, 2006.

[38] N.K. Afseth and A. Kohler. Extended multiplicative correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 117:92–99, 2012.

[39] Nikolaos Kourkoumelis, Alexandros Polymeros, and Margaret Tzaphlidou. Background estimation of biomedical raman spectra using a geometric approach. *Journal of Spectroscopy*, 27(5-6):441–447, 2012.

[40] A. Lieber and A. Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57:1363–1367, 2011.

[41] Brooke D Beier and Andrew J Berger. Method for automated background subtraction from raman spectra containing known contaminants. *Analyst*, 134(6):1198–1202, 2009.

[42] LT Kerr and BM Hennelly. A multivariate statistical investigation of background subtraction algorithms for raman spectra of cytology samples recorded on glass slides. *Chemometrics and Intelligent Laboratory Systems*, 158:61–68, 2016.

[43] Kristian Hovde Liland, Trygve Almøy, and Bjørn-Helge Mevik. Optimal choice of baseline correction for multivariate calibration of spectra. *Applied spectroscopy*, 64(9):1007–1016, 2010.

[44] John H Mathews, Kurtis D Fink, et al. *Numerical methods using MATLAB*, volume 4. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[45] C Krafft, IW Schie, T Meyer, M Schmitt, and J Popp. Developments in spontaneous and coherent raman scattering microscopic imaging for biomedical applications. *Chemical Society Reviews*, 45(7):1819–1849, 2016.

[46] Christoph Krafft, Michael Schmitt, Iwan W Schie, Dana Cialla-May, Christian Matthäus, Thomas Bocklitz, and Jürgen Popp. Label-free molecular imaging of biological cells and tissues by linear and nonlinear raman spectroscopic approaches. *Angewandte Chemie International Edition*, 56(16):4392–4430, 2017.

[47] Kenny Kong, Catherine Kendall, Nicholas Stone, and Ioan Notingher. Raman spectroscopy for medical diagnostics: From in-vitro biofluid assays to in-vivo cancer detection. *Advanced drug delivery reviews*, 89:121–134, 2015.

[48] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, et al. Using raman spectroscopy to characterize biological materials. *Nature protocols*, 11(4):664, 2016.

[49] Ian T Jolliffe. Springer series in statistics. *Principal component analysis*, 29, 2002.

[50] T Davies and Tom Fearn. Back to basics: the principles of principal component analysis. *Spectroscopy Europe*, 16(6):20, 2004.

[51] M. Hirsch and R. Wareham et al. A stochastic model for electron multiplying charge-coupled devices - from theory to practice. *PLOS one*, 8, 2013.

[52] M.J. O'Malley and E. O'Mongain. Charge-coupled: Frame adding as an alternative to long integration times and cooling. *Optical Engineering*, 31:522–526, 1992.

[53] D. Dussault and P. Hoess. Noise performance comparison of ICCD with CCD and EMCCD cameras. *Optical Science and Technology, the SPIE 49th Annual Meeting*, pages 195–204, 2004.

[54] J.R. Janesick. *Scientific charge-coupled devices*. SPIE - The international society for optical engineering, 2000.

[55] T. Veerarajan. *Probability, statistics and random processes*. Tata McGraw-Hill Education, 2nd edition, 2002.

[56] K. Irie, A.E. McKinnon, K. Unsworth, and I.M. Woodhead. A model for measurement of noise in ccd digital-video cameras. *Measurement Science and Technology*, 19:045207, 2008.

[57] J. Biretta and M. Mutchler. Charge trapping and cte residual images in the wfpc2 ccds. *WFPC2-ISR*, 97, 1998.

[58] G.E. Healy and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *Pattern Analysis and Machine Intelligence*, 16:267–276, 1994.

[59] G.A. deVree et al. Photon-counting gamma camera based on an electron-multiplying ccd. *Nuclear Science*, 52:580–588, 2005.

[60] C.D. Mackay et al. Subelectron read noise at mhz pixel rates. *International Socety for Optics and Photonics*, 4306:289–299, 2001.

[61] L. Zhang et al. A characterization of the single-photon sensitivity of an electron multiplying charge-coupled device. *Journal of Physics B: Atomic, Molecular, and Optical Physiscs*, 42, 2009.

[62] R.N. Tubbs. *Lucky Exposures: Diffraction limited astronomical imaging through the atmosphere*. PhD thesis, 2004.

[63] A.G. Basden, C.A. Haniff, and C.D. Mackay. Photon counting strategies with low-light-level ccds. *Monthly notices of the royal astronomical society*, 345:985–991, 2003.

[64] M.J. Pelletier. Quantitative analysis using raman spectrometry. *Applied spectroscopy*, 57(1):20A–42A, 2003.

[65] J. Hynecek. Electron-hole recombination antiblooming for virtual-phase ccd imager. *Electron Devices*, 30:941–948, 1983.

[66] A.J. Berger, Itzkan, Irving, and M.S. Feld. Feasibility of measuring blood glucose concentration by near-infrared Raman spectroscopy. *Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy*, 53A:287–292, 1997.

[67] T. Dieing, O. Hollricher, and J. Toporski. *Confocal Raman Microscopy*. Springer Science and Business Media, 2011.

[68] L.T.Kerr et al. Methodologies for bladder cancer detection with raman based urine cytology. *Analytical Methods*, 8:4991–5000, 2016.

[69] C.A. Palmer and E.G. Loewen. *Diffraction Grating Handbook*. Newport Corporation New York, 2005.

[70] Bevan B Baker and Edward Thomas Copson. *The mathematical theory of Huygens' principle*, volume 329. American Mathematical Soc., 2003.

[71] EG Loewen, Maystre Nevière, and D Maystre. Grating efficiency theory as it applies to blazed and holographic gratings. *Applied optics*, 16(10):2711–2721, 1977.

[72] T.J. Harvey et al. Classification of fixed urological cells using raman tweezers. *Journal of biophotonics*, 2:47–69, 2009.

[73] Aspasia Elka, Violetta Moulia, Panagiota Spyridonos, and Nikolaos Kourkoumelis. The effect of irradiance and integration time in in vivo normal skin raman measurements assessed by multivariate statistical analysis. *Biomedical Spectroscopy and Imaging*, 5(2):217–223, 2016.

[74] G. H. Dunteman. *Principal components analyis*, volume 69. Sage, 1989.

[75] L.T. Kerr, K. Domijan, I. Cullen, and B.M. Hennelly. Applications of raman spectroscopy to the urinary bladder for cancer diagnostics. *Photonics and Lasers in Medicine*, 3:193–224, 2014.

[76] A.F. Palonpon et al. Raman and sers microscopy for molecular imaging of live cells. *Nature protocols*, 8:677, 2013.

[77] Robert H Webb. Confocal optical microscopy. *Reports on Progress in Physics*, 59(3):427, 1996.

[78] Vikram Prasad, Denis Semwogerere, and Eric R Weeks. Confocal microscopy of colloids. *Journal of Physics: Condensed Matter*, 19(11):113102, 2007.

[79] Tony Wilson. Resolution and optical sectioning in the confocal microscope. *Journal of microscopy*, 244(2):113–121, 2011.

[80] Steven L Jacques. Optical properties of biological tissues: a review. *Physics in Medicine & Biology*, 58(11):R37, 2013.

[81] Zhengyuan Tang, Sinead J. Barton, Tomas E. Ward, John P. Lowry, Michelle M. Doran, Hugh J. Byrne, and Bryan M. Hennelly. Multicomponent analysis using a confocal raman microscope. *Appl. Opt.*, 57(22):E118–E130, 2018.

[82] Harry Arthur Willis, JH Van der Maas, and Roy Gabriel Jonathan Miller. Laboratory methods in vibrational spectroscopy. 1987.

[83] Karsten Sperlich and Heinrich Stolz. Quantum efficiency measurements of (em)ccd cameras: high spectral resolution and temperature dependence. *Measurement Science and Technology*, 25(1):015502, 2014.

[84] L. Zhang and M. Henson. A practical algorithm to remove cosmic spikes in raman imaging data for pharmaceutical applications. *Applied Spectroscopy*, 61:1015–1020, 2007.

[85] P. Crow, A. Molckovsky, and N. Stone et al. Assessment of fiberoptic near-infrared raman spectroscopy for diagnosis of bladder and prostate cancer. *Urology*, 65:1126–11230, 2005.

[86] P. Jess, D. Smith, and M. Mazilu et al. Early detection of cervical neoplasia by raman spectroscopy. *International Journal of Cancer*, 121, 2007.

[87] A. Haka et al. Identifying micro-calcifications in benign and malignant breast lesions by probing differences in their chemmical composition using raman spectroscopy. *Journal of Cancer Research*, 62, 2002.

[88] W. Hill and D. Rogalla. Spike-correction of weak signals from charge-coupled devices and its application to raman spectroscopy. *Analytical Chemistry*, 64:2575–2579, 1992.

[89] G.R. Phillips and J.M. Harris. Polynomial filters for data sets with outlying or missing observations: Application to charge coupled device detected raman spectra contaminated by cosmic rays. *Analytical Chemistry*, 62:2351, 1990.

[90] F. Ehrentreich and L. Sümmchen. Spike removal and denoising of raman spectra by wavelet transform methods. *Analytical Chemistry*, 73:4364–4373, 2001.

[91] M. Soneira, R. Perez-Pueyo, and S. Ruiz-Moreno. Raman spectra enhancement with fuzzy logic approach. *Journal of Raman spectroscopy*, 33:599–603, 2002.

[92] Y. Katsumoto and Y. Ozaki. Practical algorithm for reducing convex spike noises on a spectrum. *Applied Spectroscopy*, 57:317–322, 2003.

[93] L.A. Quintero, S.D. Hunt, and M. Diem. Denoising of raman spectroscopy signals. *Research Thrust R2 Presentations*, 6, 2006.

[94] D. Zhang, K. Jallad, and D. Otz. Stripping of cosmic spike spectral artifacts using a new upper-bound spectrum algorithm. *Applied Spectroscopy*, 55:1523–1531, 2001.

[95] D. Zhang and D. Ben-Amotz. Removal of cosmic spikes from hyper-spectral images using a hybrid upper-bound spectrum method. *Applied spectroscopy*, 56:91–98, 2001.

[96] D. Zhang, J. Hanna, and D. Ben-Amotz. Single scan cosmic spike removal using the upper bound spectrum method. *Applied Spectroscopy*, 57:1303–1305, 2003.

[97] U. Cappel, I. Bell, and L. Pickard. Removing cosmic ray features from raman map data by a refined nearest neighbor comparison method as a precursor for chemometric analysis. *Applied Spectroscopy*, 64:195–200, 2010.

[98] S. Li and L. Dai. An improved algorithm to remove cosmic spikes in raman spectra for online monitoring. *Applied Spectroscopy*, 65:1300–1306, 2011.

[99] S. Mohzharov, A. Nordon, and D. Littlejohn et al. Automated cosmic spike filter optimized for process raman spectroscopy. *Applied Spectroscopy*, 66:1326–1333, 2012.

[100] J. Zhao. Image curvature correction and cosmic spike removal removal for high-throughput dispersive raman spectroscopy. *Applied Spectroscopy*, 57:1368–1375, 2003.

[101] D. Liu, H.J. Byrne, L. O'Neill, and B. Hennelly. Investigation of wavenumber calibration for raman spectroscopy using a polymer standard. *Optical Sensing and Detection V: International Society for Optics and Photonics*, 10680:1080627, 2018.

[102] Claire Molony, Jennifer McIntyre, Adrian Maguire, Roya Hakimjavadi, Denise Burtenshaw, Gillian Casey, Mariana Di Luca, Bryan Hennelly, Hugh J. Byrne, and Paul A. Cahill. Label-free discrimination analysis of de-differentiated vascular smooth muscle cells, mesenchymal stem cells and their vascular and osteogenic progeny using vibrational spectroscopy. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1865(2):343 – 353, 2018.

[103] P. Heraud, B.R. Wood, J. Beardall, and D. McNaughton. Effects of pre-processing of raman spectra on *in vivo* classification of nutrient status of microalgal cells. *Journal of Chemometrics*, 20:193–197, 2006.

[104] F.W. Scholz. *Maximum Likelihood Estimation*. Encyclopedia of statistical sciences, 1985.

[105] H.C. Burger, B. Schölkopf, and S. Harmeling. Removing noise from astronomical images using a pixel-specific noise model. *IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2011.

[106] M. Gomez-Rodriguez, J. Kober, and B. Schölkopf. Denoising photographs using dark frames optimized by quadratic programming. *IEEE International Conference on Computational Photography (ICCP)*, pages 1–9, 2009.

[107] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, MA, 1985.

[108] R. Kiselev, I.W. Schie, S. Aškrabić, C. Krafft, and J. Popp. Design and first applications of a flexible raman micro-spectroscopic system for biological imaging. *Biomedical Spectroscopy and Imaging*, 5:115–127, 2016.

[109] Z. Movasaghi, S. Rehmen, and I.U. Rehmen. Raman spectroscopy of biological tissues. *Applied spectroscopy*, 42:493–541, 2007.

[110] Hai Liu, Zhaoli Zhang, Jianwen Sun, and Sanya Liu. Blind spectral deconvolution algorithm for raman spectrum with poisson noise. *Photon. Res.*, 2(6):168–171, 2014.

[111] Tingting Liu, Hai Liu, Zengzhao Chen, and Alan M. Lesgold. Fast blind instrument function estimation method for industrial infrared spectrometers. *IEEE Transactions on Industrial Informatics*, PP:1–1, 2018.

[112] Yao Tian and Kenneth Burch. Automatic spike removal algorithm for raman spectra. *Applied Spectroscopy*, 70:861–871, 2016.

[113] Alex Cao, Abhilash K Pandya, Gulay K Serhatkulu, Rachel E Weber, Houbei Dai, Jagdish S Thakur, Vaman M Naik, Ratna Naik, Gregory W Auner, Raja Rabah, et al. A robust method for automated background subtraction of tissue fluorescence. *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering*, 38(9):1199–1205, 2007.

[114] Hugh J Byrne, Peter Knief, Mark E Keating, and Franck Bonnier. Spectral pre and post processing for infrared and raman spectroscopy of biological tissues and cells. *Chemical Society Reviews*, 45(7):1865–1878, 2016.

[115] Peter Lasch. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems*, 117:100–114, 2012.

[116] T Dieing, O Hollricher, and J Toporski. Confocal raman spectroscopy, 2010.

[117] L.T. Kerr et al. Applications of raman spectroscopy to urinary bladder for cancer diagnostics. *Photonics and Lasers in Medicine*, 3:2193–0643, 2014.

[118] Franck Bonnier, Damien Traynor, Padraig Kearney, Colin Clarke, Peter Knief, Cara Martin, John J O'Leary, Hugh J Byrne, and Fiona Lyng. Processing thinprep cervical cytological samples for raman spectroscopic analysis. *Analytical Methods*, 6(19):7831–7841, 2014.

[119] Tim J Harvey, Caryn Hughes, Andrew D Ward, Elsa Correia Faria, Alex Henderson, Noel W Clarke, Mick D Brown, Richard D Snook, and Peter Gardner. Classification of fixed urological cells using raman tweezers. *Journal of biophotonics*, 2(1-2):47–69, 2009.

[120] Ronald OP Draga, Matthijs CM Grimbergen, Peter LM Vijverberg, Christiaan FP van Swol, Trudy GN Jonges, J Alain Kummer, and JLH Ruud Bosch. In vivo bladder cancer diagnosis by high-volume raman spectroscopy. *Analytical chemistry*, 82(14):5993–5999, 2010.