HEAD OF DEPARTMENT

BSc Thesis Task Description

Homoki Márton János
candidate for BSc degree in Electrical Engineering

Mixture models in Raman spectroscopy

Raman spectroscopy is based on the phenomenon that the wavelength of the light changes when scattered from objects. The change in the wavelength depends on the actual material. Measuring the intensity at different wavelengths gives us the spectrum of light. The spectrum can be modeled with the help of a mixture of distributions, to be more concrete, with a mixture of Gaussian and Lorentz distributions. The goal of this topic is to develop algorithms, which can identify and fit models to actual spectra in practical applications.

Tasks to be performed by the student will include:

- Model Development: Create robust and flexible mathematical models that can automate the currently manual process of fitting peaks onto the spectra. This model will consider factors such as the initial parameters of the Lorentz and Gaussian components and the data points obtained from the spectra.
- Optimization: The accuracy, speed and complexity of the models will be considered.
- Model Analysis: Apply the developed model to test data provided by experts. Evaluate the fitness and practicality of the modeling solution by analyzing the results obtained. Check the accuracy of the fit with experts in the field. Present a comparison of the automated and the manual fitting.
- Visualization: Visualize the model consisting of the Lorentz and Gaussian components on top of the spectra.

Supervisor at the department: László György Grad-Gyenge, assistant research fellow
External supervisor:

Budapest, 23 September 2023

Dr. Hassan Charaf
professor
head of department

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Automation and Applied Informatics
H-1117 Budapest, Magyar Tudósok krt 2. Q.B.207
Tel: +3614632870, Fax: +3614632871
www.aut.bme.hu

Budapest University of Technology and Economics

Faculty of Electrical Engineering and Informatics

Department of Automation and Applied Informatics

Homoki Márton János

MIXTURE MODELS IN RAMAN SPECTROSCOPY

BSc thesis

SUPERVISOR

László György Grad-Gyenge

BUDAPEST, 2023

Table of Contents

Abstract **1**

Összefoglaló **2**

1 Introduction **3**

1.1 The Gauss-Lorentz Model 4

1.2 The target algorithm 4

2 Related work **6**

2.1 Work related to the model and algorithm 6

2.1.1 GMM 6

2.1.2 Partial data 9

2.2 Work related to fitting in the domain of spectroscopy 12

2.2.1 Introduction 12

2.3 Polynomial fitting techniques 12

2.3.1 BubbleFill 13

2.4 Modified Multi-Polynomial Fitting (ModPoly) 13

2.5 Peak removal techniques 13

2.6 Statistical Methods in Fluorescence Removal 14

2.7 Advancements in Automated Algorithms: 14

3 Data **14**

3.1 Raman spectroscopy 14

3.2 Raman spectra 15

4 Method **16**

4.1 EM algorithm 16

4.1.1 Introduction 16

4.1.2 E-step, M-step and stop criteria 17

4.1.3 EM in the context of mixture models 17

4.1.4 Extending EM with GMM to include Lorentz distributions 18

4.2 Partial results 19

4.2.1 Synthetic data 19

4.2.2 Code to generate synthetic data 19

4.2.3 EM applied to synthetic data 19

4.2.4 EM onto partial data 20

4.3 Gradient descent 22
4.3.1 Introduction 22
4.3.2 Adaptive moment estimation 23
4.4 Implementation 23
4.4.1 Extended Probability density functions 23
4.4.2 Quadratic loss function 24
4.4.3 Partial derivatives 24
4.4.4 Penalty term 27
4.4.5 The algorithm 29
4.4.6 The program 30
4.5 Priory information 30
4.5.1 Given parameters 30
4.5.2 Intervals without peaks 31
4.5.3 Gamma values 32
5 Results 33
5.1 The mixture fit 33
5.1.1 The fit 33
5.1.2 Validation 36
5.1.3 The cost over iteration 36
6 Conclusion 39
6.1 The mixture fitting 39
6.1.1 The baseline 39
6.1.2 The Peaks 39
6.1.3 Conclusion 39
6.2 Future work 40
6.2.1 Adjustment of model 40
6.2.2 Broad Applicability: Extending Beyond Raman Spectra 41
6.2.3 Extending the EM algorithm 41
6.2.4 Automating Identification 42
6.2.5 Improvement of the initialization of parameters 42
6.2.6 Final thoughts 43
Acknowledgment 44

References..... 45

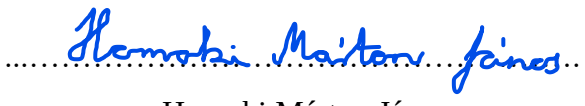
STUDENT DECLARATION

I, **Márton János Homoki**, the undersigned, hereby declare that the present BSc thesis work has been prepared by myself and without any unauthorized help or assistance. Only the specified sources (references, tools, etc.) were used. All parts taken from other sources word by word, or after rephrasing but with identical meaning, were unambiguously identified with explicit reference to the sources utilized.

I authorize the Faculty of Electrical Engineering and Informatics of the Budapest University of Technology and Economics to publish the principal data of the thesis work (author's name, title, abstracts in English and in a second language, year of preparation, supervisor's name, etc.) in a searchable, public, electronic and online database and to publish the full text of the thesis work on the internal network of the university (this may include access by authenticated outside users). I declare that the submitted hardcopy of the thesis work and its electronic version are identical.

Full text of thesis works classified upon the decision of the Dean will be published after a period of three years.

Budapest, 8 December 2023



Homoki Márton János

Abstract

This study explores the challenges of fitting Gaussian Mixture Models (GMM) and Lorentzian distributions onto random skin tissue Raman spectra samples. Initial attempts to employ the Expectation-Maximization (EM) algorithm revealed suboptimal performance, prompting the exploration of alternative methodologies. Subsequently, a novel fitting algorithm based on gradient descent was developed, specifically tailored for the unique characteristics of Raman spectra in skin tissue. Comparative analyses between the EM-based approach and the proposed gradient descent algorithm demonstrated the latter's superior efficacy in accurately capturing the intricate features of the Raman spectra. The findings underscore the significance of algorithmic adaptation to the distinctive nature of Raman spectra in biomedical applications, offering a promising avenue for improved analysis and interpretation in skin tissue studies.

Összefoglaló

Ez a tanulmány a Gaussian Mixture Model (GMM) és Lorentz féle eloszlások illesztési kihívásait vizsgálja a véletlenszerű bőr Raman-spektrum mintákon. Az elsődleges kísérletek az ExpectationMaximization (EM) algoritmussal történő illesztésre alacsony hatékonyságot mutattak, ezért alternatív módszerek felé fordultunk. Ezt követően egy új illesztési algoritmust fejlesztettünk ki, amely a gradiens descent alapján működik, kifejezetten a bőrspektrum Raman-spektrum sajátosságaihoz igazítva. Az EM-alapú megközelítés és a javasolt gradiens descent algoritmus közötti összehasonlító elemzések kimutatták az utóbbi kiemelkedő hatékonyságát a Raman-spektrumok bonyolult jellemzőinek pontos rögzítésében. Az eredmények hangsúlyozzák az algoritmikus alkalmazkodás fontosságát a biomedikai alkalmazásokban a Raman-spektrumok egyedi jellegéhez, ígéretes lehetőséget kínálva a bőrszöveti tanulmányokban történő analízis és értelmezés javítására.

1 Introduction

In this study, we address the problem of optimizing a model for the representation of Raman spectra in skin tissue. The model involves a mixture distribution consisting of Lorentz and Gauss distributions. We confront a critical decision: the choice of a feature extraction algorithm. We transition from the Expectation Maximization (EM) algorithm, which proved ineffective in this context, to the gradient descent algorithm. This transition is motivated by the need to enhance the model's capability to accurately capture the unique characteristics of Raman spectra geometrically. The objective of this introduction is to elucidate the significance of algorithm selection in data modeling and highlight the specific context of our investigation.

The field of biomedical Raman spectroscopy has witnessed significant advancements in recent years, driven by the noninvasive nature of this optical technique and its capacity to detect subtle molecular or biochemical signatures within tissues.[1][2] This technology holds great promise for clinical applications, particularly with the development of hardware systems that have substantially reduced spectral acquisition times, making real-time Raman spectroscopy feasible in clinical settings. However, a major challenge in biomedical Raman spectroscopy is the effective removal of intrinsic autofluorescence background signals, which often overshadow the relatively weaker Raman scattering signals. The separation of these signals is crucial for obtaining accurate and meaningful molecular information.

Various methods have been proposed for fluorescence background removal, encompassing both instrumental and computational approaches. While instrumental methods involve modifications to the spectroscopic systems, computational methods, facilitated by software, include shifted excitation, time gating, Fourier transformation, and polynomial fitting. Among these, polynomial fitting has gained popularity due to its simplicity and convenience, especially for in vivo biomedical applications. However, the efficacy of traditional polynomial fitting methods is compromised in real-time and low signal to noise ratio (SNR) environments.

1.1 The Gauss-Lorentz Model

The incorporation of the Gauss-Lorentz model in our study is specifically tailored to address the dual challenges posed by fluorescence background and distinct Raman peaks in skin tissue spectra. The Gaussian distribution within the model is strategically employed to capture and model the broader, symmetrical features associated with fluorescence background signals. This component is adept at effectively characterizing the fluorescence contribution, which tends to exhibit a more uniform and widespread distribution.

Conversely, the Lorentzian distribution is chosen to accurately represent the narrower and often asymmetric peaks in Raman spectra. Lorentzian profiles are well suited for modeling the distinct vibrational modes and peaks associated with Raman scattering. By combining these two components, the Gauss-Lorentz model enables our algorithm to effectively disentangle the overlapping fluorescence background from the specific Raman peaks, providing a more nuanced and accurate representation of the molecular information present in skin tissue samples. This strategic selection ensures that our model is equipped to handle the unique challenges posed by both fluorescence and Raman features, contributing to the robustness and precision of our data modeling approach.

It is notable to mention that by default the peaks are usually themselves comprised of both Lorentz and Gaussian distribution percentage wise, but in this domain of spectroscopy as pointed out by experts in the field, the peaks are plain Lorentz distributions.[2]

1.2 The target algorithm

The goal of our study is to develop an algorithm designed to fit the Gauss-Lorentz model onto partial data using priori information of the Lorentz and the Gaussian distributions. In other word the extraction of features from the spectra such as location, width and height of peaks and the background fluorescence.

This targeted algorithm is specifically crafted to optimize the representation of Raman spectra in skin tissue. The Gauss-Lorentz model, selected for its capacity to address

both fluorescence background and distinct Raman peaks, employs Gaussian distributions to capture the broader, symmetrical features associated with fluorescence signals and Lorentzian distributions for the narrower, often asymmetric peaks in Raman spectra.

The algorithm's primary function is to effectively disentangle the overlapping fluorescence background from the specific Raman peaks, providing a nuanced and accurate representation of molecular information present in skin tissue samples. Leveraging priori information of the Lorentz distributions and the Gaussians, the algorithm aims to iteratively refine its parameters and fit the model to partial data, enhancing its adaptability and precision.

By combining the strengths of both Gauss and Lorentz components within the model, the algorithm ensures a robust and versatile approach to data modeling. The ultimate objective is to create a tool that can navigate the challenges posed by fluorescence and Raman features, contributing to the overall advancement of biomedical Raman spectroscopy for noninvasive molecular characterization within biological tissues.

2 Related work

2.1 Work related to the model and algorithm

2.1.1 GMM

Gaussian Mixture Model (GMM) is a probabilistic model that represents a mixture of multiple Gaussian (normal) distributions. This is related to our work in such a way that the Gaussian-Lorentz mixture model is an extension of this by incorporating an extended mixture of Lorentz distributions. GMM has seen uses in the Expectation Maximization algorithm (EM) algorithm, being one of the most common models utilized in.

GMM with the Expectation Maximization algorithm (EM) has seen a wide variety of uses in [3][4][5][6][7][8][9][10][11]. The traditional algorithm usage with GMM model is well summarized in [4]. Where an explanation for the whole process is provided, specifically with GMM and is later touched on in section 5. In [5] the traditional GMM with EM model is applied to concatenated vectors of high- and low-resolution patches sampled from a large database of pairs of high-resolution and the corresponding low-resolution images in order to achieve Single image super-resolution (SR).

In [3] Wi-Fi fingerprinting based indoor positioning systems where an extension of GMM is used in order to deal with dropped data such as those happening from unanticipated operations of equipment or the temporary switching off state of APs for the energy-saving purpose. In [3] the model is extended by incorporating missing a binary assignment to a value being dropped, thereby extending the traditional GMM into a Censored and Dropped Gaussian Mixture Model (CD-GMM). The underlying algorithm used in this fitting is the EM algorithm. Another spectacular part of the paper is that when collected RSS (Received Signal Strength) values which didn't fall that data will still follow the traditional GMM model.

On the other hand, there have been cases [7][8] where the EM algorithm itself was modified. Such as the GMM applied to text-independent speaker recognition in [7]. The proposed method, referred to as Maximum Normalized Likelihood Estimation

(MNLE), aims to maximize the frame-level normalized likelihoods of the training data. Thereby improving speaker identification rates and verification equal error rates compared to Maximum Likelihood Estimation (MLE) and MCE/GPD methods. This MNLE algorithm modified the E step of the EM algorithm by altering the maximum likelihood Estimate. In the standard EM algorithm, the E step involves computing the expected value of the latent variables, given the observed data and the current parameter estimates. However, in MNLE, the likelihood normalization is introduced, altering the formulation of the E step. This normalization is performed to ensure that the likelihoods are normalized probabilities, and this modification is crucial for the discriminative nature of MNLE. It adjusts the likelihoods based on the competing models and the target model, contributing to the enhanced separation between speakers in the training process. In one other case [8] EM for a sensor system the EM algorithm is extended in such a way that data, read from sensors from unattended sensor systems are prone to fault thereby avoiding this issue, a Fault-Tolerant Expectation maximization (FEM) yet again similarly to the previous case introduces a modification to the expectation maximization by reading data one by one from the sensors in a separate subspace, thereby the same time allowing for the lower power consumption of sensors as the sensors can now switch off more frequently due to the lower prone to an error happening in the system.

Another concurrent theme in this area is the idea to combine 2 or more, more than one traditional algorithm, in some systems a necessary step, In [9] we saw the combination of the traditional multidimensional GMM with EM combined with a support vector machine (SVM) this step was a necessary combination in order to detect the probability of cancer the classification of the extracted features into three classes of tissues as normal, benign or malignant, was crucial. In a different case in [6] EM was used along with the genetic algorithm a global stochastic search in order to solve an issue known as the genetic drift problem.

Yet a different approach in this domain is the underlying hardware modification presented [10], where the hardware structure is optimized for running such an algorithm as EM. This can be achieved by Field-Programmable Gate Arrays (FPGAs) which are commonly used to optimize hardware structures for running specific algorithms.

FPGAs are programmable integrated circuits that can be configured to implement custom digital circuits and functions. They provide flexibility and parallelism, making them well-suited for accelerating specific algorithms and achieving performance improvements compared to general-purpose processors. FPGAs excel at parallel processing, allowing multiple operations to be performed simultaneously such as the computation for one batch of data. Algorithms that can be parallelized benefit significantly from FPGAs, as they can exploit the parallel architecture to achieve higher throughput. Additionally, FPGAs can be more power-efficient for certain workloads compared to general-purpose processors. This contrasts with traditional processors, which are designed for general-purpose computation and may not be as efficient for certain types of algorithms. This was the case in [10] where the paper presents an optimized implementation of EM algorithm on Stratix V and Arria 10 FPGAs using Intel FPGA Software Development Kit (SDK) for Open Computing Language (OpenCL).

In summary we can see that in most cases either the model, the algorithm is modified, simply the traditional method is applied, multiple traditional methods are combined, or the hardware is optimized and used in order to show feedback, improvement, novelty and innovation in the domain of GMM.

Some other unmentioned general examples: Speech and Speaker Recognition, Image and Video Processing: Segmentation, object recognition, and video analysis based on pixel intensity distributions. Natural Language Processing (NLP): Text classification, language identification. Biometrics: Face recognition, fingerprint identification, Anomaly Detection: Identifying anomalies in various fields, such as network intrusion detection or fraud detection, Medical Image Analysis: Segmentation and feature extraction in medical imaging, Financial Modeling: Modeling financial data for risk assessment, fraud detection, Data Compression: Used in speech and audio coding for efficient data compression, Robotics: Applied to robot perception tasks, including object recognition and localization, Genetics and Bioinformatics: Modeling and analyzing biological data, such as DNA sequences.

2.1.2 Partial data

By partial data in the context of statistical modeling, we mean that only a subset or portion of the complete dataset is available, and crucial information, especially related to the distribution of the data, is missing. This scenario poses challenges because the missing part may contain valuable insights for accurately characterizing the underlying distribution.

In the pursuit of advancing biomedical Raman spectroscopy for noninvasive molecular characterization within biological tissues, the optimization of spectral modeling algorithms remains a critical focal point. This study addresses the intricate challenge of fitting partial data, a scenario frequently encountered in Raman spectroscopy applications involving skin tissue. The employed model integrates a Gauss-Lorentz distribution, strategically chosen for its efficacy in simultaneously capturing fluorescence background and distinct Raman peaks. Fitting partial data introduces inherent complexities, including the risk of information loss, potential overfitting or underfitting, heightened parameter sensitivity, and challenges in selecting representative subsets for modeling.

The fitting algorithm's adaptability to partial data necessitates a meticulous consideration of initialization procedures and robust parameter tuning. The challenge is magnified by the iterative nature of the algorithm, where the convergence to an optimal solution becomes contingent upon the quality of the initial estimates. Moreover, the representativeness of partial data poses a significant concern, as biased subsets may compromise the generalizability of the model to the entire spectrum, thereby impacting the validity of the fitted results.

Furthermore, the computational intensity of fitting partial data cannot be understated, as the algorithm must contend with potential data gaps, necessitating extrapolation or interpolation methods that introduce computational overhead. In the presence of noisy data, a prevalent issue in Raman spectroscopy, the algorithm must exhibit resilience in distinguishing genuine signal from noise, particularly in scenarios characterized by low signal-to-noise ratios.

In addressing these complexities, the algorithm's design should mandate a careful balance between computational efficiency and robustness. Rigorous validation, encompassing diverse datasets and comprehensive statistical analyses, is imperative to ensure the algorithm's reliability in capturing the intricate features of Raman spectra in skin tissue. This scientific inquiry contributes to the broader understanding of algorithmic challenges in the field, paving the way for enhanced methodologies in the precision-driven realm of biomedical Raman spectroscopy.

Within the domain of addressing challenges posed by incomplete datasets, researchers have explored methodologies for estimating the joint distribution, particularly emphasizing the Gaussian copula and marginal distributions in the context of partial data.

While existing studies have focused on imputation methods for handling missing values, this specific approach stands out for its innovative use of an EM algorithm based on Monte Carlo integration [11]. Notably, this algorithm adopts an iterative process, concurrently updating both copula and marginal distributions. This design acknowledges the inherent complexities associated with incomplete data, where the absence of values may not follow a random pattern.

A distinctive feature of this work is its explicit consideration of the missing data mechanism. The incorporation of semiparametric mixture models provides a robust solution, especially when there is a lack of prior knowledge about the marginals. This aspect contributes to the broader understanding of how to navigate challenges posed by incomplete datasets in statistical modeling.

The practical application of the proposed methodology in [11] assessing key performance indicators at BMW's Battery Cell Competence Center adds real-world relevance. By demonstrating the utility of the method in a tangible scenario with missing data, the approach proves valuable for addressing challenges in various domains where accurate joint distribution estimation is critical.

In essence, this related work significantly contributes to the ongoing discourse on incomplete datasets. By introducing a methodology that effectively handles the complexities of estimating joint distributions under missing data scenarios, the study

offers valuable insights and practical implications for researchers and practitioners navigating similar challenges.

The problem with such solutions is the need to modify the underlying model. This can be a problem since Dealing with incomplete datasets poses a significant challenge in statistical modeling, and various strategies have been proposed to address this issue. One common approach involves modifying the underlying model to accommodate missing data. However, this solution is not without its own set of potential drawbacks and considerations.

One primary concern associated with modifying the model is the risk of model misspecification. Adjustments made to the original model must accurately reflect the underlying data-generating process; otherwise, biased parameter estimate, and inaccurate inferences may result. Striking the right balance between model flexibility and fidelity to the data is crucial.

Another consideration is the increased complexity that often accompanies model modifications. Introducing adjustments may lead to a more intricate model, making it challenging to interpret and implement. Moreover, increased complexity can pose computational challenges, especially when dealing with large datasets, potentially impacting the efficiency and scalability of the solution.

The need for assumptions about the missing data mechanism when modifying the model introduces a risk of assumption violation. If these assumptions do not hold in practice, the model's validity may be compromised, affecting its ability to accurately represent the true underlying patterns in the data.

Generalization challenges also arise when modifying models to handle incomplete data. Adjustments made for one dataset may not necessarily generalize well to new or unseen data, limiting the applicability of the model across diverse datasets.

Furthermore, the computational intensity associated with some modified models can be a practical concern. The increased demand for computational resources and time may pose limitations on the feasibility of implementing such solutions, particularly in scenarios with large datasets.

2.2 Work related to fitting in the domain of spectroscopy

2.2.1 Introduction

In this section we would like to talk about those models and algorithms which are a far breath away from GMM and EM, which were discussed in chapter 2.1. Those tools which are now relevant in Raman spectroscopy namely the type used for fitting auto-fluorescence background signals and the peaks.

The field of biomedical Raman spectroscopy has seen considerable advancements in addressing the challenges posed by auto-fluorescence background signals. Various methods [12][13] have been proposed and implemented to improve the accuracy of fluorescence removal, particularly in the context of real-time applications and scenarios characterized by low signal-to-noise ratios (SNR). In this section, we review relevant literature and highlight key approaches that have paved the way for the development of the Improved Modified Multi-Polynomial Fitting (I-ModPoly) algorithm.

2.3 Polynomial fitting techniques

Traditional polynomial fitting methods have been widely employed for fluorescence background removal. These methods, although straightforward, are sensitive to the choice of polynomial order and the spectral range used for fitting. Prior studies [12][13] have investigated the efficacy of different polynomial orders and their impact on the accurate separation of Raman and fluorescence signals. While these methods provided a foundation, challenges persisted in achieving optimal results in real-time applications. The field of biomedical Raman spectroscopy has seen considerable advancements in addressing the challenges posed by autofluorescence background signals. Various methods have been proposed and implemented to improve the accuracy of fluorescence removal, particularly in the context of real-time applications and scenarios characterized by low signal-to-noise ratios (SNR). In this section, we review relevant literature and highlight key approaches that have paved the way for the development of the Improved Modified Multi-Polynomial Fitting (I-ModPoly) algorithm.

2.3.1 BubbleFill

This study [12] presents a new and open-sourced spectroscopic data pre-processing package that includes a novel morphological baseline removal technique called BubbleFill. The primary goal of the package is to improve adaptability to complex baseline shapes, a challenge not adequately addressed by current gold standard techniques. Additionally, the package features a versatile tool for simulating spectroscopic data with varying characteristics, allowing researchers to explore different scenarios.

The results indicate that the BubbleFill technique outperforms established algorithms such as iModPoly and MorphBR when applied to simulated data, showcasing its superior baseline removal capabilities. Further validation on four independent in-human datasets demonstrates the package's efficacy in ensuring compatibility across different spectroscopic systems.

In conclusion, this open-sourced package provides a valuable resource for researchers and clinicians working with Raman spectroscopy. Its innovative BubbleFill technique, along with the capability for simulating diverse data scenarios, positions the package as a promising tool for developing new clinical applications in the field.

2.4 Modified Multi-Polynomial Fitting (ModPoly)

The introduction of ModPoly marked a significant improvement in fluorescence background removal. ModPoly addressed some limitations of conventional polynomial fitting by incorporating an iterative procedure. This approach demonstrated enhanced rejection of fluorescence signals and improved accuracy in capturing Raman spectra. However, challenges persisted in adapting ModPoly to real-time scenarios and environments with low SNR, prompting the need for further refinements.[13]

2.5 Peak removal techniques

Recognizing the importance of addressing peak-related issues in fluorescence removal, certain studies focused on peak-removal techniques. These techniques aimed to eliminate artificial peaks introduced during the fitting process, ensuring a more precise separation of Raman and fluorescence contributions. While these approaches showed

promise, a comprehensive solution that integrates peak removal with an efficient algorithm suitable for real-time applications remained elusive.[13]

2.6 Statistical Methods in Fluorescence Removal

Some studies explored the integration of statistical methods to enhance fluorescence rejection in Raman spectra. These methods considered the statistical characteristics of signal noise, contributing to a more robust algorithm. However, their applicability in real-time settings and their effectiveness in scenarios with intense Raman peaks required further investigation.[13]

2.7 Advancements in Automated Algorithms:

The emergence of automated algorithms for fluorescence background removal marks a pivotal shift in the landscape of biomedical Raman spectroscopy. Researchers have increasingly recognized the need for solutions that not only deliver high accuracy in separating Raman and fluorescence signals but also offer efficiency and user-friendliness, particularly in real-time applications. This transition reflects a broader acknowledgment of the practical challenges faced in experimental settings and the growing demand for adaptable and streamlined methodologies.[13] In this paper we will not consider polynomial algorithms any further as we would like to avoid using a polynomial based model, instead opting for the model mentioned in section 1 consisting of Gauss and Lorentz distributions.

3 Data

3.1 Raman spectroscopy

Raman spectroscopy stands as a powerful scientific technique employed for investigating molecular vibrations and rotations within a sample. Its fundamental principle, the Raman shift, captures the energy difference between incident laser light and the resulting scattered light. This technique offers a unique window into the chemical composition, molecular structure, and physical properties of diverse materials.

Raman spectroscopy is a widely used scientific technique that explores molecular vibrations and rotations by measuring the inelastic scattering of laser light in a sample. The key principle is the Raman shift, the energy difference between the incident laser light and the scattered light. It provides insights into the chemical composition, molecular structure, and physical properties of the sample.

Raman spectroscopy is versatile and finds applications in chemistry, materials science, pharmaceuticals, forensics, and geology. It is non-destructive, suitable for various sample types, and complements other techniques such as infrared spectroscopy. Recent advancements include Raman imaging and resonance Raman spectroscopy, enhancing its capabilities.

However, challenges include sensitivity to fluorescence interference and the relatively weak signal, which may require longer measurement times.[1]

3.2 Raman spectra

Samples differ in composition, size quantity, duration of the measurement and other factors. In this thesis the Raman spectrum, of skin tissue is considered. Due to fluorescence, a proportionally increasing slope below the spectrum appears, this is a point of interest. As well as the peaks, whose location, amplitude and width house important information about the molecules.

This thesis centers its exploration on the Raman spectrum of skin tissue. Skin, being a complex biological material, introduces unique considerations and challenges. The fluorescence interference manifests as a proportionally increasing slope below the spectrum, capturing the attention of researchers. Beyond this slope, the identification and characterization of peaks become crucial. The peaks' attributes, including their location, amplitude, and width, carry vital information about the molecules present in the skin tissue.[2] For example:

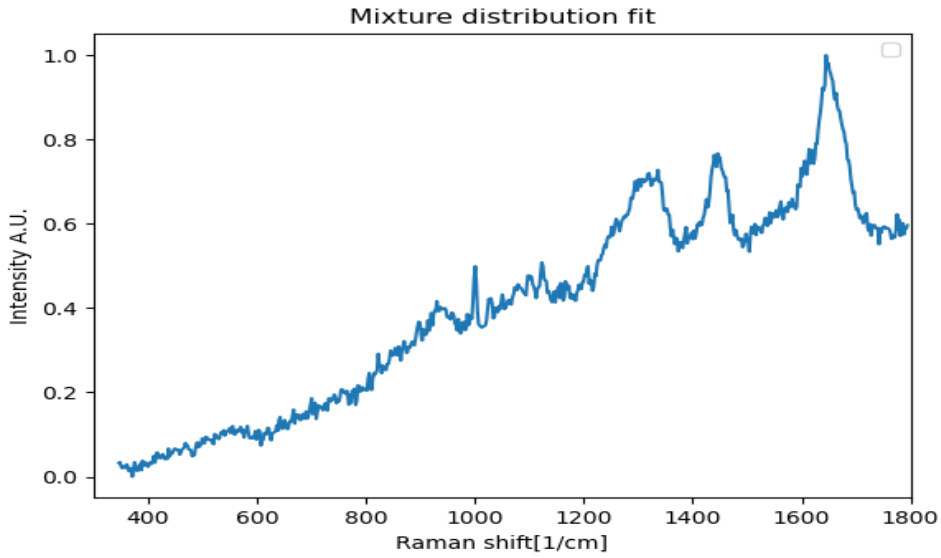


Figure 3.1: A skin tissue sample

4 Method

4.1 EM algorithm

4.1.1 Introduction

The Expectation Maximization (EM) algorithm emerges as a powerful tool for fitting distributions onto data, especially in scenarios where inherent complexities, such as missing data or mixture distributions, complicate traditional modeling approaches just as we saw in section 2.

The general idea being the assignment of weighted probabilities to data points based on the current parameter estimates then updating the parameters in order to maximize the loglikelihood function.

The likelihood function being a function of the parameters which describes how well a model with given data fits. Since the likelihood function involves a long multiplication, for computational purposes it is a general rule of thumb to take the log of the likelihood function which in turn due to the laws of logarithms becomes a long summation which is more computationally efficient.

Kullback-Leibler divergence can be used to show that the EM algorithm monotonically increases the log likelihood function at each iteration thereby converging to a local maximum, a lower bound of the log likelihood function.[4] This factor of the EM only converging to a local minimum point was exploited in [6] where EM was used to find a local extremum specifically due to another algorithm coming only in vicinity of the global maximum point, sometimes failing to find it used for initializing the EM which in turn then finds a global extremum.

Its iterative nature and reliance on the E-step and M-step render it adaptable to a myriad of applications, making it a cornerstone in statistical modeling and machine learning endeavors [3][4][5][6][7][8][9][10][11]. The EM is an alternative tool that can be used for feature extraction, specifically for fitting arbitrary models onto data, by adjusting the model parameters from initial parameter values.

4.1.2 E-step, M-step and stop criteria

The algorithm consists of 2 iterated steps, namely the E-step(expectation) and M-step(maximization) steps. In the E-step the probability of the data points lying within the model with the currently available parameters is considered. While in the M-step the model parameters are updated in a way to maximize the likelihood function.

The stop criteria for the iterations can be either the sufficient number of iterations or the arbitrarily small difference between the resulting parameters of 2 consecutive terms.

The iterative nature of the EM algorithm introduces the need for convergence criteria to determine when the algorithm has sufficiently optimized the model parameters. This can be achieved by either reaching a predetermined number of iterations or when the change in parameters between consecutive iterations falls below a predefined threshold.

4.1.3 EM in the context of mixture models

EM finds particular relevance in the realm of mixture models, where the observed data is assumed to be generated by a mixture of several underlying probability distributions. In the context of Raman spectra representation, the mixture of Gaussian and Lorentz functions becomes pertinent. EM adeptly navigates the challenges posed by such complex models, effectively estimating the parameters of each component distribution.

There are countless models which can be used for EM. In our case the mixture of the Gaussian and Lorentz functions are relevant.

4.1.4 Extending EM with GMM to include Lorentz distributions

$$q_{c_i}(\mathcal{L}_m) = \frac{\pi_m \mathcal{L}_m(x_i, \eta_m, \gamma_m)}{\sum_{n=0}^{N} \pi_n \mathcal{N}_n(x_i, \mu_n, \sigma_n) + \sum_{m=0}^{M} \pi_m \mathcal{L}_m(x_i, \eta_m, \gamma_m)} \quad \textbf{(4.1)}$$

$$q_{c_i}(\mathcal{N}_n) = \frac{\pi_n \mathcal{N}_n(x_i, \mu_n, \sigma_n)}{\sum_{n=0}^{N} \pi_n \mathcal{N}_n(x_i, \mu_n, \sigma_n) + \sum_{m=0}^{M} \pi_m \mathcal{L}_m(x_i, \eta_m, \gamma_m)} \quad \textbf{(4.2)}$$

In equations (4.1) and (4.2) an extension to the E-step of the algorithm is shown for the 1-dimensional x vector, these equations as we know from [4] are necessary in updating the parameters in the M-step, as these are non-linear exponential terms that will be handled as constants in the M-step.

In the M-step the partial derivatives with respect to the corresponding parameters of the loglikelihoods are taken to maximize the likelihood function, which results in the denominator of (4.1) every time and a numerator which we know from (4.11) - (4.16) this follows from the rules of derivatives for the natural logarithmic function. For the GMM terms the computation stays the same except for the change in the denominator due to the Lorentz terms.

Due to the partial derivates with respect to each parameter being of the form of containing the original functions and additionally linear terms (see (4.11) - (4.16)), For each corresponding parameter update in the M-step, exponential terms that include the functions itself will be written of the from (4.1) and (4.2), while the parameters which appear as linear terms in the equation are solved for.

Or yet another way to look at it also from [4] is finding the posterior distribution (4.1) and (4.2), terms which appear in the expected joint probabilities in the M-step also encompassing partial derivatives with the respected parameters. This concludes the algorithm.

4.2 Partial results

4.2.1 Synthetic data

The effectiveness of the EM algorithm is often tested using synthetic data, providing a controlled environment to evaluate its performance. The synthesis involves creating an ideal scenario, in this case, a mixture distribution of Gaussian and Lorentz components. The algorithm's ability to fit the model to synthetic data serves as a testament to its capabilities.

Synthetic data can be a controlled way to test out an algorithm by creating an ideal scenario, in this case the EM algorithm on a mixture distribution. The EM had been tested out on synthetic data which consisted of the mixture of Gaussian and Lorentz components, generated from histograms of the probability density function (pdf) of Lorentz and Gaussian components.

4.2.2 Code to generate synthetic data

```
import numpy as np
from scipy.stats import norm
from scipy.stats import cauchy
data=np.concatenate((np.random.normal(10,20,5000),cauchy.rvs(100,5,5000)))
```

4.2.3 EM applied to synthetic data

The application of the EM algorithm to the histogram of the synthetic dataset, composed of a Gaussian normal distribution with a mean of 10, a standard deviation of 20 (comprising 5000 samples), and a Lorentz (Cauchy) distribution with a median of 100, a gamma parameter of 5 (also containing 5000 samples), is visually demonstrated in Figure 3. The intricate interplay between these distributions is effectively captured by the EM algorithm, showcasing its capability to discern and model complex mixture distributions. The convergence of the algorithm to the underlying data distribution, as evidenced by the fidelity of Figure 4.1 to the original synthetic dataset, reinforces the robustness and adaptability of the EM algorithm in fitting intricate mixture models. In figure 4.1 a Gaussian component on the left while a Lorentz component on the right is shown. The outline is the sum of the 2 distributions.

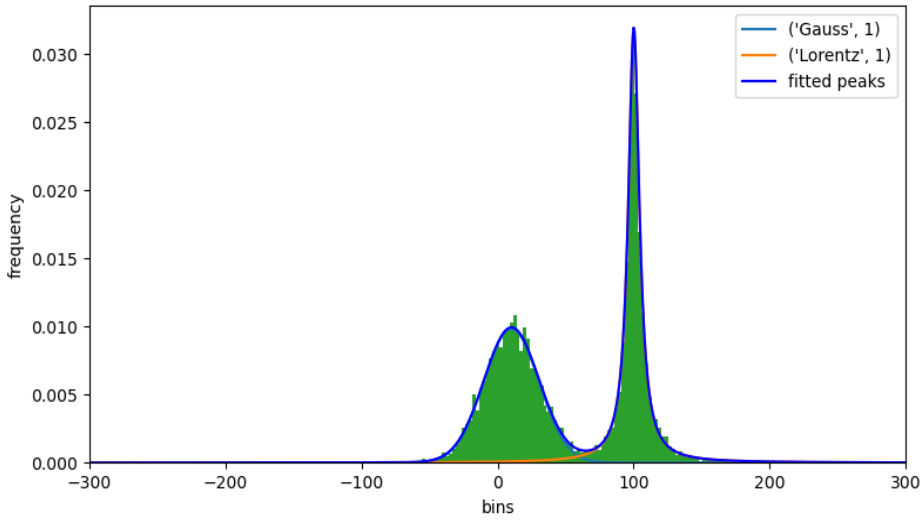


Figure 4.1: EM fitting onto synthetic data

4.2.4 EM onto partial data

In Figure 4.2, where the x-axis extends beyond 2300, the limitation becomes evident as the Gaussian component of the EM algorithm does not extend to cover the unknown region after 2300. This discrepancy raises concerns about the algorithm's competency in handling scenarios with incomplete data, especially when abrupt changes or zero values are encountered.

The root of the issue lies in the fundamental nature of the EM algorithm, which relies on mean and median computations, making it less adaptive to situations where data exhibits sudden disruptions or partial data. In this context, the inadequacy of the Gaussian distribution to seamlessly extend into the unobserved region compromises the overall fidelity of the fitted mixture model.

It is worthy to note here that the result encountered in figure 4.2 where the model of the Gauss curve being fit onto the Raman spectra, leads to a binary decision of a good and bad fit, this one being the latter as the slope due to the fluorescence is clearly not captured as we wanted, which in turn leads us to be skeptical of the algorithm at hand leading us to believe that the problem that we formulated in this context is incorrect.

Recognizing the need for a more resilient approach, an alternative algorithm is chosen, one that circumvents the limitations posed by the incomplete data in Figure 4.2. This alternative algorithm, as later showcased in Figures 5.1-5.4, proves to be more adept at accommodating the intricacies of the mixture model. Its ability to handle partial data and accurately model the distribution, even in regions with zero amplitudes, highlights its suitability for the task at hand for partial data.

The decision to deviate from the EM algorithm in favor of this alternative reflects a pragmatic response to the specific challenges posed by the dataset. By leveraging an algorithm better aligned with the characteristics of the mixture model and the nuances of the incomplete data, the revised approach ensures a more accurate and robust representation of the underlying distribution. This strategic adaptation underscores the importance of selecting an algorithm that not only fits the task requirements but also demonstrates resilience in the face of diverse data patterns and complexities.

The algorithm of choice for the rest of this paper is gradient descent, from this section onwards. Which should better suit our problem with geometrical fitting of the model.

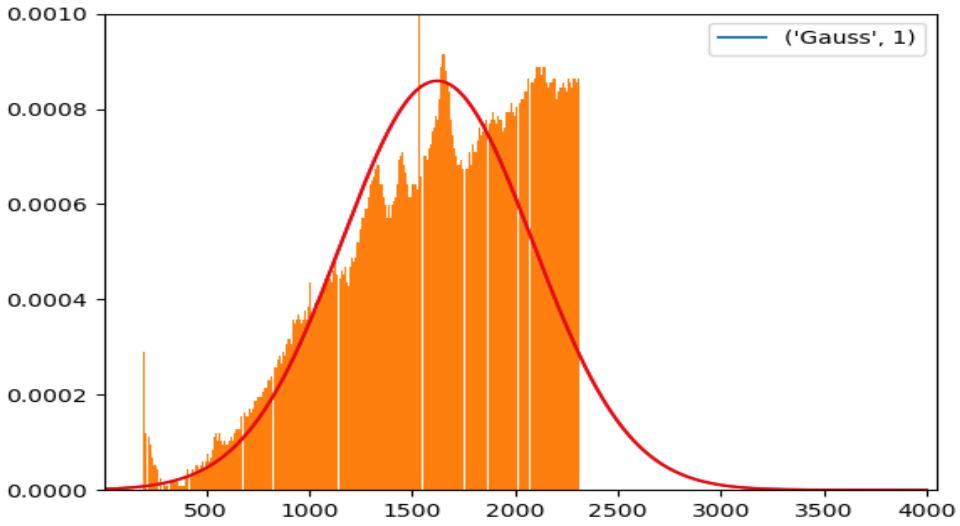


Figure 4.2: Showing bad results for EM

4.3 Gradient descent

4.3.1 Introduction

Gradient descent, a fundamental optimization algorithm, plays a pivotal role in locating minimum points in mappings of arbitrary dimensions by iteratively adjusting parameters. This iterative process aims to minimize the error originating from a given loss function. In our specific case, the goal is to find the minimum point on the quadratic loss function, representing a mapping form for a multi-variable function. Optimization involves selecting an ideal learning rate (α), a parameter (b), and a multi-variable function (f).

Gradient descent can be used in order to locate minimum points on mappings of arbitrary dimensions by adjusting parameters. Generally, we want to minimize the error originating from a loss function. Specifically in our case finding the minimum point on the quadratic loss function, which is a $R^n \rightarrow R^1$ mapping, in other words a multi variable function. For an ideal alpha, a parameter b and a multivariable function f:$b_{n+1} = b_n - \nabla f(b_n)$[14]

4.3.2 Adaptive moment estimation

An extension of stochastic gradient descent, Adaptive Moment Estimation (Adam), introduces adaptability to the learning rate. The key components of Adam include the moment variables (M and V), which are exponentially moving averages of the gradients and the squared gradients, respectively. The update step involves adjusting the weights (w) based on these moments, ensuring adaptive and efficient optimization.

An extension of stochastic gradient descent is the Adaptive moment estimation [15]:

$$m_w^{t+1} = \beta_1 m_w^t + (1 - \beta_1)\nabla_w L^t$$

(4.3)

$$v_w^{t+1} = \beta_2 v_w^t + (1 - \beta_2)(\nabla_w L^t)^2$$

(4.4)

$$V = \frac{v_w^{t+1}}{1 - \beta_2^t}$$

(4.5)

$$M = \frac{m_w^{t+1}}{1 - \beta_1^t}$$

(4.6)

$$w^{t+1} = w^t - \alpha \frac{M}{\sqrt{V} + \varepsilon}$$

(4.7)

4.4 Implementation

4.4.1 Extended Probability density functions

The implementation involves extended probability density functions, incorporating amplitudes for Gaussian (G) and Lorentzian (L) distributions within the mixture. The quadratic loss function is defined, where t represents the ideal target function consisting of given data points, and y is the model function to be fit onto t. The model y, is a

mixture distribution/function, comprising an arbitrary number of Gaussian and Lorentzian probability density functions, each with its amplitude.

Extended probability distribution functions (4.8) and (4.9), to be used in the mixture with amplitudes such that:

$$\mathcal{L}(x, L, \eta, \gamma) = L/\pi\gamma[1 + ((x - \eta)/\gamma)\text{^}2\ ]$$

(4.8)

$$amplitude: L, median: \eta, gamma: \gamma$$

$$\aleph(x, G, \mu, \sigma) = \frac{G}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(4.9)

$$amplitude: G, mean: \mu, std: \sigma$$

4.4.2 Quadratic loss function

We define a quadratic loss function: $f(y) = (t - y)^2$ where t, is the ideal target function consisting of the given data points, while y is the model function to be fit onto t.[16] In our case the model y which is the mixture distribution/function, consists of arbitrary number of Gaussian pdfs and Lorentz pdfs, from terms 0 up to N and terms 0 up to M. Amplitudes of G and L, n and m components respectively for each pdf included:

$$y = \sum_{n=0}^{N} \frac{G_n}{\sigma_{n\sqrt{2\pi}}} e^{\frac{-1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2} + \sum_{m=0}^{M} \frac{L_m}{\pi\gamma_m\left[1 + \left(\frac{x-\eta_m}{\gamma_m}\right)^2\right]}$$

(4.10)

4.4.3 Partial derivatives

A multivariable function is said to be partially differentiable if it has partial derivatives with respect to each of its variables. Both the Gaussian and Lorentzian functions are smooth and well-behaved functions, and their derivatives of all orders exist. To facilitate the underlying algorithm, partial derivatives of the loss function are crucial. For simplicity, the assumption is made that y and the loss function are partially

differentiable. Terms unrelated to the current component (Gaussian or Lorentzian) can be extracted, given their distinct parameters. Mathematical function solvers are employed to compute these partial derivatives, enhancing the efficiency of the optimization process.

The underlying algorithm requires the usage of partial derivatives of the loss function, consequently it is beneficial to provide the partial derivatives of y and then of the loss function.[14] The partial derivative of a function with respect to a variable, is its derivative with respect to that chosen variable while the others are kept constant.[17] For simplicity we assume that y and the loss function are partially differentiable. Terms which are unrelated to the current term n or m, can be taken out in all cases, since they have different parameters all together (different variables), while the Lorentz terms in case of Gaussian terms can be taken out as well and vice versa. A mathematical function solver is used to partially differentiate y.[18]

Partial derivatives of the mixture function y, in equations (4.11) to (4.16):

$$\frac{\partial y}{\partial G_n} = \frac{1}{\sigma_{n\sqrt{2\pi}}} e^{\frac{-1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2}$$

(4.11)

$$\frac{\partial y}{\partial \sigma_n} = e^{\frac{-(x-\mu_n)^2}{2\sigma_n^2}} \left[\frac{(x-\mu_n)^2}{\sqrt{2\pi}\sigma_n^4} - \frac{1}{\sqrt{2\pi}\sigma_n^2}\right]$$

(4.12)

$$\frac{\partial y}{\partial \mu_n} = (x-\mu_n) \frac{e^{\frac{-(\mu_n-x)^2}{(2\sigma_n^2)}}}{\sqrt{2\pi}\sigma_n^3}$$

(4.13)

$$\frac{\partial y}{\partial \gamma_m} = \frac{(x-\eta_m)^2 - \gamma_m^2}{\pi[\gamma_m^2 + (x-\eta_m)^2]^2}$$

(4.14)

$$\frac{\partial y}{\partial \eta_m} = 2\gamma_m \frac{x - \eta_m}{\pi[\gamma_m^2 + (\eta_m - x)^2]^2}$$

(4.15)

$$\frac{\partial y}{\partial L_m} = \frac{1}{\pi\gamma_m \left[1 + \left(\frac{x - \eta_m}{\gamma_m}\right)^2\right]}$$

(4.16)

Partial derivatives of the loss function with respect to θ the general parameter of y encompassing any of the above. We get this by combining the chain rule and partial differentiation.[17][19] Consequently, the partial derivative of the quadratic loss function with respect to an arbitrary parameter θ of y:

$$\frac{\partial f(\theta)}{\partial \theta} = -2\frac{\partial y}{\partial \theta}\big(t - y(\theta)\big)$$

(4.17)

In order to partially differentiate the quadratic loss function, specific to arbitrary parameters, we get the following if we only consider the Gaussian distributions for the loss function in (4.18) to (4.20):

$$\frac{\partial f(\theta)}{\partial G_n} = \frac{-2}{\sigma_{n\sqrt{2\pi}}} e^{\frac{-1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2} \left(t - \sum_{n=0}^{N} \frac{G_n}{\sigma_{n\sqrt{2\pi}}} e^{\frac{-1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2}\right)$$

(4.18)

$$\frac{\partial f(\theta)}{\partial \sigma_n} = -2G_n e^{\frac{-(x-\mu_n)^2}{2\sigma_n^2}} \left[\frac{(x-\mu_n)^2}{\sqrt{2\pi}\sigma_n^4} - \frac{1}{\sqrt{2\pi}\sigma_n^2}\right] \left(t - \sum_{n=0}^{N} \frac{G_n}{\sigma_{n\sqrt{2\pi}}} e^{\frac{-1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2}\right)$$

(4.19)

$$\frac{\partial f(\theta)}{\partial \mu_n} = -2G_n(x - \mu_n) \frac{e^{\frac{-(\mu_n - x)^2}{(2\sigma_n^2)}}}{\sqrt{2\pi}\sigma_n^3} \left(t - \sum_{n=0}^{N} \frac{G_n}{\sigma_{n\sqrt{2\pi}}} e^{\frac{-1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2}\right)$$

(4.20)

In order to partially differentiate the quadratic loss function, specific to arbitrary parameters, we get the following if we only consider 1 of the Lorentz distributions for the loss function in (4.21) to (4.23):

$$\frac{\partial f(\theta)}{\partial L_m} = \frac{-2}{\pi\gamma_m\left[1+\left(\frac{x-\eta_m}{\gamma_m}\right)^2\right]}\left(t - \frac{L_m}{\pi\gamma_m\left[1+\left(\frac{x-\eta_m}{\gamma_m}\right)^2\right]}\right)$$

(4.21)

$$\frac{\partial f(\theta)}{\partial \gamma_m} = -2L_m\frac{(x-\eta_m)^2-\gamma_m^2}{\pi[\gamma_m^2+(x-\eta_m)^2]^2}\left(t - \frac{L_m}{\pi\gamma_m\left[1+\left(\frac{x-\eta_m}{\gamma_m}\right)^2\right]}\right)$$

(4.22)

$$\frac{\partial f(\theta)}{\partial \eta_m} = -4L_m\gamma_m\frac{x-\eta_m}{\pi[\gamma_m^2+(\eta_m-x)^2]^2}\left(t - \frac{L_m}{\pi\gamma_m\left[1+\left(\frac{x-\eta_m}{\gamma_m}\right)^2\right]}\right)$$

(4.23)

These partial derivatives contribute to the iterative update of parameters, ensuring the convergence of the optimization algorithm towards the minimum point. The combination of gradient descent and Adam optimization [15], tailored to the intricacies of the mixture model, forms a robust framework for efficient and accurate parameter optimization in the context of the given model for the spectra.

4.4.4 Penalty term

In the pursuit of refining the Raman spectra analysis model, an additional layer of sophistication is introduced through the incorporation of a penalty term. This strategic augmentation is designed to ensure that the model remains in proximity to specified parameter values, effectively imposing a penalty for deviations. The inclusion of a penalty term serves a dual purpose.

Firstly, it acts as a mechanism to constrain the model within a predefined parameter space, aligning with the known characteristics of the analyzed Raman spectra. This is

particularly crucial in scenarios where maintaining fidelity to specific parameter values is imperative for accurate representation.

Secondly, the penalty term bolsters the robustness of the model by discouraging drastic deviations from expected parameter values. By introducing a controlled penalty for parameter divergence, the model gains a nuanced understanding of the significance of staying close to the specified values. This is especially pertinent in instances where adherence to prior knowledge or experimental expectations is paramount. The penalty term, when seamlessly integrated into the error minimization framework, transforms the optimization process into an adaptive mechanism. Instead of solely focusing on minimizing the fit error, the model now navigates the parameter space with an awareness of the penalty incurred for straying too far from the predefined values. This adaptive strategy lends a nuanced balance between fitting the data optimally and honoring a priori knowledge.

Due to partial derivatives, the corresponding Penalty term to each parameter will be of the form F=-2A(X-P) where X is the actual parameter, P is the parameter we are currently estimating and A is the consideration factor, which specifies how much the entire model should depend on the penalty term. The -2 is obtained when partially differentiating the quadratic loss function of the penalty term of added loss function D/DP(A(XP)^2)=-2A(X-P).

4.4.5 The algorithm

```
y=Raman_spectra, t=intervals_without_peaks, centrs=around_center_of_peaks

First:

Input:y,t,GaussParam

While(i<I)

       Gauss_partial_derivatives=derivatives(y,t,GaussParam)

       GaussParam=Adam(Gauss_partial_derivatives)

Output:GaussParam

Second:

Input:y,centrs,GaussParam,LorentzParam

While(i<I)

       Lorentz_partial_derivatives=derivatives(y,centrs,GaussParam,LorentzParam)

       LorentzParam=Adam(Lorentz_partial_derivatives)

Output: LorentzParam
```

In the pseudocode above, the algorithm unfolds in two distinct steps. Initially, the partial derivatives of the quadratic loss function with respect to the Gaussian components are computed. This calculation involves the input initial parameters and the t vector, representing the Raman spectra. Notably, the Raman spectra are initialized with suspected positions without peaks. Subsequently, gradient descent is applied, optionally utilizing the Adam optimization algorithm for updating the parameters but traditional method could also be used, until a satisfactory number of iterations (I) is reached.

In the second step, the outputs from the first phase become inputs for the subsequent stage. Additional inputs now encompass the Lorentz parameters. In this step, a singular Lorentz distribution is considered, a vector 5-10 indices centered around the peak positions of the Lorentz distributions. Importantly, the Gaussian components are treated as constants during this phase, streamlining the focus to precisely fit the peaks through the incorporation of the Lorentz distribution. This two-step process illustrates the algorithm's strategy in iteratively refining the model to accurately capture the underlying structure of the Raman spectra. To conclude, for every iteration the whole

vectors are checked, additionally in the second step adjacent peaks to the current peak can be taken as constants in order to avoid overlaps.

Note that the penalty term can optionally be added at the calculation of partial derivatives for the respected parameters but was removed due to the parameters not deviating enough from the initial values.

4.4.6 The program

The implementation of the program, detailed in this section, underscores the utilization of Python for its simplicity and versatility. The primary focus lies in fitting a model to Raman spectra data, incorporating both Gaussian and Lorentzian distributions. This process involves careful consideration of given parameters, specific intervals, and other decisions to optimize the model's accuracy.

Combining the information is section 4, The program is written in a programming language of choice. The vector t is separately handled for the Gaussian and Lorentzian distributions. For the Gausses, the points without peaks are considered. While for the Lorentz distributions a pivotal approach is taken around the centers. The intervals and initial parameters are discussed in section 4.5.

4.5 Priory information

4.5.1 Given parameters

The initial step involves identifying and incorporating given parameters into the model. This includes data for two Gaussian distributions and ten Lorentz distributions. These distributions are characterized by means, standard deviations, amplitudes for Gaussians, and medians for the Lorentzian, laying the foundation for subsequent fitting processes.

Initially we are given the following data to look for within the model; 2 Gaussian distributions:

	Mean	*Standard dev.*	*Amplitude*
Gauss1	1600	500	1
Gauss2	2500	1000	1

Table 4.1 Showing given approximate parameters of 2 Gauss distributions.

10 Lorentz distributions with given medians:

Median
566
853
936
1003
1095
1124
1318
1457
1607
1653

Table 4.2: Showing Lorentz centers

4.5.2 Intervals without peaks

The identification of intervals without peaks in the Raman spectra is pivotal for accurate model fitting. The specified Raman shift intervals devoid of significant peaks are carefully outlined, guiding the subsequent fitting of Gaussian intervals.

We are given the following intervals without peaks onto which we fit Gaussian intervals:

shift [1/cm]	300-500	600-800	1130-1200	1480-1500	1740-1790

Table 4.3: Showing intervals without peaks

4.5.3 Gamma values

Lorentz distributions were chosen to be fit one by one in the model specifically around the medians. This avoids fitting to unnecessarily many points, saving computational power while focusing on the prioritization of the locations.

Addressing uncertainties, particularly concerning the gamma variable of the Cauchy distributions, necessitates speculative decisions. For each peak, a meticulous exploration of one to five indices around the center points in the data vector is performed. This choice is taken due to the fact that the underlying model has the largest peak at around 5-10 index around the median therefore choosing a lower bound decreases the chance of peaks overlapping. Random batches of values in these ranges were taken using rand int function from the Numpy library.[20]

Additionally, initial vectors for gammas and amplitudes of Lorentz distributions, as well as amplitudes of Gaussian distributions, are chosen with eight random values, setting the stage for subsequent fitting endeavors.

Since the gamma variable of the Cauchy-s is unknown, for each peak 5-10 indices are checked around the center points in the data vector. Moreover, an initial vector is chosen for the gammas and the amplitudes of the Lorentz distributions with 8 random values. Same goes for the amplitudes of the Gauss distributions.

The intricate interplay of given parameters, identified intervals without peaks, and our intuitive decisions collectively shape the program's methodology. The program's efficacy hinges on a meticulous consideration of these factors, ensuring a robust and accurate fit of the model to the intricate nuances of Raman spectra data.

5 Results

5.1 The mixture fit

5.1.1 The fit

In Figures 5.1 to 5.4, the outcomes of the developed algorithm are visually represented on four randomly selected skin tissue samples. Each figure showcases the mixture fit, where individual components are amalgamated to form a comprehensive representation. The distinctive baseline under the curve, coupled with the identification of Lorentz components at specified median locations, illustrates the algorithm's effectiveness in capturing the nuanced features of Raman spectra. Importantly, the variations in the parameters result in diverse baseline shapes, distinct peaks, and consequently, unique mixture models across the samples.

On these figures results of the fit are visible on 4 randomly chosen skin tissue samples. The mixture consisting of individual components summed up. A baseline is formed under the curve, while Lorentz components are detected in the specified median locations. The results for the samples differ in terms of parameters resulting in diversly shaped base lines, different peaks and accordingly a contrasting mixture model.

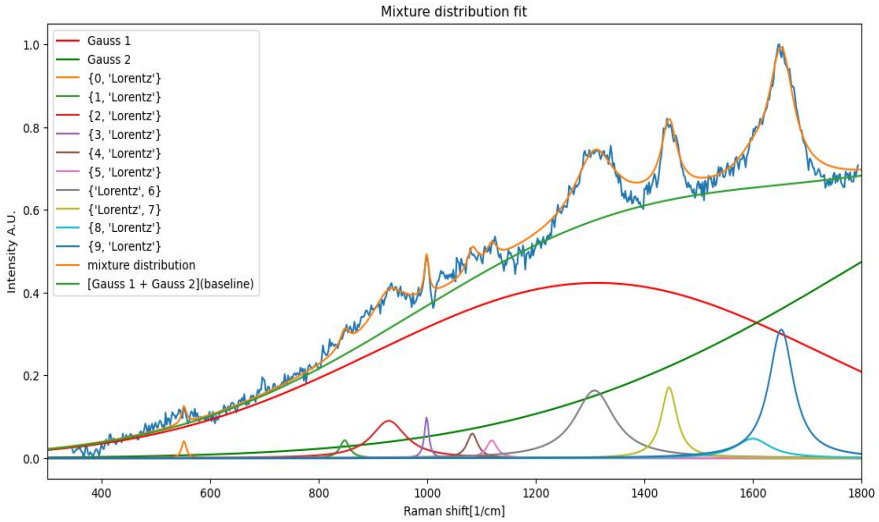


Figure 5.1: Showing mixture fit onto sample 1

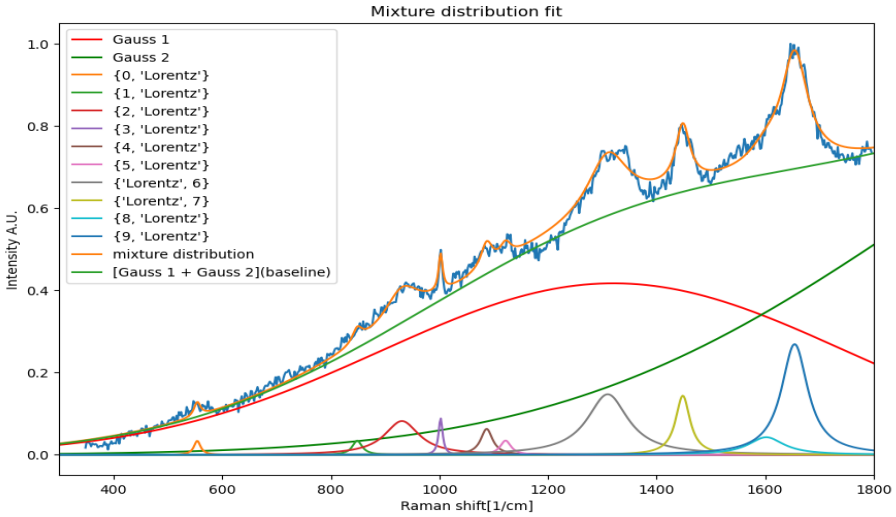


Figure 5.2: Showing mixture fit onto sample 2

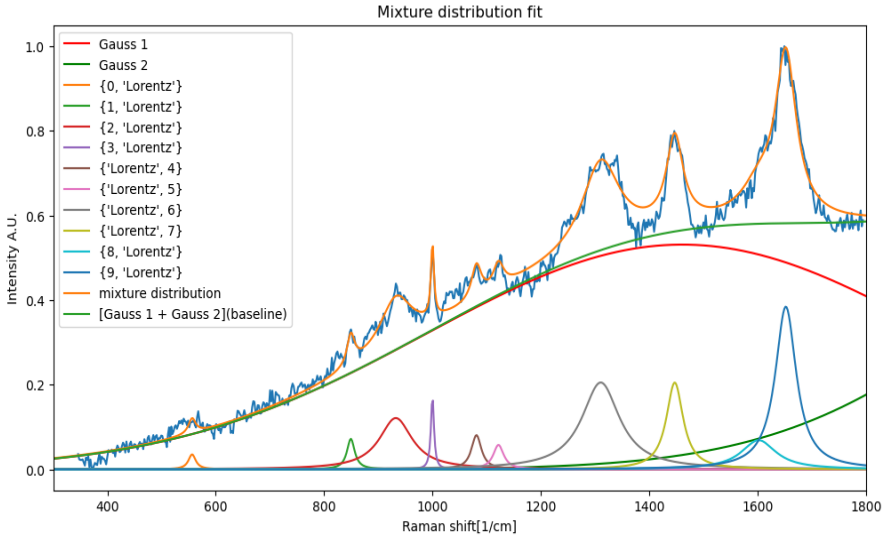


Figure 5.3: Showing mixture fit onto sample 3

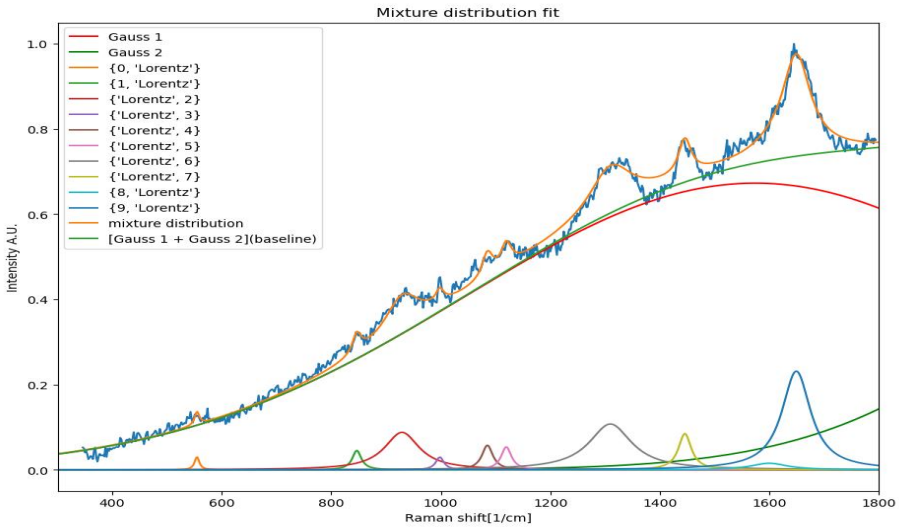


Figure 5.4: Showing mixture fit onto sample 4

5.1.2 Validation

The robustness and efficacy of the developed algorithm in capturing the intricate features of Raman spectra on skin tissue samples were validated by domain experts.

The amalgamation of individual components into a comprehensive representation, as depicted in Figures 5.1 to 5.4, was systematically examined and confirmed by experts in the field. It is worthwhile to note here that the verification is not a binary decision. The distinct baseline under the curve, coupled with the precise identification of Lorentz components at specified median locations, underscores the algorithm's ability to accurately model the complexities within the Raman spectra. This validation process revealed that variations in algorithm parameters led to diverse baseline shapes, distinct peaks, and consequently, unique mixture models across the randomly selected skin tissue samples.

The expert verification not only reinforces the credibility of the outcomes but also positions the developed algorithm as a solid starting point for further advancements in Raman spectroscopy analysis for skin tissue characterization. The algorithm's capability to adapt to different sample variations and provide consistent and meaningful results demonstrates its potential utility in diverse research and diagnostic applications within the field.

5.1.3 The cost over iteration

Figure 5.5 provides insight into the optimization process, depicting the cost function's dependency on the learning rate (alpha) and the number of iterations. The cost per iteration, assessed using Mean Squared Error (MSE) on points within intervals considered devoid of peaks, offers a quantitative measure of optimization progress. Observing the relationship between the cost and alpha reveals that a moderate increase in alpha accelerates the convergence to a constant, indicating a faster attainment of minimum points. It is essential to note the non-linearity of the model, leading to multiple local minimum points. Consequently, the appearance of bumps on the line reflects the inherent noise from error calculations during the optimization process.

These visual representations not only serve as a testament to the algorithm's efficacy in fitting complex mixture models onto Raman spectra but also offer valuable insights into

the dynamics of the optimization process, showcasing the delicate balance between learning rate and convergence in the pursuit of accurate and robust model representation. We can see that in this case choosing a higher learning rate is not necessarily an issue as long as the algorithm converges into the same state faster.

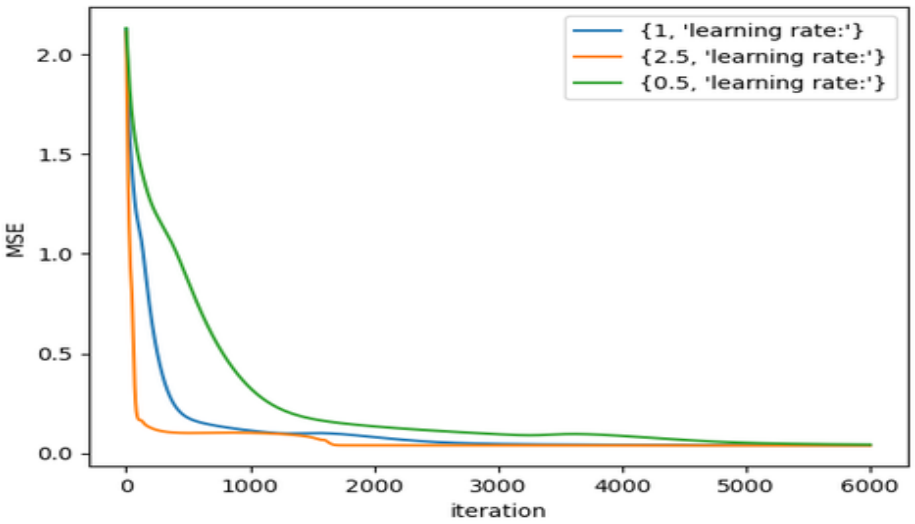


Figure 5.5 Depicting the learning rates as MSE over iteration

On Figure 5.6 we can see the cost function, dependent on alpha which is the learning rate and the number of iterations. For the cost per iteration, MSE is used on the points that are on the intervals, considered to be with peaks.

Slightly increasing alpha will result in the line reaching a constant faster meaning a minimum point faster, up to a certain extent. Bumps on the line may appear due to the noise from the error calculation, the bigger the learning rate, since the points we are fitting to contain noise in the form of “hair” on the top of the lines of the spectra which we addressed in 6.2.5.

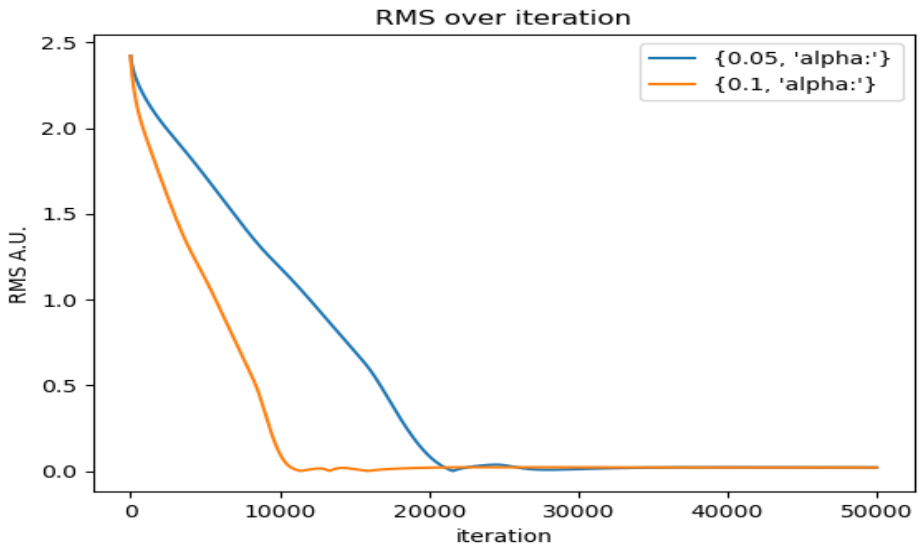


Figure 5.6: Showing MSE over iteration

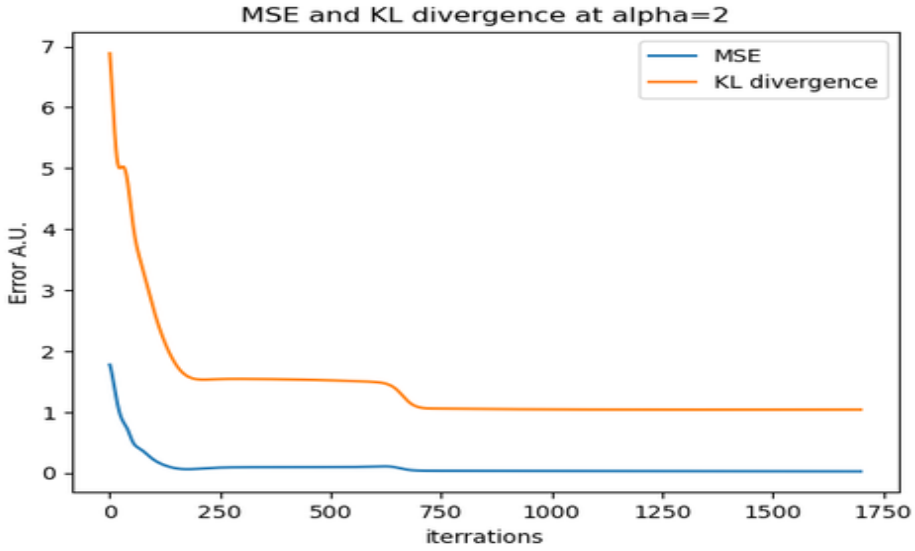


Figure 5.7 Depicting the Kullback-Leibler divergence of the model per iteration

In 5.7 the Kullback-Leibler divergence of the model is a non-negative valued function depicted for discrete data points, a relationship between the fitted model of the Gaussians, and the intervals without peaks. The lower this value the better, a value of 0 indicating no difference in the model. In 5.7 this is calculated for a randomly chosen sample at a learning rate of 2 along with the MSE.

6 Conclusion

6.1 The mixture fitting

6.1.1 The baseline

For each different Sample the weights, widths and locations of the Gaussians are adjusted accordingly, their sum forming a baseline. The sum of the two separate Gaussian components provided a sufficient geometric outline for the slope originating from the fluorescence.

In our exploration of optimizing a model for representing Raman spectra in skin tissue, the baseline fit demonstrated promising results, aligning well with the expected model and lying just below the spectrum as specified by the expert. The constraint of only two Gaussian distributions in some cases posed exceptions where more were anticipated, affecting the overall fit.

6.1.2 The Peaks

The parameters in the Lorentz distributions are adjusted appropriately according to the sample for which they had been adapted to. Their sum outlines the peaks in the spectra.

Little to no information about the peaks, except for their centers, led to a suboptimal performance compared to the baseline. The interval of fitting which is based on the decision process outlined in Section 4.3.3, contributed to this challenge. The absence of underlying information on which points to consider for peak analysis compounded the difficulty. The potential existence of more than eight peaks in a Raman spectrum added another layer of complexity.

6.1.3 Conclusion

In conclusion, the primary objective of this task was to leverage the gradient descent algorithm to optimize a model for the specific challenge posed by Raman spectra in skin tissue. This endeavor stemmed from the inadequacy of another algorithm, namely the EM algorithm, for this specific model and data. Focusing on the Raman spectrum

for skin tissue and the mixture distribution comprising a specific number of Lorentz and Gauss distributions, we chose the gradient descent as an alternative approach.

We furthermore extended the EM and gradient descent algorithms, in order to be applicable for this type of model which consisted of Gaussian and Lorentzian distributions.

This decision underscored the critical importance of selecting the right optimization algorithm tailored to the nuances of the task and dataset. Not all algorithms are universally applicable, and the choice significantly influences the model's performance. The application of gradient descent showcased its adaptability and efficacy in enhancing the model's capability to capture the unique characteristics of Raman spectra in skin tissue.

The presented results serve to emphasize the potential of extending and adapting optimization algorithms for specific data types, shedding light on both known and unknown aspects of model parameters. This exploration contributes to the broader understanding of algorithmic choices in data modeling and sets the stage for further advancements in biomedical Raman spectroscopy.

6.2 Future work

6.2.1 Adjustment of model

In the future we may use a different loss function other than the quadratic loss function, for example KL divergence which we have previously mentioned, as other loss functions may capture the fit in a better way.

The groundwork laid in this thesis with two Gaussian and eight Lorentzian distributions opens the door for future adjustments and expansions. Raman spectra are complex, and the algorithm's support for arbitrarily many parameters suggests the potential inclusion of more Gaussian distributions and the accommodation of spectra with hundreds of peaks. This avenue could lead to a more nuanced and adaptable model, better suited to the intricate nature of Raman spectra.

In this thesis we were previously acquainted with 2 Gaussian and 8 Lorentzian distributions. In the future more could be added, as such Raman spectra may contain

more Gaussian distributions and up to hundreds of peaks since the algorithm supports arbitrarily many parameters.

6.2.2 Broad Applicability: Extending Beyond Raman Spectra

The methodology and algorithms developed in this thesis, tailored for the intricate analysis of Raman spectra in skin tissue, hold significant promise for broader applications. Beyond the realm of Raman spectroscopy, the framework established here can be seamlessly extended to spectra obtained from other sources, such as Nuclear Magnetic Resonance (NMR).

The adaptability of the algorithm allows for a seamless transition to NMR spectra analysis. NMR, a powerful analytical technique employed in various scientific domains, provides insights into the structure and dynamics of molecules. The challenges encountered in biomedical Raman spectroscopy, such as background signal removal and feature extraction, are paralleled in NMR. By applying the lessons learned and methodologies developed in the context of Raman spectra, the algorithm can be adeptly repurposed for NMR data.

The synergies between different spectroscopic modalities create a unique opportunity for cross-modal collaboration. As advancements are made in the analysis of Raman spectra, they can be leveraged to enhance methodologies in other spectroscopic techniques, and vice versa. The interdisciplinary nature of this approach positions the research at the forefront of a broader scientific landscape.

Extending the application beyond Raman spectra to NMR opens avenues for comprehensive biomedical research. The ability to extract meaningful information from diverse spectral sources contributes to a more holistic understanding of molecular compositions and interactions. This, in turn, has implications for disease diagnosis, drug development, and a myriad of other applications in the biomedical domain.

6.2.3 Extending the EM algorithm

While the EM algorithm was discussed and its limitations acknowledged, future work might involve exploring its potential in a modified context. The constraints associated with the EM algorithm, such as the necessity to convert data into a histogram, could be

addressed by adapting the model to align with the algorithm's requirements. Alternatively, investigating alternative algorithms that align with the desired geometric fit could provide valuable insights into refining feature extraction methodologies.

As discussed before the EM algorithm maybe used but constrained by several factors such as the model we are fitting, and the fact that the data has to be converted into a histogram, due to the fact that our goal is a geometric fit, meanwhile the EM algorithm relies on the frequency of points in the 1-dimensional case. We may change the model to one that provides the desired geometric fit.

6.2.4 Automating Identification

With the mixture model successfully extracting features from Raman spectra, the next logical step is the development of automated identification systems. Neural networks (NNs) present a promising avenue for this purpose. By leveraging the extracted features, NNs could be trained to recognize and identify underlying materials in Raman spectra. This automation holds the potential to streamline and expedite the analysis of complex spectral data, making it more accessible and efficient.

What we now covered with the mixture model is feature extraction, meaning in the next step we may create systems, such as neural networks that can identify the underlying material based on the extracted features from the spectra. For example: a method similar to [9].

6.2.5 Improvement of the initialization of parameters

Addressing the randomness in the initialization of unknown parameters is a crucial aspect for further refinement. Introducing algorithms for improved detection of initial parameters can contribute to faster convergence to the minimum and enhance the robustness of the overall model.

For example: k-means clustering can be combined with the EM algorithm, which leads to the discussion of a similar traditional algorithm being first run on the data before entry into our algorithm, in turn improving the robustness of our algorithm.

Furthermore, preprocessing techniques such as kernel density estimation could be explored to provide a more informed and strategic initialization, by removing noise

geometrically present on top of the spectrum if we consider a different point of view to initialize the parameters as the geometric “hair” on the top of the function could cause multipolar results.

6.2.6 Final thoughts

In summation, the outlined future work encapsulates a trajectory of refinement, expansion, and automation. These prospective directions not only aim to enhance the current model's capabilities but also to contribute to the broader landscape of data modeling in biomedical Raman spectroscopy. By addressing specific challenges and leveraging emerging technologies, the research journey initiated in this thesis sets the stage for continued advancements in the quest for precise and reliable molecular characterization within biological tissues.

Acknowledgment

I would like to acknowledge and give my warmest thanks to my supervisor László György Grad-Gyenge at the BME Faculty of Electrical Engineering and Informatics, Department of Automation and Applied Informatics and to my external supervisor Dr. Miklós Veres at the Wigner Research Centre for Physics in Budapest who made this work possible. Their guidance and advice carried me through all the stages of writing my thesis.

References

[1] Movasaghi, Z., Rehman, S., & Rehman, I.U., Raman Spectroscopy for Biological and Biochemical Samples, Applied Spectroscopy Reviews, Vol. 42, 2007, pp. 493-541.

[2] Tuschel, D., 2021. Peak Shape and Closely Spaced Peak Convolution in Raman Spectra. Spectroscopy, 36(9), pp.10-14.

[3] Vu, T.K., Hoang, M.K. and Le, H.L., 2019. An EM algorithm for GMM parameter estimation in the presence of censored and dropped data with potential application for indoor positioning. ICT Express, 5(2), pp.120-123.

[4] Sridharan, R., 2014. Gaussian mixture models and the EM algorithm. Avilable in: http://people. csail. mit. edu/rameshvs/content/gmm-em. Pdf.

[5] Sandeep, P. and Jacob, T., 2016. Single image super-resolution using a joint GMM method. IEEE Transactions on Image Processing, 25(9), pp.4233-4244.

[6] Zablotskiy, S., Pitakrat, T., Zablotskaya, K. and Minker, W., 2011. GMM parameter estimation by means of EM and genetic algorithms. In Human-Computer Interaction. Design and Development Approaches: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part I 14 (pp. 527-536). Springer Berlin Heidelberg.

[7] Markov, K.P. and Nakagawa, S., 1998. Discriminative training of GMM using a modified EM algorithm for speaker recognition. In ICSLP.

[8] Xia, Y., Zhang, C., Weng, S. and Liu, R., 2005, June. Fault-Tolerant EM Algorithm for GMM in Sensor Networks. In DMIN (pp. 166-172).

[9] Arafa, A.A.A., El-Sokary, N., Asad, A. and Hefny, H., 2019. Computer-aided detection system for breast cancer based on GMM and SVM. Arab Journal of Nuclear Sciences and Applications, 52(2), pp.142-150.

[10] Momen, M.A., Khalid, M.A. and Oninda, M.A.M., 2019, October. Fpga-based acceleration of expectation maximization algorithm using high-level synthesis. In 2019 Conference on Design and Architectures for Signal and Image Processing (DASIP) (pp. 41-46). IEEE.

[11] Kertel, Maximilian & Pauly, Markus. (2022). Estimating Gaussian Copulas with Missing Data.

[12] Sheehy, G., Picot, F., Dallaire, F., Ember, K., Nguyen, T., Petrecca, K., Trudel, D. and Leblond, F., 2023. Open-sourced Raman spectroscopy data processing package implementing a baseline removal algorithm validated from multiple datasets acquired in human tissue and biofluids. Journal of Biomedical Optics, 28(2), pp.025002-025002.

[13] Zhao, J., Lui, H., McLean, D.I. and Zeng, H., 2007. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. Applied spectroscopy, 61(11), pp.1225-1232.

[14] Wikipedia: Gradient descent, https://en.wikipedia.org/wiki/Gradient_descent (revision 17:58, 23 October 2023)

[15] Kingma, Diederik; Ba, Jimmy (2014). "Adam: A Method for Stochastic Optimization". arXiv:1412.6980 [cs.LG].

[16] Wikipedia: Loss function, https://en.wikipedia.org/wiki/Loss_function (revision 17:58, 23 October 2023)

[17] Wikipedia: Partial derivative, https://en.wikipedia.org/wiki/Partial_derivative (revision 17:58, 23 October 2023)

[18] Wolfram Alpha LLC. 2024. Wolfram|Alpha., https://www.wolframalpha.com/ (revision 17:58, 23 October 2032)

[19] Wikipedia: Chain rule, https://en.wikipedia.org/wiki/Chain_rule (revision 17:58, 23 October 2023)

[20] Harris, C.R. et al., 2020. Array programming with NumPy. Nature, 585, pp.357–362.