Machine Learning for Spectroscopy in Biological Applications

Miao Zhao

Master of Science

School of Informatics

University of Edinburgh

2023

Abstract

The prevalence of pulmonary nodules is on the rise due to increased rates of smoking, prompting an urgent need for rapid and accurate disease detection methods. Raman spectroscopy has emerged as a promising solution for in vivo assessment of living tissue. However, it confronts several hurdles. One particular note is the dominance of fluorescence intensity over Raman scattering signals, a significant obstacle that impedes the extraction and analysis of Raman signals.

This project seeks to address these concerns by using an open-source ssersBayes package[28] within the R programming. Leveraging an existing Bayesian hierarchical model, we employ it to analyse our collected data. Additionally, a series of experiments are conducted to evaluate the model's fitting performance under variations in prior information. This approach aims to mitigate background interference, enhancing the qualitative and quantitative analysis of spectra for improved disease detection and assessment.

The conducted experiments reveal that the combination of prior knowledge encompassing peak locations, the median of amplitudes and scales, and the standard derivation of amplitudes and scales, significantly enhances the reliability and interpretability of the obtained results. Additionally, the paper delves into the model's flexibility by introducing the inaccurate prior peak locations. Despite the overall fitting effect of the models appears to be notably compromised, the model tend to capture the spectral signal more effectively when intervals are set at 10.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Miao Zhao*)

Acknowledgements

Words cannot express my gratitude to my supervisor, Dr. Sohan Seth, for his insightful guidance and patience. Our collaborative brainstorming during meetings has been instrumental in hastening the progress of my paper.

Equal appreciation extends to Dr. Daga Panas, whose expert assistance in addressing experimental challenges and providing invaluable suggestions for my draft dissertation resonates deeply with me. I still remember the initial visit to Dr. Daga Panas's office, where her welcoming demeanor and encouraging words solidified my commitment to pursuing this captivating, yet unfamiliar, topic. Despite the demand for meetings twice a week, which consumes her considerable time, her patience in resolving my problems and giving guidance when I was confused is remarkable. Beyond academia, she also cares about my daily life, which creates an atmosphere of ease and joy during our meetings.

Furthermore, I would also like to thank my parents, who have given me unconditional support through the whole process of my postgraduate study.

And finally, I owe a debt of gratitude to my friends. During the last month in Edinburgh, an unexpected episode of food poisoning left me on the brink of fainting at home. Fortunately, my friends came to my aid, accompanying me to the hospital and providing much-needed support during the injections. Without their help, I can't recover from this serious illness.

Table of Contents

Chapter 1

Introduction

Raman spectroscopy, named in honor of the Indian scientist C.V. Raman, is regarded as a powerful technique for detecting and analyzing molecular structures [8]. It generates Raman spectrum using the scattering of light, which provides valuable insights into molecular vibrations and rotations. By analyzing the frequency shifts in the scattering spectrum of the light, researchers can extract a wealth of information regarding molecular properties and inter-molecular bond energies.

These unique capabilities of Raman spectroscopy have led to its widespread application in diverse scientific disciplines, including chemistry, physics, biology, and materials science. In chemistry, it enables the characterization of molecular structures, conformational changes, and reaction kinetics [27]. In the realm of physics, it proves indispensable for probing condensed matter, vibrational modes, and lattice dynamics [20]. Additionally, in the fields of biology and materials science, Raman spectroscopy serves as a powerful tool for identifying biomolecules and characterizing materials [29].

However, its applicability has been seriously limited to specific applications by the presence of a complex background. Recognizing the cruciality of Raman spectroscopy and its potential impact, researchers have turned their attention towards improving background elimination techniques for spectral data, obtaining considerable attention. Several methods, such as those proposed by Mazilu [26], Gebrekidan [11], and Zhang [35], have been developed to effectively address the issue. While these methods demonstrate efficacy, they are not without their drawbacks, making them suboptimal solutions. Thus, a pressing need exists to explore novel approaches that bridge this gap and overcome the limitations of existing methods.

This paper is dedicated to the exploration of an established Bayesian model and a sequential Monte Carlo algorithm. Our objective is to assess the model's adaptability

through comprehensive experiments, probing their performance across a spectrum of conditions that diverge from those presented in previous literature.

1.1 Motivation and Problem Statement

The development of spectroscopy that could deal with the high complexity of biological molecules and their unique functions, has called for a more cost-effective alternatives. Such alternatives should be capable of handling large datasets, enabling the fitting of computational models for further research and analysis.

While traditional techniques like immunofluorescence have historically served as the tool for molecular analysis, they come with certain limitations. These methods often involve complex and time-consuming procedures, including labelling and tagging molecules of interest [30]. In addition, they are susceptible to issues such as photobleaching and limited spectral resolution. In contrast, Raman spectroscopy presents a label-free approach that alleviates these drawbacks. By directly probing the inherent vibrational properties of molecules, Raman spectroscopy eliminates the need for specific molecular labeling, reducing the potential for alterations to the sample.

However, Raman spectroscopy has emerged as a promising method that offers significant advantages in meeting these requirements. It provides a non-invasive and non-destructive [28] approach to scientific investigation, which allows for minimal disruptions to the biological sample being analyzed. Moreover, this technique provides detailed molecular-level information, with minimal sample preparation, no labeling, and is water-insensitive, making it suitable for assessing the physiological and molecular signatures of tissue. Recent advancements in Raman imaging, such as surface-enhanced Raman scattering(SERS), have successfully identified and discriminated bacteria [16].

The merits of Raman spectroscopy have captivated scientists, prompting dedicated efforts to address its challenges. Notably, a significant challenges in quantitative analysis of Raman spectroscopy, particularly in in-vivo applications, is the weak intensity of the Raman scattering signal when concurrent sample fluorescence and background scattering from the optical fiber are present[17]. This challenge becomes more pronounced when visible light is employed, which can significantly impact the accuracy and sensitivity of Raman measurements. The phenomenon can be attributed, in part, to the low cross-sectional scattering of Raman spectroscopy. As a result, the fluorescence intensity typically surpasses the Raman scattering signal by several orders of magnitude, and the spectrum is completely dominated by fluorescence, posing a substantial obstacle

to capturing and extracting Raman information effectively. An example depicting the challenge within Raman spectrum is presented in Figure 1.1. Evidently, the data appears noisy, and the peaks lack distinct clarity, thereby augmenting the difficulty of accurately fitting the Raman peaks.

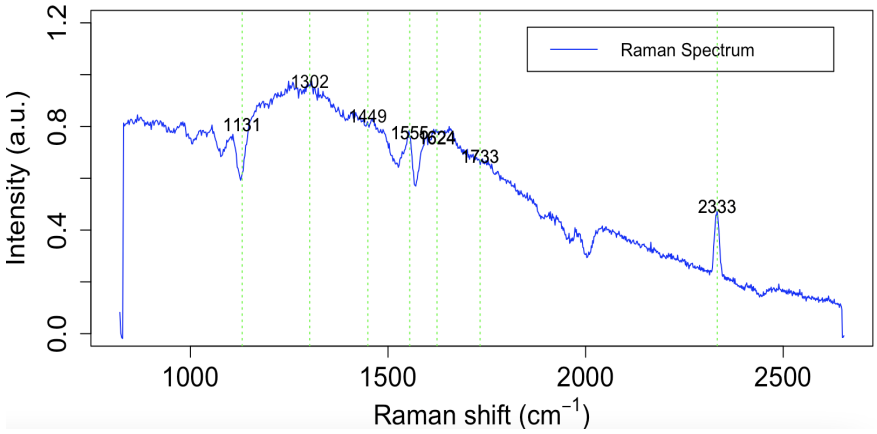

Figure 1.1: Raman spectrum of lung tissue

Raman scattering and fluorescence have fundamental differences. Raman scattering is characterized by the weak interaction between a molecule and an incident photon, leading to a molecule being elevated to a transient, or ”virtual,” excited state. After relaxation by a radiative transition, the molecule then returns to a different vibrational energy level within its electronic ground state [4]. In contrast, fluorescence is a significantly more intense phenomenon. It combines the absorption of light by a molecule, which attributes excitation to a higher electronic energy level. Subsequently, the molecule undergoes vibrational relaxation followed by a radiative transition, returning to the ground energy state.

Therefore, noise removal is a crucial step in data processing, particularly in the context of fluorescence background removal. The strong fluorescence signal in Raman spectroscopy directly affects the accuracy and sensitivity of Raman measurements. While several methods have been developed, there is currently no optimal approach. Existing techniques, involving hardware and software, have been devised to improve fluorescence background removal. These include polynomial fitting [23] [2], wavelet transform [22] [37], and the use of first- and second-order derivatives [34]. While these approaches have shown merits in addressing the issue, they each come with limitations.

The polynomial fitting model, for example, can be effective in removing background noise. However, it often requires user intervention, making the process time-consuming and potentially subjective. As for wavelet transform-based methods, although they

can be powerful for processing signals with low-frequency components, they are less effective when applied to other types of signals, for which more sophisticated methods should have been explored.

Thus, there is a need for further research and development to explore more robust and efficient approaches to remove fluorescence background. Such advances should advance accuracy and reliability so that enhance analysis and interpretation of molecular information.

1.2 Objectives and Methodology

In our paper, we first focus on handling the challenge of extracting Raman signals from a fluorescence background, particularly when involving multiple spectra.

To accomplish this goal, the project utilizes an existing Bayesian model [28] designed for extracting Raman signals from a fluorescence background, to analyse our collected data, including cyclohexane, sesame oil, and ex vivo human lung tissue samples. In this step, we combined prior knowledge like peak locations, the median of scales and amplitudes, and the standard derivation of scales and amplitudes, with our model to refine the model's performance.

In order to enhance the model's flexibility and sensitivity, we then introduce incorrect prior peak locations into the hierarchical framework to assess the model's fitting effect. This situation closed to the real-world applications, offering valuable insights for refining the model's performance.

Chapter 2

Background

The task of removing fluorescence background from Raman signal observations has led to the application of various approaches based on instrumental, experimental, and computational methods.

2.1 Research Gaps

Previous work has explored various techniques for florescence removal in Raman spectra, including wavelet transform, shifted-excitation wavelength method and principal component analysis. These approaches could be classified into three groups based on the number of shifted measurements involved: single spectrum, two spectra, and multiple spectra. Each group poses challenges and opportunities for improvement, prompting extensive research to be conducted to explore their merits and drawbacks. Additionally, the methods related to two spectra and multiple spectra based on a fundamental knowledge: while fluorescence emission remains relatively stable under minor changes, the Raman spectrum performs shifts corresponding to alterations in excitation photon energy.

2.1.1 Single Spectrum

Hu [15] introduced a novel method based on the power of multi-resolution through the wavelet transform. By decomposing signals into localized contributions characterized by a scale parameter, the approach holds great promise for fluorescence removal. Nevertheless, the wavelet transform process may also lead to the loss of significant spectra information. Despite its ability to highlight localized features, some crucial

details might be obscured during the decomposition process, impacting the overall results. Furthermore, the technique also relies on the assumption that the background is well separated in the transformed domain from the signal.

In 2009, an improved intelligent background-correction algorithm [36] was proposed, representing a significant step forward in spectral data processing. It encompassed a three-pronged approach to achieve the goal of background correction: employing the wavelet transform method for accurate peak location, the signal-to-noise method for peak-width estimation, and penalized least squares for background fitting. The combination of techniques promised to deliver a more generalized and accurate background correction, addressing a wide range of spectroscopic applications. Despite the impressive potential to enhance the results, it is essential to acknowledge the complexity and computation time associated with this technique.

Additionally, another novel algorithm [35] named adaptive iteratively reweighted Penalized Least Squares(airPLS) has emerged as an improvement for the intelligent background-correction algorithm. The key innovation of airPLS lies in its iterative approach, which iteratively adjusts the weights of sum squares errors between the fitted baseline and the original signals. This algorithm takes on the role of a smoothing algorithm since it hasn't setting zeros to the weight vector. By iteratively refining the weight, it fine-tunes the performance, leading to accurate baseline estimates. The algorithm also leveraged the penalized least squares technique [13] to develop a smoother process. In this way, the penalized least squares algorithm can be categorized as a method that achieves smoothness through a balance between squares fidelity to the original data and the roughness of the fitted data[35]. While this algorithm shows promise, it falls short of being fully optimal as it only considers certain facets of spectral behavior.

2.1.2 Two Spectra

One of the ingenious applications of Kasha's rule [18] has paved the way for a fluorescence removal technique known as the shifted-excitation wavelength method [11]. Based on the knowledge introduced before, they subtracted two spectra obtained with different excitation wavelengths and normalized them. Then, it effectively reduced the impact of fluorescence background. However, it might not represent the optimal choice. While it capitalizes on the spectral behavior of fluorescence and Raman signals, it does not leverage their unique spectral characteristics. Moreover, it is crucial to acknowledge that subtracting two noisy signals may also amplify the noise in the estimated difference

spectrum, leading to broader peaks. This method may compromise the accuracy and precision of the results. Hence, further research should be conducted to improve this drawback.

2.1.3 Multiple Spectra

Shixuan et al. [14] introduced an improved asymmetric least squares method designed for correcting the baseline of Raman spectra. Their method concerns smoothness of derivative, leading to enhanced baseline correction performance. However, it is difficult to interpret the preprocessed spectra. Feng et al. [10] proposed an improved polynomial fitting method with an automatic threshold, implementing an iterative process to adjust the signal peaks. Through step by step iterations, they aimed to find the best estimation. Moreover, a baseline subtraction model was proposed in 1999 [31], which utilized robust local regression estimation and minimized the need for human intervention. It improved the sensitivity to outliers compared to other conventional approaches.

Additionally, combining Principal Component Analysis(PCA) with modulation techniques discussed in two spectra, holds tremendous promise, offering several key advantages. PCA is primarily employed to discern the most significant basis for reexpressing a given dataset. This fresh basis unveils hidden data structures and effectively eliminates noise [21].

The author[26] extracted the maximal variation from a set of N spectra through the first principal component, which denotes the eigenvector with the largest eigenvalue, denoted by μ, from the covariance matrix C_{ik}. The technique effectively determines the major axis of rotation in a multidimensional object comprising scattered points, each corresponding to a spectrum in an n-dimensional space, where n represents the number of measured wavenumbers.

When compared to alternative mathematical approaches like Least Squares fit(LSQ), Fourier Filtering(FF), and Standard Deviation(SD) analysis, Principal Component Analysis stands out in terms of signal-to-noise ratio, reducing noise in Raman signals. Leveraging the ability to discern meaningful variations from noise, PCA excels at enhancing the quality of Raman spectra, significantly improving results. One of the advantages of PCA is its independence from a synchronization process, minimizing the need for user intervention during analysis. This feature lends itself well to real-time applications.

Chapter 3

Methodology

This chapter is dedicated to outlining the research methodology employed in my paper. This chapter consists five parts, which are experimental data, data processing, models, reliability, and code of ethics. The primary aim of our project is to apply and adapt an existing model to our novel data, thereby examining its versatility and effectiveness.

3.1 Experimental Data

The data used in our project is the same reported in [17], collected using a custom-built optical setup. The data used in my study has undergone ethics assessment for a previous study protocol, and that covered also future re-use of data, so no project-specific ethics assessment was needed.

We investigated three distinct samples:

1) Cyclohexane: Cyclohexane was choosen owing to its well known Raman spectrum and absence of fluorescence background. An example of the observed spectrum of cyclohexane is displayed in Fig 3.1.

2) Sesame oil: The Raman spectrum of sesame oil exhibits prominent, well known peaks, but it is also characterized by a high level of fluorescence background. An example observed spectrum of sesame oil is illustrated in Fig 3.2.

3) Lung tissue: This unique data was obtained from the Royal Infirmary of Edinburgh (NHS Lothian BioResource, reference 15/ES/0094). An ex vivo human lung tissue sample was collected from a patient diagnosed with suspected or confirmed lung cancer. An example observed spectrum of tissue is shown in Fig 3.3.

Furthermore, we standardized the data by dividing it with the maximum value to eliminate disparities between features. To reduce experimental variability, We relied

on 10 experiments results on each of the three samples and computed their average. Moreover, we determined the shift value for the x-axis using the formula:

$$shift = (\frac{1}{excitationwavelength} - \frac{1}{wavelength}) * 1e7 \tag{3.1}$$

where excitation wavelength is 785.423297 nm.

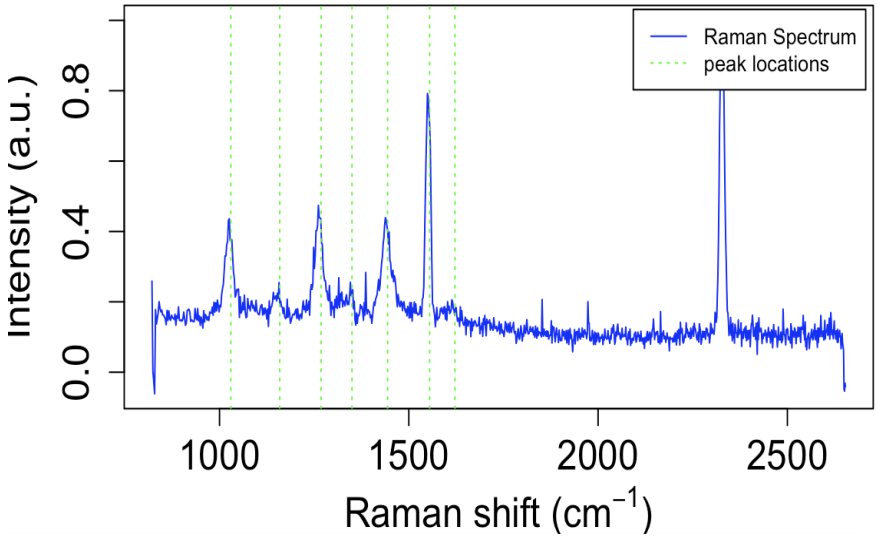

Figure 3.1: Raman spectrum of Cyclohexane

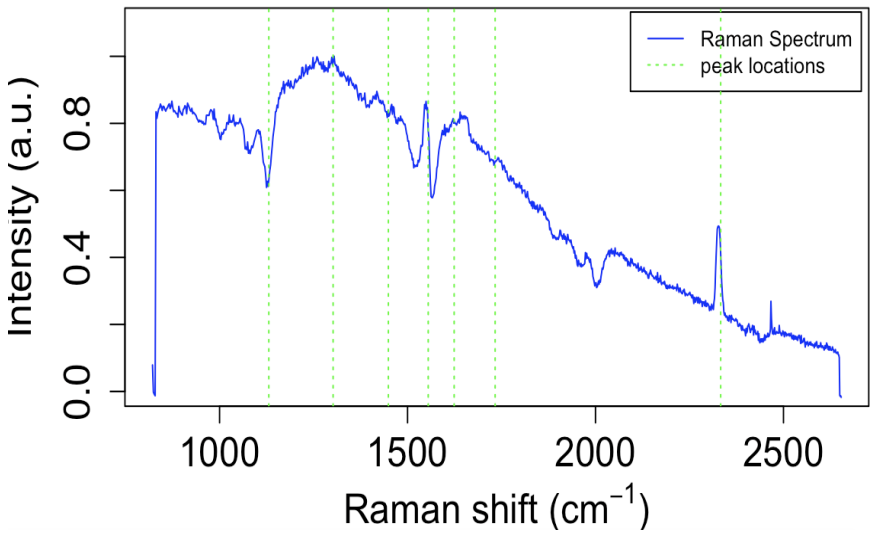

Figure 3.2: Raman spectrum of Sesame Oil

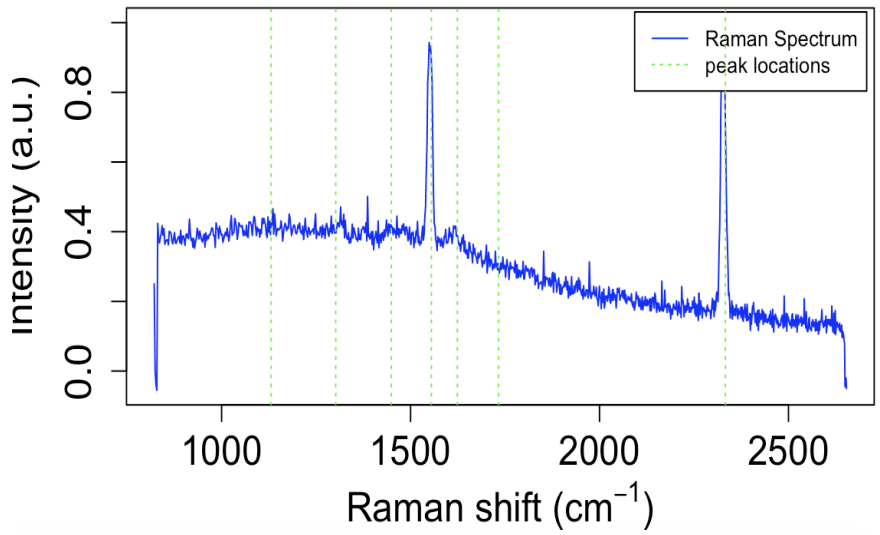

Figure 3.3: Raman spectrum of Lung Tissue

3.2 Models

In this section, we delve into the introduction of our models. Figure 3.7 provides a visual representation of how our model is applied, and we will proceed to elaborate on the intricacies of two key components (Hierarchical Model and Bayesian Computation)in the subsequent subsections.

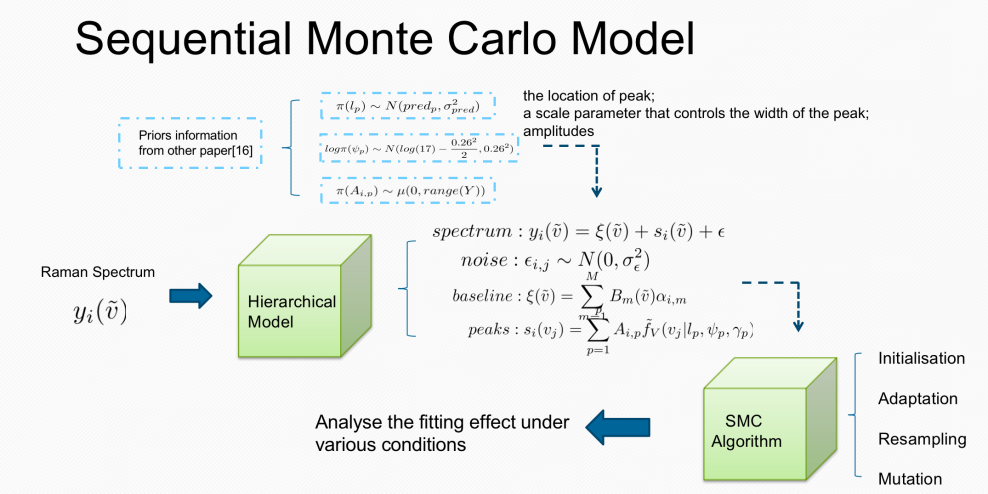

Figure 3.4: Flow Chart of Our Model

3.3 Theoretical Foundation

In this paper, we make an assumption that our signal data can be represented using continuous functions. Additionally, a foundation of our model lies in the discretization of the Raman spectrum into a multivariate observation, exhibits high collinearity [28]. This crucial basis enables us to employ our hierarchical model, allowing for further analysis and interpretation.

De Rooi and Eilers [6] have introduced a baseline model that employs P-splines, which combines B-splines with a penalty to tune smoothness. This model is complemented by a mixture model consisting of two component distributions, one dedicated to modeling noise and the other focused on capturing the peaks. They assumed that the noise followed a Gaussian distribution, while the peaks were modeled using a uniform distribution. This representation of peaks as a continuous function aligns with the properties of Raman spectroscopy and can be applied effectively in our project. To meet the Brownian motion, we apply a radial basis function(RBF):

$$f_G(\nu_j|l_p,\psi_p) \propto exp - \frac{(\nu_j - l_p)^2}{2\psi_p^2} \tag{3.2}$$

where ν_j is the jth wavenumber in the spectrum, l_p denotes the location of peak p, and ψ_p represents a scale parameter that determines the width of the peak. The wavenumber $\tilde{\nu}$ measured in units of reciprocal centimeters (cm^{-1}). It is derived from the corresponding wavelengths (λ in nm) in unit length ($\tilde{\nu} = \frac{10^7}{\lambda}$). Peak location l, peak amplitudes γ, and the scale parameter ψ, play crucial roles in the task of fluorescence removal. Peak location refers to the specific position where a peak occurs in a spectrum. It represents the horizontal axis value corresponding to the peak's maximum intensity. Peak amplitude is the magnitude of a peak, representing the vertical axis value at the peak's maximum point. Peak scale, also known as peak width, determines the width of a peak, influencing its broadness and shape.

In addition, it is important to consider collisional broadening, which occurs due to particle collisions, and may cause a shift in the energy of emitted photons. This phenomenon follows a Lorentzian function:

$$f_G(\nu_j|l_p,\psi_p) \propto \frac{\gamma_p^2}{(\nu_j - l_p)^2 + \gamma_p^2} \tag{3.3}$$

Matthew T et al. [28] subsequently proposed a Vogit function which is a combination

of a RBF and a Lorentzian function, to accurately represent the spectral process:

$$f_V(\nu_j) = (f_G * f_L)(\nu_j) = \int_{-\infty}^{+\infty} f_G(\tilde{\nu}) f_L(\nu_j - \tilde{\nu}) d\tilde{\nu}$$

Wertheim and Dicenzo [33] also proposed a pseudo-Voigt function, using additive Gaussian-Lorentzian function to estimate the value:

$$f_V(\nu_j) \approx \tilde{f}_V(\nu_j | l_p, \psi_p, \gamma_p) = \eta_p f_L(\nu_j | l_p, \gamma_p) + (1 - \eta_p) f_G(\nu_j | l_p, \psi_p)$$

where η_p could be regarded as an average factor: $\eta_p = 0$ equal to a RBF, while $\eta_p = 1$ equal to a Lorentzian function.

3.3.1 Hierarchical Model

Our approach involves decomposing the Raman spectrum into three components: spectral signature $s_i(\tilde{\nu})$, baseline $\xi(\tilde{\nu})$, and additive white noise ε. This decomposition grants us a comprehensive understanding of the underlying structure and captures essential features within our data. The hyperspectral observation $y_i(\tilde{\nu})$ can be represented as:

$$y_i(\tilde{\nu}) = \xi_i(\tilde{\nu}) + s_i(\tilde{\nu}) + \varepsilon \tag{3.4}$$

where multiple observations can be expressed as a matrix $[Y]_{1:n_y, 1:n_{\tilde{\nu}}}$.

The first components, $s_i(\tilde{\nu})$ represents the spectral signature and holds significant information about the Raman peaks. We estimated this component by utilizing pseudo-Voigt functions [9]:

$$s_i(\tilde{\nu}) = \sum_{p=1}^{P} A_{i,p} \tilde{f}_V(\nu_j | l_p, \psi_p, \gamma_p) \tag{3.5}$$

where $A_{i,p}$ denotes the amplitude of peak p in the ith observation. While Moores[28] proposed incorporating the molecule's concentration into the model, it is not the primary interest of our paper, as our experiments were conducted under consistent concentrations.

The second component is the baseline $\xi(\tilde{\nu})$. In our paper, we utilized a penalised B-spline [7] as the baseline:

$$\xi(\tilde{\nu}) = \sum_{m=1}^{M} B_m(\tilde{\nu}) \alpha_{i,m} \tag{3.6}$$

where B_m are the basis functions, M is the total number of splines, and $\alpha_{i,m}$ are the coefficients of the baseline for the ith observation.

Table 3.1: Average log marginal likelihood for models with varied knots parameter

knots	average log marginal likelihood
10	13.65
30	16.77
50	14.77
100	14.21

The B-spline presents an appealing approach for nonparametric modeling, serving to generate smooth curves. A B-spline composed of polynomial segments, an illustration of a B-spline of degree 1 is depicted in figure 3.5. On the left side of Figure 3.5, the knots are x_1, x_2, and x_3. One segment is from x_1 to x_2, while the other is from x_2 to x_3. On the right side, there are three B-splines, overlapping with each other.

However, determining the optimal number of knots can be challenging: an abundance of knots can result in overfitting, while an insufficient number can lead to underfitting. To address this, we conducted a series of experiments to determine the knots' value. Through Table 3.1, we selected the best number of knots as 30, as it corresponds to the highest average log marginal likelihood.

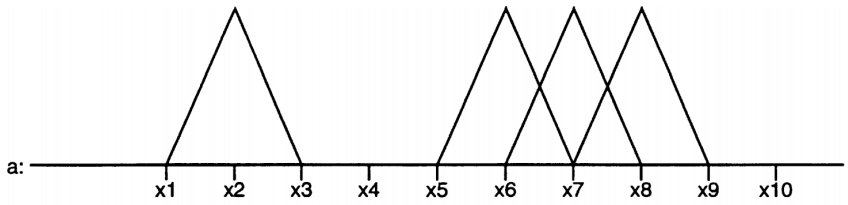

Figure 3.5: An example of B-spline of degree 1[7]

The smoothing parameters in the model are used to make the fitting curve better fit the original complex curve. Similarly, the selection of the smoothness value is determined through the computation of the average log marginal likelihood.

Additionally, we assume that the last component represents the additive white noise ε, with zero mean and constant variance:

$$\varepsilon_{i,j} \sim N(0, \sigma_\varepsilon^2) \tag{3.7}$$

To satisfy this assumption, the model also introduced autocorrelated errors [3]. These adjustments were implemented to ensure the accuracy of our model.

Table 3.2: Average log marginal likelihood for models with varied smoothness parameter

smoothness value	average log marginal likelihood
0.5	4.96
1	5.30
10	16.77
20	6.83

3.3.2 Bayesian Computation

In our paper, we employed an existing ssersBayes package which implements the algorithm described in detail in Moores et al.[28]. The core of this Bayesian model lies a framework for sampling from a target variable with a probability distribution $\pi(\cdot)$ through the generation of a Markov chain. This algorithm designed to iteratively update a population of weighted particles, denoted as $[\Theta]_{q=1}^{Q}$ through a series of intermediate target distributions. To begin with, the weighted particles is generated based on priors information:

$$\pi_0(\Theta) = \pi(l)\pi(A)\pi(\psi)\pi(\gamma)\pi(\alpha)\pi(\sigma_\varepsilon^2) \tag{3.8}$$

The SMC algorithms comprise four stages: initialisation, adaptation, resampling, and mutation. In the initial phase, the weighted particles and the measure of effective sample size(ESS) [25] are initialised. One challenge associated with SMC methods is the inherent autocorrelation exhibited by chain samples [12].Consequently, the sample size plays a crucial role within the Bayesian model, especially when computing the estimated log-likelihood for the model [1].

Adaptation and resampling stages carry significance within this algorithm. During the adaptation stage, it computes each intermediate distribution taking advantages of a marginal likelihood and a power parameter $k_t \in [0,1]$. This step is grounded in the assumption of additive Gaussian noise, rendering the likelihood as follows:

$$p(y_i(\tilde{\nu})|\Theta) \sim \pi_{\nu_j \in \tilde{\nu}} N(y_{i,j}; \xi_i(\nu_j) + s_i(\nu_j), \sigma_\varepsilon^2) \tag{3.9}$$

To simplify the formula, the algorithm uses conjugate priors to obtain a marginal likelihood with the parameters of components of the spectral signature: $p(y_i(\tilde{\nu})|l, A, \psi, \gamma)$.

Our main focus in this step is to compute the marginal likelihood at each intermediate distribution. The algorithm desires to raise the probability by a power $k_t \in [0,1]$:

$$\pi_t(l, A, \psi, \gamma | y_i(\tilde{\nu}) \propto p(y_i(\tilde{\nu}) | l, A, \psi, \gamma)^{k_t} \pi_0(l, A, \psi, \gamma) \tag{3.10}$$

Consequently, the marginal likelihood can be regarded as a metric to evaluate the models' fitting effect.

The resampling stage begins when the effective sample size(ESS) descends below the threshold value of Q/2. During this phase, the weighted particles undergo resampling, a process determined by the multinomial distribution[19]. In pursuit of minimizing the Monte Carlo variance, the algorithm employs the residual resampling method [24].

However, this resampling step introduces the possibility of duplicates among weighted particles. To alleviate this concern, the mutation step intervenes. This phase entails updating the parameters using a Markov chain Monte Carlo(MCMC) kernel, as represented by the formula(3.10). Employing a random walk Metropolis step[32], it effectively updates the parameters of the peak.

Chapter 4

Data Analysis and Findings

4.1 Point Estimates

In this section, our objective is to analyze the fitting models' effect by evaluating the fitting plots and other metrics. Specifically, Figures 4.1 to 4.3 showcase the best-fitting models, while Figures 4.4 to 4.6 illustrate the Credible Intervals(CI) for the signatures and baselines of average spectra. These visualizations provide valuable insights into the performance and reliability of our model. Additionally, Tables 4.1 to 4.4 illustrate 95 % highest posterior density(HPD) intervals for the regression coefficients β_p and the average value of log marginal likelihood.

4.1.1 Model Fitting

For figures 4.1 to 4.6, we observe that sesame oil data stands out as notably noisier in comparison to the other two one. The sesame oil dataset exhibits a high degree of volatility, characterized by an abundance of peaks that bear resemblance but lack authenticity. The model is hard to distinguish the peaks albeit given the prior information. As a result, the fitting outcome of this data is notably subpar when contrasted with the other plots. The model encounters difficulty in precisely fitting the peaks present. Conversely, our models demonstrate superior fitting performance on the cyclohexane and lung tissue.

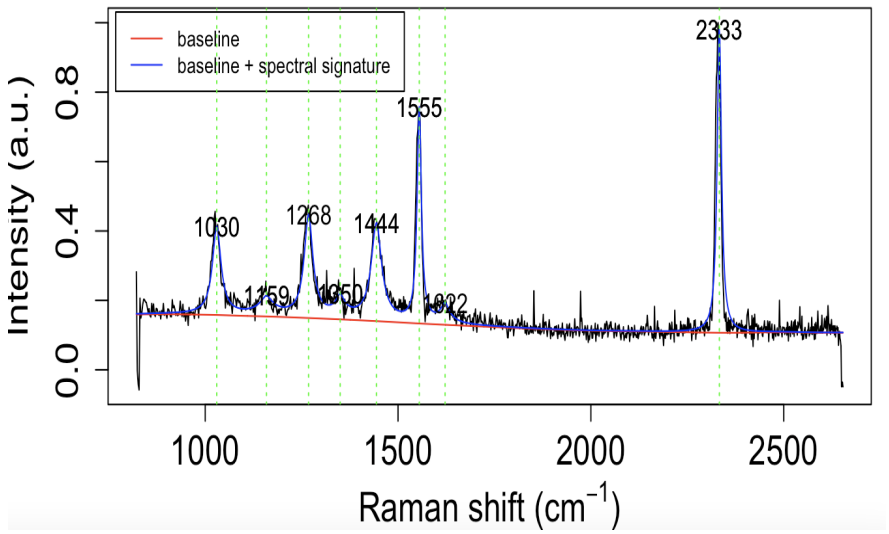

Figure 4.1: Point estimates of the baseline and signature for Cyclohexane

Additionally, Figures 4.2, 4.4, and 4.6 showcase the Credible Intervals(CI) of the models. The primary purpose of these intervals is to describe and present the uncertainty inherent in the estimation of unknown parameters. Within our study, we calculate the 95% credible interval, which is the central portion of the posterior distribution containing 95% of the values.

Comparing the CI results, a distinct pattern emerges: those for Cyclohexane and Lung Tissue, as depicted in plots 4.2 and 4.6, exhibit narrower intervals in contrast to CI for Sesame Oil. This outcome signifies that the fitting of Cyclohexane and Lung Tissue models carries more credibility due to the confinement of the estimated parameter values within these tighter intervals.

Table 4.1: 95% highest posterior density(HPD) intervals for the regression coefficients β_p (inverse nanomolar, nM^{-1}) for Cyclohexane

$l_p(cm^{-1})$	$\beta_p(nM^{-1})$
1030	[0.02, 0.09]
1159	[0.25, 0.33]
1268	[0.01, 0.30]
1350	[0.16, 0.31]
1444	[0.47, 0.70]
1555	[0.03, 0.12]
1622	[0.51, 0.94]

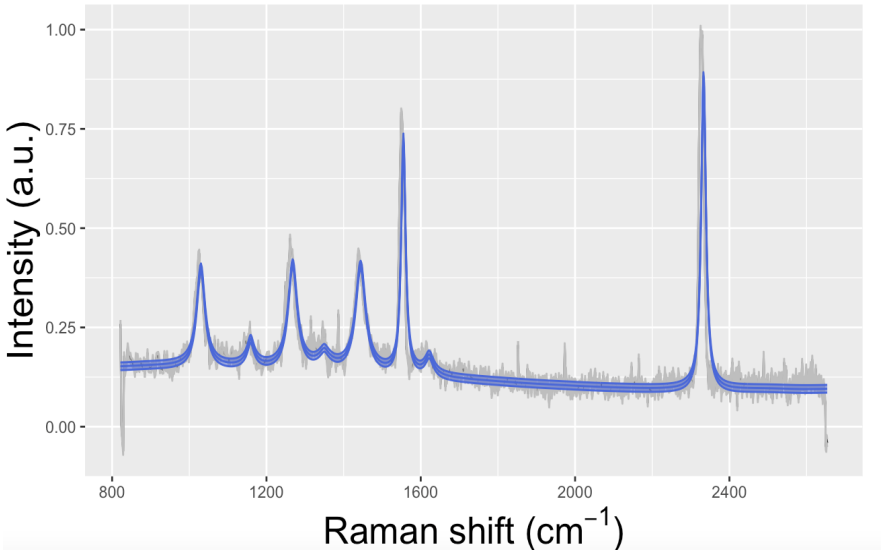

Figure 4.2: Credible Intervals(CI) for the peaks and baselines of average Cyclohexane spectra

Moreover, Tables 4.1 to 4.3 depict the 95% highest posterior density interval(HDI), where points within this interval carry a higher probability density than those outside it. The HDI value is employed as the Credible Interval(CI) in Figures 4.2, 4.4, and 4.6. Notably, the CI for Sesame Oil appears broader compared to the others, illustrating its suboptimal fitting effect.

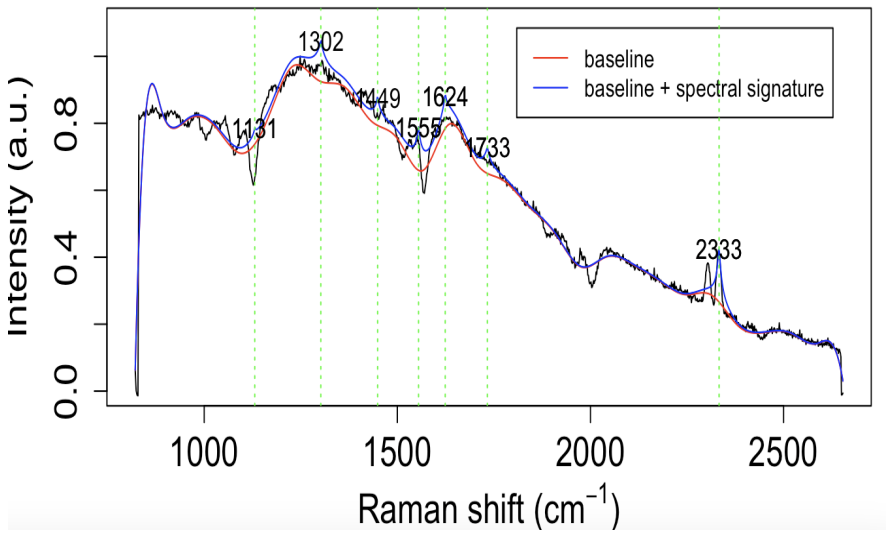

Figure 4.3: Point estimates of the baseline and signature for Sesame Oil

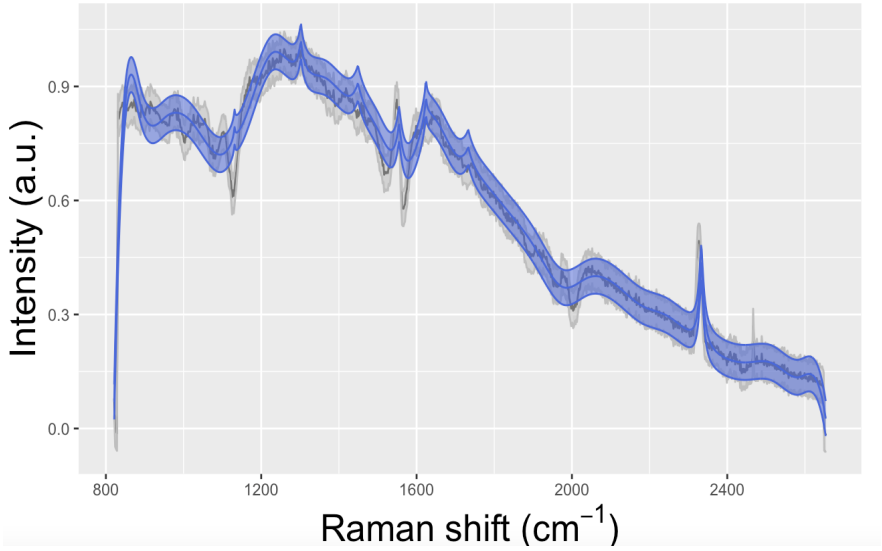

Figure 4.4: Credible Intervals(CI) for the peaks and baselines of average Sesame Oil spectra

Table 4.2: 95% highest posterior density(HPD) intervals for the regression coefficients β_p (inverse nanomolar, nM^{-1}) for Sesame Oil

$l_p(cm^{-1})$	$\beta_p(nM^{-1})$
1030	[0.01, 0.41]
1159	[0.02, 0.82]
1268	[0.04, 0.69]
1350	[0.00, 0.49]
1444	[0.02, 0.35]
1555	[0.02, 0.51]
1622	[0.02, 0.50]

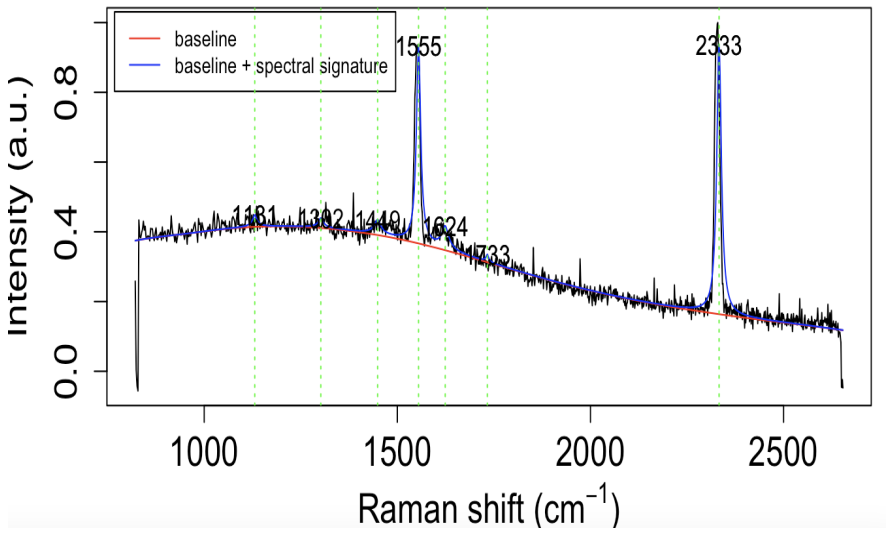

Figure 4.5: Point estimates of the baseline and signature for Lung Tissue

Table 4.3: 95% highest posterior density(HPD) intervals for the regression coefficients β_p (inverse nanomolar, nM^{-1}) for Lung Tissue

$l_p(cm^{-1})$	$\beta_p(nM^{-1})$
1131	[0.00, 0.14]
1302	[0.01, 0.50]
1449	[0.01, 0.08]
1555	[0.46, 0.63]
1624	[0.04, 0.12]
1733	[0.00, 0.05]
2333	[0.77, 0.89]

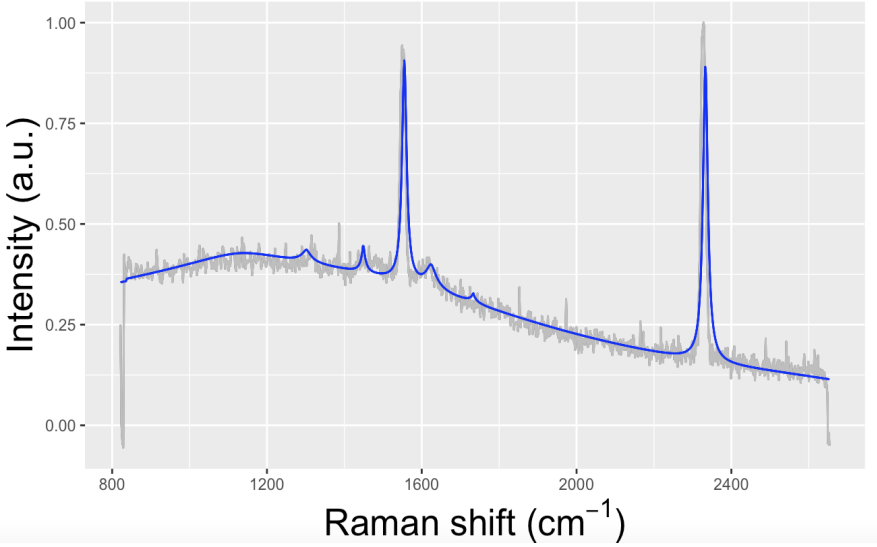

Figure 4.6: Credible Intervals(CI) for the peaks and baselines of average Lung Tissue spectra

4.1.2 Evaluation Metrics

In this subsection, we aim to evaluate the models. We utilized a marginal likelihood as a metric. The likelihood function is computed at the position corresponding to the parameter estimates, typically at the maximum likelihood location. In Bayesian statistics, there is no parameter vector representing the fit of the model, instead, the model can be evaluated based on the likelihood and the prior, the full posterior distribution of the parameters is derived. In our model, the marginal likelihood function is

Table 4.4: Model Evaluation

media	average value of log marginal likelihood
Cyclohexane	12.60
Sesame Oil	5.01
Lung Tissue	10.77

$$p(y_i(\tilde{\nu})|l,A,\psi,\gamma) = \frac{p(y_i(\tilde{\nu})|\Theta)\pi(\alpha)\pi(\sigma_\varepsilon^2)}{p(\alpha\sigma_\varepsilon^2|y_i(\tilde{\nu}),l,A,\psi,\gamma)} \tag{4.1}$$

In this metric, our aim is to achieve a higher average value of the log marginal likelihood, which indicates superior model fitting. As shown in Table 4.4, it is evident that the Cyclohexane data yields the best-fitting model, followed by Lung Tissue and Sesame Oil. Particularly, the log marginal likelihood values for Cyclohexane and Lung Tissue are more than twice that of Sesame Oil.

4.2 Exploring the Impact of Prior Data on Models

In the previous section, we utilized peak information from paper [17] as our prior knowledge to enhance the fitting models. However, we are curious about how our models perform when supplied with incorrect prior information. In this section, we conducted a thorough analysis by generating various plots using inaccurate peak information. This investigation allows us to assess the influence of prior data on the model's performance.

Moreover, we focus on Cyclohexane as our primary subject due to its stability and obvious peak characteristics in contrast to the other two media (sesame oil and lung tissue).

By conducting this experiment, we gained valuable insights into the sensitivity of our models to prior information and to ascertain the robustness of their predictions in the real-world where inaccuracies may be present.

4.2.1 Peak Locations

In this part, we investigated the impact of changing peak locations on our models. To achieve this, we designed an experiment where we generated inaccurate peak locations by segmenting the data between 800 to 1800 into intervals of 5, 10, 30, 50, 80, and 100, respectively. These artificially created peak locations allow us to explore how such

changes influence the performance of our models. By conducting this analysis, we gained valuable insights into the sensitivity and adaptability of models on variations in peak positions, thus enhancing our understanding of their overall robustness.

Figures 4.7 to 4.16 show that when a substantial number of incorrect peak locations are provided, visually represented by the dense arrangement of green lines, the overall fitting effect of the models is notably compromised. This effect is particularly pronounced when the data is generated every 5 between 800 and 1800.

Nevertheless, a noticeable enhancement in model performance is observed when the peak intervals are increased to 10. As depicted in Figures 4.9 and 4.10, the model effectively captures the peak characteristics, with the blue fitted line closely aligning with the grey line representing the original data.

Additionally, an interesting finding is that the model's performance seems to degrade when the intervals exceed 10. This phenomenon can be attributed to two factors. Firstly, the number of incorrect peak locations is twice as high for intervals of 5 compared to intervals of 10, impeding the model's ability to learn peak information accurately.

Secondly, as illustrated in Figure 4.11, fake peaks emerge on both sides of the spectral signal, generating broad, flat peaks. For those intervals of more than 30, despite the number of inaccurate peaks having decreased, the intervals are broader and the artificial peaks are far away from the real value, contributing to the worse effect.

Overall, while achieving commendable fitting results with 10 intervals, the findings underscore the importance of accurate peak location information in attaining satisfactory model fitting outcomes.

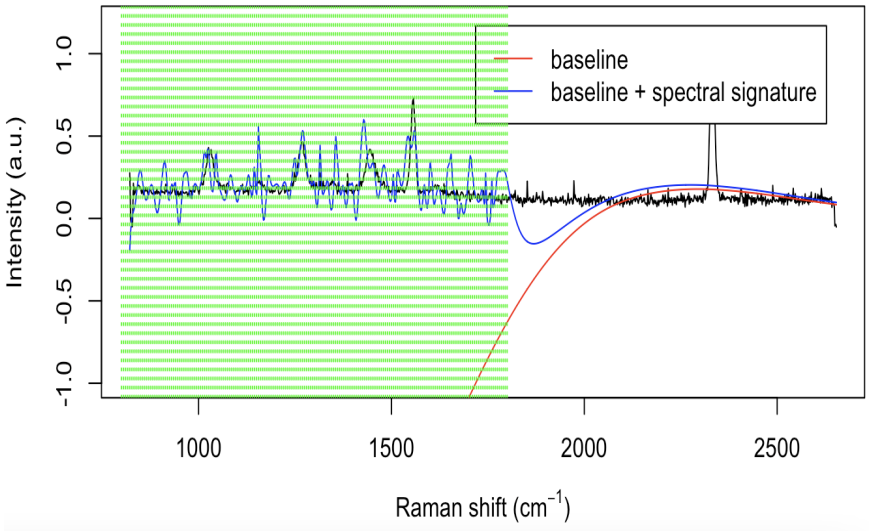

Figure 4.7: Impact of Incorrect Peak Locations (generating new peak locations every 5 between 800 and 1800)

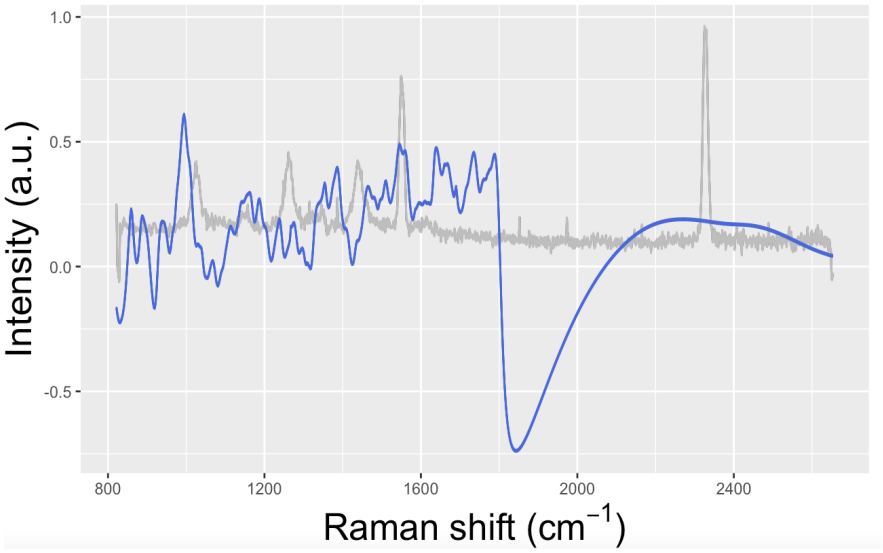

Figure 4.8: Credible Intervals(CI) for Incorrect Peak Locations (generating new peak locations every 5 between 800 and 1800)

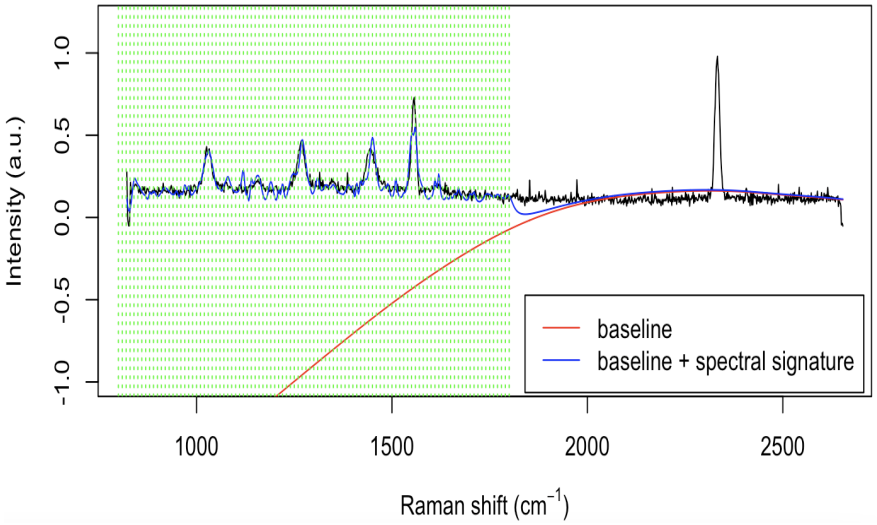

Figure 4.9: Impact of Incorrect Peak Locations (generating new peak locations every 10 between 800 and 1800)

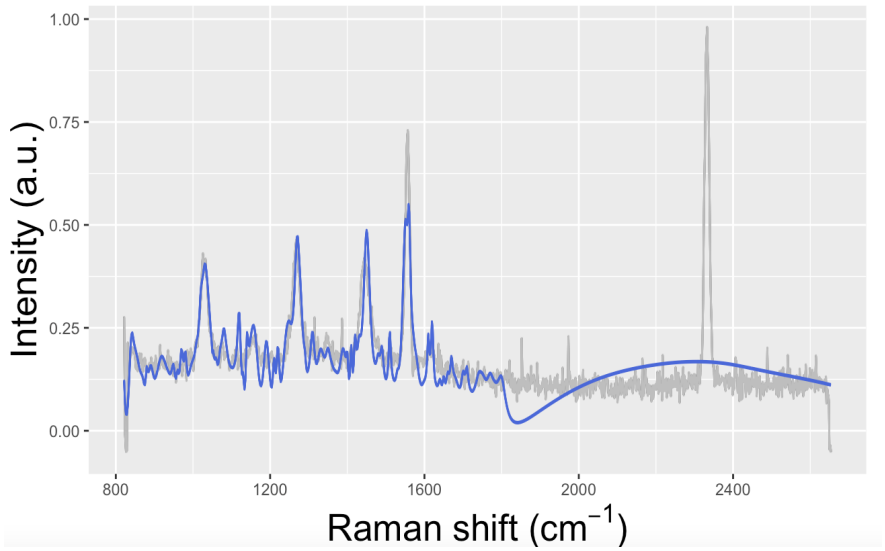

Figure 4.10: Credible Intervals(CI) for Incorrect Peak Locations (generating new peak locations every 10 between 800 and 1800)

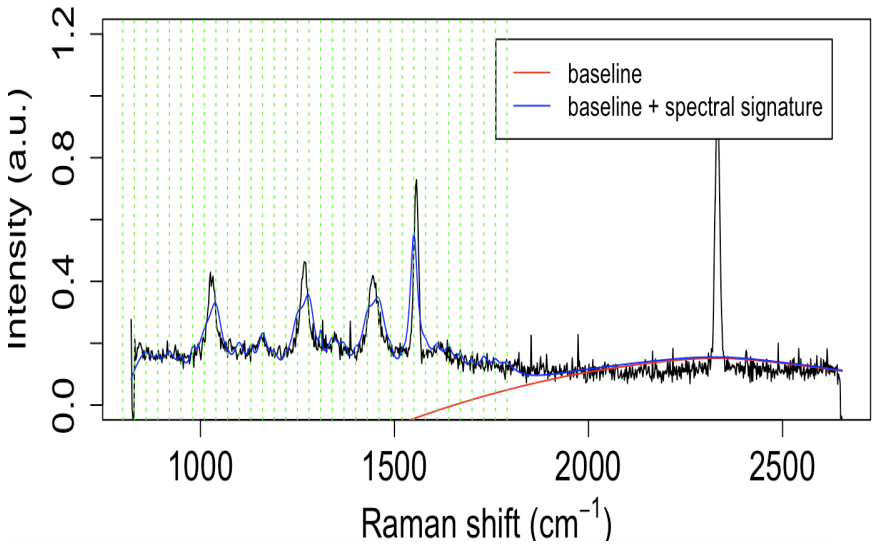

Figure 4.11: Impact of Incorrect Peak Locations(generating new peak locations every 30 between 800 and 1800)

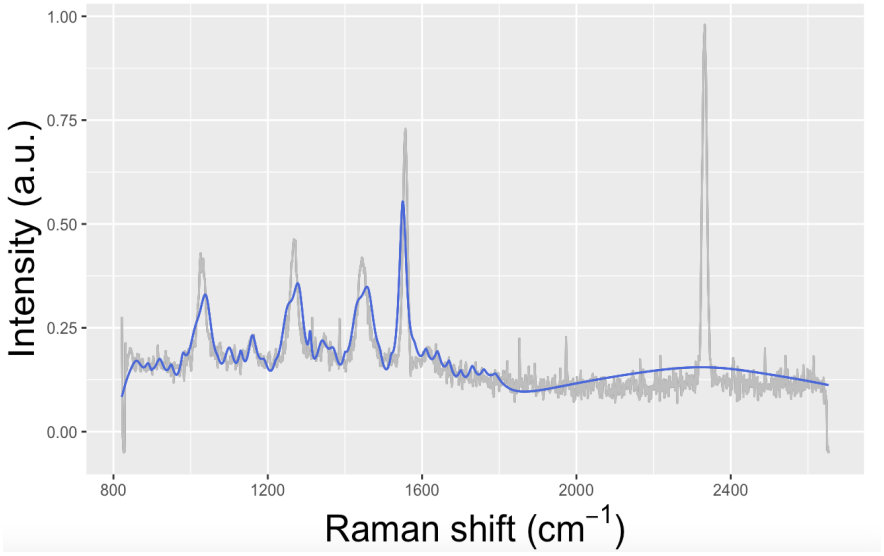

Figure 4.12: Credible Intervals(CI) for Incorrect Peak Locations (generating new peak locations every 30 between 800 and 1800)

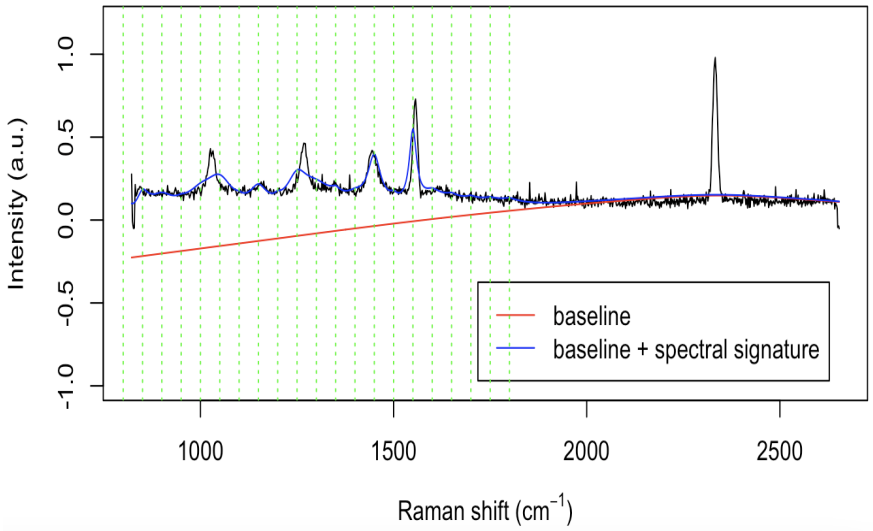

Figure 4.13: Impact of Incorrect Peak Locations (generating new peak locations every 50 between 800 and 1800)

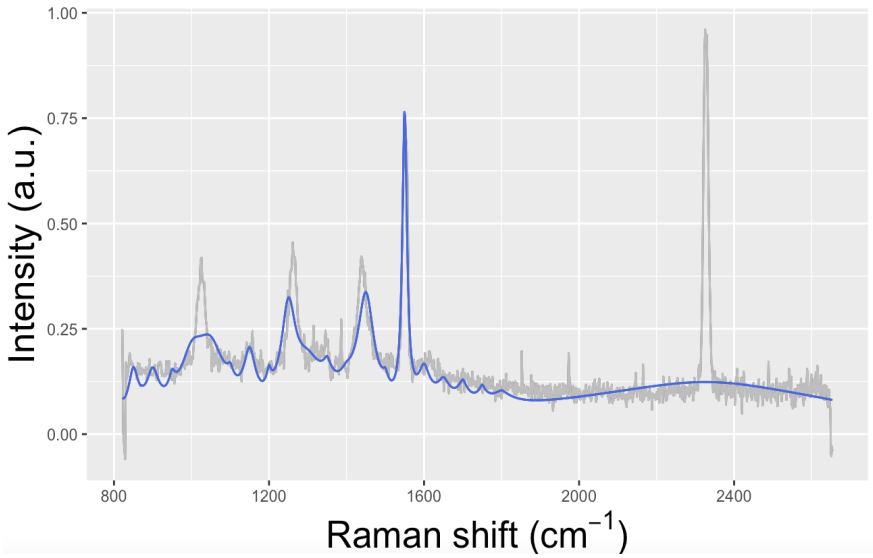

Figure 4.14: Credible Intervals(CI) for Incorrect Peak Locations (generating new peak locations every 50 between 800 and 1800)

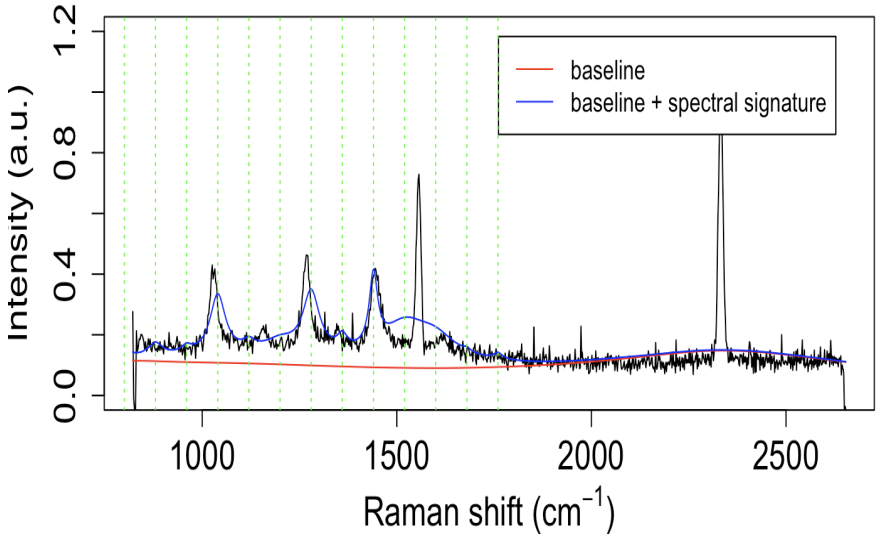

Figure 4.15: Impact of Incorrect Peak Locations(generating new peak locations every 80 between 800 and 1800)

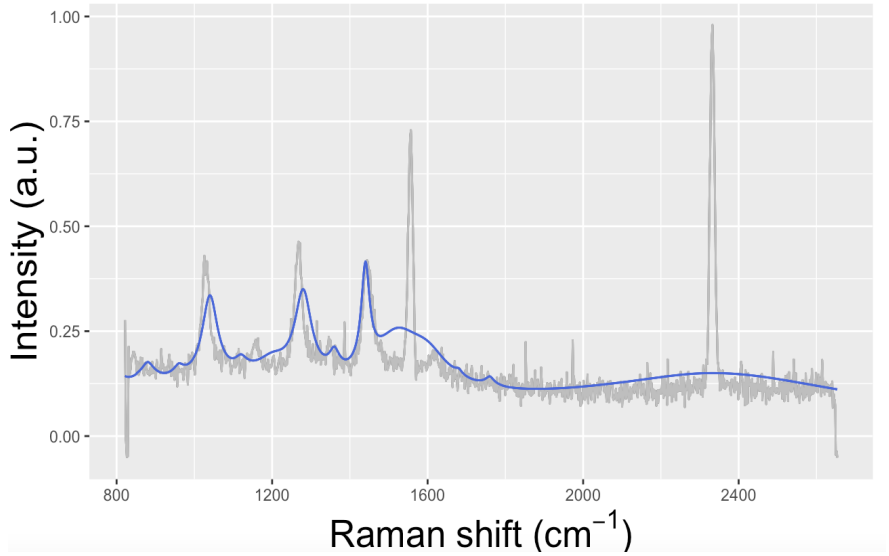

Figure 4.16: Credible Intervals(CI) for Incorrect Peak Locations (generating new peak locations every 80 between 800 and 1800)

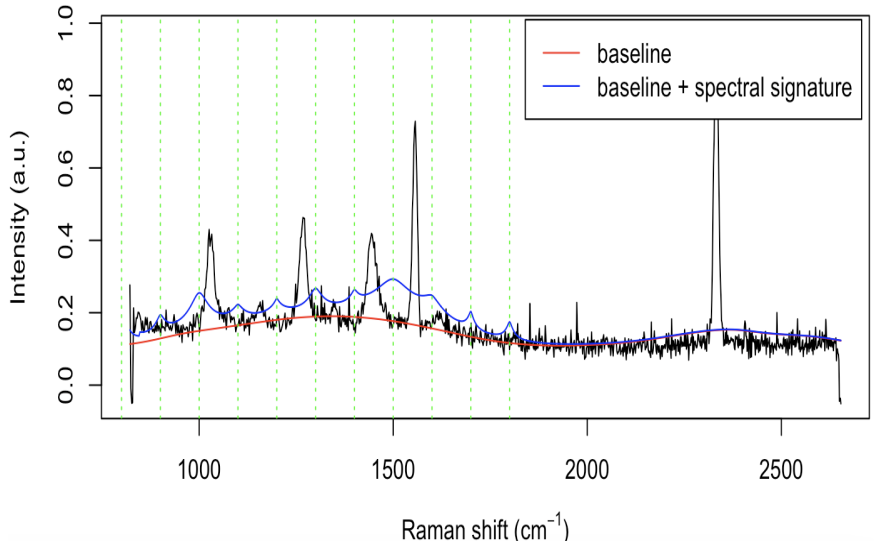

Figure 4.17: Impact of Incorrect Peak Locations(generating new peak locations every 80 between 800 and 1800)

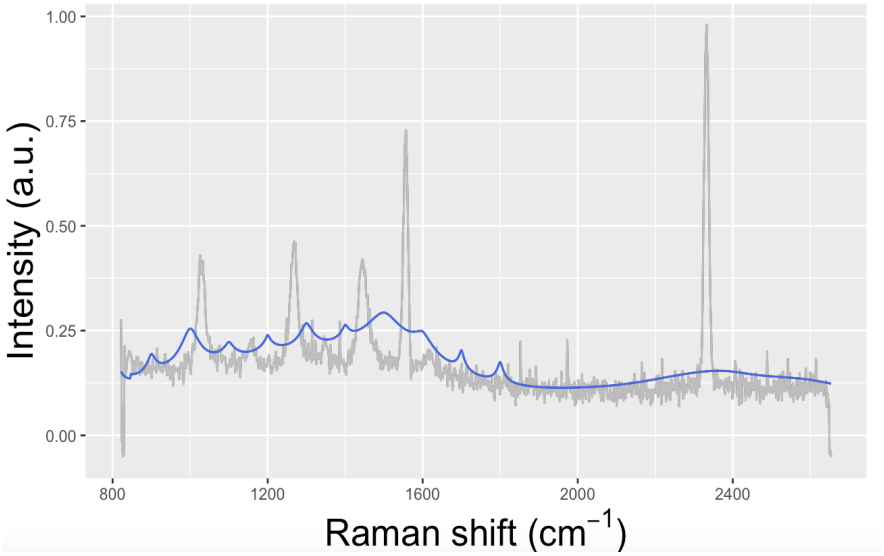

Figure 4.18: Credible Intervals(CI) for Incorrect Peak Locations (generating new peak locations every 80 between 800 and 1800)

Table 4.5: Model evaluation given inaccurate peak locations

number of intervals	average value of log marginal likelihood
5	-0.51
10	3.09
30	4.85
50	6.87
80	8.80
100	10.05

Chapter 5

Discussion

5.1 Result Interpretation

One main challenge of Raman spectroscopy is the interference of fluorescence background generated by biological molecules. This background can be several orders of magnitude stronger than the comparatively weak Raman scatter, making it difficult to observe and analyse the Raman spectrum. In our research, we tackled this issue by implementing an existing Bayesian model to analyse the Raman data gathered from another paper[17]. The primary objectives of our study are to investigate the fitting outcomes and assess their robustness in handling the fluorescence background. We anticipate that by doing this, we would improve the precision and reliability of Raman spectroscopy's applications in various industries, such as biotechnology, forensic science, and diagnostics [5].

Through a series of experiments in Chapter 4, we successfully tackled the research issues and attained our stated goals. Notably, after modifying the model parameters, such as smoothness and the number of knots, our models show increased fitting impact.

Moreover, it's evident that our models exhibit superior fitting capabilities for Cyclohexane and Lung Tissue compared to Sesame Oil. This discrepancy in performance can be attributed to the fact that Sesame Oil data is noisier and more contaminated, which consequently impedes the model's ability to accurately capture the peak amplitudes and scales.

A noteworthy aspect to highlight is that we only provided detailed peak locations, while amplitudes and scales are specified only by their median and standard derivation value. This also contributes to the challenge of fitting the signal signature. Despite the less-than-ideal conclusion for Sesame Oil, the outcomes obtained from Cyclohex-

ane and Lung Tissue demonstrate how effective our models are at capturing peaks' characteristics.

Additionally, we thoroughly investigated the influence of prior information on the model's performance. We discovered that when peculiar values, such as identical values, zeros, or extracted values at regular intervals, were introduced as prior information, the fitting effectiveness of the models diminished significantly.

Conversely, we made an intriguing observation: when we supplied incorrect but plausible prior information to the models, they exhibited notably better fitting results. This finding highlights the models' capacity to adapt to variations in prior data and underscores the importance of providing realistic and reasonable prior information for optimal performance.

Chapter 6

Conclusions

In conclusion, this paper was undertaken to explore an existing Bayesian hierarchical model[28] combined a Sequential Monte Carlo(SMC) algorithm on Raman spectroscopy in capturing features of signals from the complex fluorescence background, assess its sensibility in distinct datasets, and explore its availability when providing inaccurate prior knowledge. Through a set of experiments and analysis, we have illuminated several key insights that contribute to the theoretical and practical domains of this field.

Our investigation into the Bayesian hierarchical model, as implemented through the ssersBayes package, has shed light on the emphasis of combining prior information(peak locations, the median of peak scales and amplitudes, the standard derivation of scales and amplitudes) in enhancing the model fitting. The paper has also discussed the reason for the suboptimal fitting effect of Sesame Oil, which may be of assistance to further improvements. Moreover, we underscored the sensitivity of the model to the prior knowledge it relies upon by inputting the incorrect prior peak locations. While compromised fitting was observed, the model exhibited resilience and accuracy when provided with 10 intervals.

Our paper provides valuable practical contributions to an existing Bayesian hierarchical model [28]. While earlier research theoretically introduced the model for implementation in Raman data, it lacked a thorough exploration of the models' sensitivity and parameter variations. In contrast, our paper delves into flexibility using distinct data collected by us.

Additionally, the significance of this research extends beyond the laboratory. The potential of Raman spectroscopy to address clinical diagnostics and provide real-time insights into biological issues is evident. We research the impact of inaccurate prior knowledge on the model's performance, shedding light on real-world applications where

obtaining accurate real data is challenging. This investigation holds significance for practical applications, offering insights into the model's behavior in less-than-ideal circumstances.

One concern about the findings is that our samples are not enough. Among the three media, only Sesame Oil data emerged as notably noisier and more contaminated. Thus, it is hard to conclude that the model can better fit other abnormal data. Future investigations should be undertaken to explore a broader range of noisier data to comprehensively assess the model performance.

Moreover, we have yet to enhance the core code for the Bayesian model, which could potentially yield improved fitting outcomes for noisier datasets. Further efforts are required to advance the code quality. Analyzing the fitting impact calls for a refinement of the codebase. For instance, exploring whether the Markov chain Monte Carlo(MCMC) algorithm outperforms the Sequential Monte Carlo(SMC) algorithm could provide more valuable insights.

Bibliography

[1] James Berger, MJ Bayarri, and LR Pericchi. The effective sample size. *Econometric Reviews*, 33(1-4):197–217, 2014.

[2] PJ Cadusch, MM Hlaing, SA Wade, SL McArthur, and PR Stoddart. Improved methods for fluorescence background subtraction from raman spectra. *Journal of Raman Spectroscopy*, 44(11):1587–1595, 2013.

[3] Siddhartha Chib. Bayes regression with autoregressive errors: A gibbs sampling approach. *Journal of econometrics*, 58(3):275–294, 1993.

[4] Ronald J Clarke and Anna Oprysa. Fluorescence and light scattering. *Journal of chemical education*, 81(5):705, 2004.

[5] Ruchita S Das and YK Agrawal. Raman spectroscopy: Recent advancements, techniques and applications. *Vibrational spectroscopy*, 57(2):163–176, 2011.

[6] Johan J de Rooi and Paul HC Eilers. Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117:56–60, 2012.

[7] Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121, 1996.

[8] John R Ferraro. *Introductory raman spectroscopy*. Elsevier, 2003.

[9] Kasper Bayer Frøhling, Tommy Sonne Alstrøm, Michael Bache, Michael Stenbæk Schmidt, Mikkel Nørgaard Schmidt, Jan Larsen, Mogens Havsteen Jakobsen, and Anja Boisen. Surface-enhanced raman spectroscopic study of dna and 6-mercapto-1-hexanol interactions using large area mapping. *Vibrational Spectroscopy*, 86:331–336, 2016.

[10] Feng Gan, Guihua Ruan, and Jinyuan Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):59–65, 2006.

[11] Medhanie Tesfay Gebrekidan, Christian Knipfer, Florian Stelzle, Juergen Popp, Stefan Will, and Andreas Braeuer. A shifted-excitation raman difference spectroscopy (serds) evaluation strategy for the efficient isolation of raman spectra from extreme fluorescence interference. *Journal of Raman spectroscopy*, 47(2):198–209, 2016.

[12] Charles J Geyer. Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo*, 20116022:45, 2011.

[13] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

[14] Shixuan He, Wei Zhang, Lijuan Liu, Yu Huang, Jiming He, Wanyi Xie, Peng Wu, and Chunlei Du. Baseline correction for raman spectra using an improved asymmetric least squares method. *Analytical Methods*, 6(12):4402–4407, 2014.

[15] Yaogai Hu, Tao Jiang, Aiguo Shen, Wei Li, Xianpei Wang, and Jiming Hu. A background elimination method based on wavelet transform for raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 85(1):94–101, 2007.

[16] Roger M Jarvis, Alan Brooker, and Royston Goodacre. Surface-enhanced raman spectroscopy for bacterial discrimination utilizing a scanning electron microscope with a raman spectroscopy interface. *Analytical Chemistry*, 76(17):5198–5202, 2004.

[17] Nia C Jenkins, Katjana Ehrlich, András Kufcsák, Stephanos Yerolatsitis, Susan Fernandes, Irene Young, Katie Hamilton, Harry AC Wood, Tom Quinn, Vikki Young, et al. Computational fluorescence suppression in shifted excitation raman spectroscopy. *IEEE Transactions on Biomedical Engineering*, 2023.

[18] Michael Kasha. Characterization of electronic transitions in complex molecules. *Discussions of the Faraday society*, 9:14–19, 1950.

[19] Augustine Kong, Jun S Liu, and Wing Hung Wong. Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288, 1994.

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

[21] Takio Kurita. Principal component analysis (pca). *Computer Vision: A Reference Guide*, pages 1–4, 2019.

[22] Gang Li. Removing background of raman spectrum based on wavelet transform. In *2009 ETP International Conference on Future Computer and Communication*, pages 198–200. IEEE, 2009.

[23] Chad A Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied spectroscopy*, 57(11):1363–1367, 2003.

[24] Jun S Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

[25] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001.

[26] 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.

[27] Richard L McCreery. *Raman spectroscopy for chemical analysis*. John Wiley & Sons, 2005.

[28] Matthew Moores, Kirsten Gracie, Jake Carson, Karen Faulds, Duncan Graham, and Mark Girolami. Bayesian modelling and quantification of raman spectroscopy. *arXiv preprint arXiv:1604.07299*, 2016.

[29] Zanyar Movasaghi, Shazza Rehman, and Ihtesham U Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007.

[30] Ricardo Piña, Alma I Santos-Díaz, Erika Orta-Salazar, Azucena Ruth Aguilar-Vazquez, Carola A Mantellero, Isabel Acosta-Galeana, Argel Estrada-Mondragon, Mara Prior-Gonzalez, Jadir Isai Martinez-Cruz, and Abraham Rosas-Arellano. Ten approaches that improve immunostaining: A review of the latest advances for the optimization of immunofluorescence. *International journal of molecular sciences*, 23(3):1426, 2022.

[31] Andreas F Ruckstuhl, Matthew P Jacobson, Robert W Field, and James A Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193, 2001.

[32] Chris Sherlock, Paul Fearnhead, and Gareth O Roberts. The random walk metropolis: linking theory and practice through a case study. 2010.

[33] GK Wertheim and SB Dicenzo. Least-squares analysis of photoemission data. *Journal of electron spectroscopy and related phenomena*, 37(1):57–67, 1985.

[34] Huibin Yu, Yonghui Song, Xiang Tu, Erdeng Du, Ruixia Liu, and Jianfeng Peng. Assessing removal efficiency of dissolved organic matter in wastewater treatment using fluorescence excitation emission matrices with parallel factor analysis and second derivative synchronous fluorescence. *Bioresource Technology*, 144:595–601, 2013.

[35] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[36] Zhi-Min Zhang, Shan Chen, Yi-Zeng Liang, Zhao-Xia Liu, Qi-Ming Zhang, Li-Xia Ding, Fei Ye, and Hua Zhou. An intelligent background-correction algorithm for highly fluorescent samples in raman spectroscopy. *Journal of Raman spectroscopy*, 41(6):659–669, 2010.

[37] Ruomei Zhao, Lulu An, Di Song, Minzan Li, Lang Qiao, Ning Liu, and Hong Sun. Detection of chlorophyll fluorescence parameters of potato leaves based on continuous wavelet transform and spectral analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 259:119768, 2021.