Norwegian University
of Life Sciences

Master's Thesis 2017 60 ECTS
Faculty of Science and Technology

An optimized algorithm for separating scattering and chemical absorption in biomedical infrared spectroscopy and imaging

Johanne Heitmann Solheim

Preface

The year spent working on this thesis has truly been the most exciting and inspiring of all my years of studying. I have been introduced to a whole new field, and during this time I got to participate in so many different things. At the FTIR workshop in Berlin I presented this work in a poster, and a manuscript for a paper with the title "An open source code for Mie scatter correction of infrared microscopy spectra of cells and tissues" is soon ready for submission. None of this could have been achieved alone, and so I would like to thank the people who have guided and supported me, such that I now can complete this master degree.

First of all, I want to express my deepest gratitude to my supervisor Achim Kohler for excellent guidance and support throughout this year. Thank you for your constant encouragement, for challenging me and for patiently devoting your time – I could think of no better supervisor. I would also like to thank Tania Konevskikh, whose work I have continued, for helpful thoughts and feedback.

Further, I am grateful for being included in an inspiring and diverse team, and for this I want to thank BioSpec Norway for such a warm welcome. A special thanks to Boris Zimmermann for taking your time to explain different aspects of such a broad field. I also want to thank Stanislau Trukhan for further expanding this field, and for keeping me company at the office when the work dragged out.

I would also like to thank the research groups I have had the pleasure to collaborate with. I have had many enlightening discussions with Carol Hishmugl and the research group at University of Wisconsin in Milwaukee, with Sugato Ray, Ghazal Azarfar, Alex Schofield, Nick Walter and Sarah Patch. I would also like to thank Reinhold Blümel for theoretical insight and inspiration. At Ruhr University in Bochum I've had the pleasure to work together with Dennis Petersen, Frederik Grosserüschkamp and Claus Gerwert. Thank you for engaging discussions and providing me with data.

Further I want to thank Francisco Peñaranda, Nick Stone, Ganesh Sockalingum and Josep Sulé-Suso for providing measured spectra for feedback on this work.

Last but not least, I want to express my sincere gratitude to my friends and family for their constant love and support. My parents have always encouraged me to follow my interest, and my sisters have always been great a inspiration to me. Thank you, Malin and Alise, for being faithful passengers on this roller coaster, and for making Ås the greatest place on earth. And thank you Martin, for your endless love and enthusiasm, and for proof reading and giving feedback on this thesis.

Ås, December 15th 2017

Johanne Heitmann Solheim

Abstract

Over the past decades, infrared spectroscopy of biological samples has been developed to a promising tool for non-destructive biochemical analysis. Infrared absorbance spectra provide molecular fingerprints. However, single cells and tissues cause complex Mie scattering features in infrared absorbance spectra contaminating the pure chemical signatures. Several pre-processing methods have been proposed to handle scattering in infrared spectroscopy. The Mie correction [24, 5, 28, 26] based on extended multiplicative signal correction (EMSC) [32, 18, 21, 35, 34] is currently considered as the most powerful tool for separating Mie scattering and biochemical absorption in infrared spectra of cells and tissues.

Kohler et al. [24] developed an algorithm based on EMSC that could successfully predict Mie scattering features and remove them from infrared absorbance spectra. Bassan et al. developed the Mie EMSC model further to handle the so called dispersion effect. The model was implemented in an iterative algorithm, and a compiled program for Mie correction was published [5]. This program is currently the mostly used pre-processing tool for infrared spectra of cells and tissues in the diagnosis of cancer by infrared imaging. However, the algorithm is observed to be strongly biased, since corrected spectra adapt features of the reference spectrum. During recent years, Konevskikh et al. improved the Mie EMSC model further, however a user-friendly program based on the improved algorithm is not yet available [28, 26].

The main aim of this thesis is to further develop the Mie correction algorithm, such that a user-friendly program for Mie correction can be published. This is achieved by proposing a number of improvements to the Mie correction algorithm related to stabilization and optimization. In addition, there is a need for establishing a simulated data set with known pure absorbance spectra and scatter features that mimic measured apparent absorbance spectra, in order to validate different features of the algorithm.

The improvements of the Mie EMSC correction algorithm include a number of aspects. The algorithm presented in this thesis sets the number of principal components in the Mie EMSC model automatically by the program, based on a desired level of explained variance in the Mie extinction curves. A flexible stop criterion, based on the convergence of a forward Mie EMSC model is implemented. Further, the initialization parameters are standardized by controlling the scaling of the reference spectrum. Additional stability is gained by weighting the reference spectrum and by setting negative parts of the reference spectrum to zero. A simple quality test for evaluating the correction based on the error of the forward model is implemented, which is used to optimize the initialization parameters. In order to validate the algorithm, a set of absorbance spectra mimicking measured apparent absorbance spectra was simulated. In the simulations, the underlying pure absorbance is known, and scattering features were based on measured spectra. The simulated spectra were used for validation, and to assess critical features of the algorithm. We demonstrate that the correction is not biased by the initial reference spectrum and that a more reliable amide I peak position is retrieved. Sensitivity towards the initialization parameters is further reviewed. It is further demonstrated that the estimated scatter parameters from the EMSC model are meaningful and can be used for clustering of samples with respect to morphological characteristics. The advantage of pre-processing for a subsequent multivariate analysis by chemometrics and machine learning is discussed and suggestions are made how the algorithm can be employed on big spectral data from FTIR imaging. As a result of the proposed improvements, a user-friendly code for correcting highly Mie scatter-distorted absorbance spectra is published at https://bitbucket.org/biospecnorway/mie-emsc-code.

Sammendrag

Infrarødspektroskopi av biologiske prøver har blitt utviklet til et lovende verktøy for ikke-destruktiv biokjemisk analyse gjennom de siste tiårene. Infrarøde absorbansspektre representerer molekylære fingeravtrykk. Enkeltceller og vev forårsaker imidlertid komplekse Mie-spredningsegenskaper i infrarøde absorbansspektre som forurenser de rene kjemiske signaturene. Flere prosesseringsteknikker har blitt foreslått for å håndtere spredning i infrarødspektroskopi. Mie-korreksjon [24, 5, 28, 26] basert på extended multiplicative signal correction (EMSC) [32, 18, 21, 35, 34] betraktes for tiden som det kraftigste verktøyet for å separere Mie-spredning og biokjemisk absorpsjon i infrarøde spektra av celler og vev.

Kohler et al. [24] utviklet en algoritme basert på EMSC som har vist seg å kunne predikere og fjerne Mie-spredning fra infrarøde absorbansspekter. Bassan et al. [5] utviklet Mie EMSC-modellen videre for å håndtere den såkalte dispersjonseffekten. Modellen ble implementert i en iterativ algoritme, og et kompilert program for Mie-korreksjon ble publisert [5]. Dette programmet er for tiden det mest brukte pre-prosesseringsverktøyet for infrarøde spektra av celler og vev i kreftdiagnose ved infrarød avbildning. Imidlertid er det blitt observert at algoritmen tilpasser de korrigerte spektra etter egenskaper av referansespekteret. I de senere år har Konevskikh et al. forbedret Mie EMSC-modellen ytterligere, men et brukervennlig program basert på den forbedrede algoritmen er ennå ikke tilgjengelig [28, 26].

Hovedformålet med denne oppgaven er å videreutvikle Mie-korreksjonsalgoritmen, slik at et brukervennlig program for Mie-korreksjon kan publiseres. Dette oppnås ved å foreslå en rekke forbedringer av Mie-korreksjonsalgoritmen knyttet til stabilisering og optimalisering. I tillegg er det et behov for å etablere et simulert datasett hvor de rene absorbansspektre er kjent og med spredningsavtrykk som etterligner målte absorbansspekter, for å validere forskjellige egenskaper ved algoritmen.

Forbedringene i Mie EMSC-korreksjonsalgoritmen inkluderer en rekke aspekter. Algoritmen som er presentert i denne oppgaven setter antall prinsipalkomponenter i Mie EMSC-modellen automatisk, basert på et ønsket nivå av forklart varians i Mie utslukkingskurvene. Et fleksibelt stoppkriterium, basert på konvergensen av Mie EMSC-modellen, er implementert. Videre er initialiseringsparametrene standardisert ved å skalere referansespektret. Ytterligere stabilitet oppnås ved å vekte referansespektret og ved å sette negative deler av referansespektret til null. En enkel kvalitetstest for evaluering av korreksjonen basert på feilen i fremovermodellen er implementert, og brukes til å optimalisere initialiseringsparametrene. For å validere algoritmen simuleres det et sett med målte absorbansspekter. I simuleringene er den underliggende rene absorbansen kjent, og spredningsegenskaper ble basert på målte spektra. De simulerte spektrene ble brukt til validering, og for å vurdere kritiske trekk ved algoritmen. Vi demonstrerer at korreksjonen ikke er påvirket av det opprinnelige referansespekteret, og at en mer presis amide I topposisjon er oppnådd. Følsomhet overfor initialiseringsparametrene blir ytterligere gjennomgått. Det er videre påvist at de estimerte spredningsparametrene fra EMSC-modellen er meningsfulle og kan brukes til clustering av prøver med hensyn til morfologiske egenskaper. Fordelen med forhåndsbehandlingen for en etterfølgende multivariate analyse ved kjemometrisk og maskinlæring diskuteres og forslag til hvordan algoritmen kan brukes på store spektrale data fra FTIR-bildebehandling presenteres. Som et resultat av de foreslåtte forbedringene, publiseres det en brukervennlig kode for korrigering av Mie spredning-forvrengt absorbansspektra ved https://bitbucket.org/biospecnorway/mie-emsc-code.

Contents

Chapter 1

Introduction

During the last decades, infrared spectroscopy has developed to one of the most prominent tools for non-destructive biochemical characterization of biological materials. This development started in the 1990s, when Naumann et al. [41] showed that FTIR spectra of microorganisms can be used for identification of microorganisms on different phenotypic levels. For identification by FTIR spectroscopy, microbial cells were cultivated under strictly controlled conditions and subsequently measured as dried thin films on infrared-transparent substrates. The proof that the infrared characterization of microbial cells can be used as a stable phenotypic fingerprint, allowing in some cases even the identification of microorganisms on a strain level, is today considered as one of the most important findings in the development of infrared spectroscopy in biology. In the subsequent years, FTIR spectroscopy of microorganisms developed into a high-throughput technique for phenotypic characterization of microorganisms. This trend was further promoted by the development of highly sensitive high-throughput spectrometers for characterization of thin films of dried cell solutions. High-throughput FTIR spectroscopy is today routinely used in biology for phenotyping. Examples are the identification of all types of microorganisms [41, 56, 45, 46, 50] and the evaluation of the response of microbial cells to different environmental conditions [49, 12]

A second milestone in the development of instrumentation for infrared spectroscopy of biological materials, was the invention of infrared microscopes that allowed the characterization of single cells and tissues. Infrared microspectroscopy can be used for the characterization and phenotyping of plant material, such as pollen grains [59, 22] in questions related to health [62], development of vegetation and monitoring biodiversity and climate change [2].

Infrared microspectroscopy has further been used for the characterization of single eukaryotic cells in the development of infrared spectroscopy as a tool for cancer diagnosis [29], and in general for the characterization of cancerous and healthy tissues in the medical field or for diagnosis of Alzheimer disease [40, 10, 37, 6, 15]

In both high-throughput infrared spectroscopy and infrared microspectroscopy of biological materials, infrared absorbance spectra are used, since they are approximately proportional to the concentrations of the absorbing components in the material and the thickness of the material under investigation. Infrared absorbance spectra are obtained by measuring the attenuation of infrared radiation transmitting through a sample. Experimentally, this is realized by probing the infrared radiation beam first without the sample, an then the transmitted radiation with the sample. The transmitted radiation is in an ideal case only attenuated by chemical absorption. Unfortunately, in most practical situations, attenuation of the electromagnetic radiation by the sample is not only caused by chemical absorption. Measured infrared absorbance spectra show in general two different types of attenuation: attenuation that is caused by scattering, and attenuation that is due to chemical absorption. The overlap of scattering and absorption features in infrared absorbance spectra of biological material is a major problem for the spectral interpretation and multivariate data analysis. Therefore, without processing tools that allow separating scattering and absorption features, it is impossible to distinguish between attenuation caused by chemical and physical properties of the sample.

Depending on the biological sample under investigation, different types of physical attenuation occur. The most prominent absorbance variation caused by physical properties of the sample is the variation due to sample thickness resulting in a scaling effect in absorbance spectra. In neither high-throughput FTIR spectroscopy, nor in infrared microspectroscopy, the samples thickness can be easily controlled in most practical cases. The variation in sample thickness leads to multiplicative effects in the absorbance spectra [25]. Thus, scaling is an inevitable part of the pre-processing of absorbance spectra. Further, diffuse scattering caused by for example rough surfaces and variations in the infrared radiation source, may lead to baseline shifts in the absorbance spectra [25]. In addition, in absorbance measurements of almost perfect thin films, so-called fringes are observed [27]. Multiplicative effects and baseline effects occur typically in high-throughput spectroscopy of thin films. Fringes may occur, although they are rare.

In infrared microspectroscopy, single cells and tissues are observed to be effective scatterers, and the scattering has been interpreted as Mie scattering [38]. In infrared absorbance spectra, Mie scattering introduces gross baseline variations and shifts in peak heights and positions. Mie scattering is caused by the samples being approximately spherical, and with a size comparable to the wavelength in the mid and near infrared region of the electromagnetic spectrum, which is used to probe the samples. Without pre-processing, shifts in peak position can be misinterpreted for example in the analysis of protein structure, and lead to erroneous conclusions.

In order to handle situations where scattering contributes significantly to the infrared absorbance spectra, several pre-processing methods have been proposed and are in use. A standard approach for suppressing broad baseline variations is by taking the first or second derivative of the spectra [47, 61]. Scaling variations can be successfully removed by vector normalization [3]. Combining derivatives with vector normalization is a often used approach in high-throughput spectroscopy of microorganisms, where spectra have a high quality and scatter effects are a minor problem. A different approach that allows the removal of scaling variations and constant baseline shifts in a model-based approach is multiplicative signal correction (MSC) [32, 21, 18].

In infrared microspectroscopy, where more complex scattering features are observed, the standard pre-processing approaches that handle baseline shifts and scaling effects only, are not sufficient. Situations where scattering leads to non-constant baseline shift could be handled with adding additional terms to the MSC model, such as linear and quadratic wavenumber dependent baseline shifts [35, 34]. The extended model is named extended multiplicative signal correction (EMSC), and is a powerful pre-processing method favoured for its modularity. In principle, any terms can be added to the model, enabling suppression of different chemical constituents or physical phenomena. One of the great advantages of the EMSC model is that the scatter information is not lost after correction. As the baseline variations are a direct consequence of the optical properties of the sample, the parameters from the EMSC correction can carry important information about the sample [25, 60, 59].

In the past decade, several pre-processing methods has been proposed for handling the case of Mie scattering in FTIR spectroscopy [24, 5, 16, 27]. For rare cases where the sample is almost perfectly spherical and homogeneous, and both the refractive index and radius is know *a priori*, a model-based pre-processing method based on an iterative algorithm was proposed by van Dijk et al. in 2013 [16]. For biological samples however, this model is not applicable, as the samples are in reality not perfectly spherical and homogenous. In addition, the refractive index and radius are in most practical cases unknown. A pre-processing method which has proven to yield good results for correcting Mie scattering in absorbance spectra from biological samples, is the method proposed by Kohler et al. in 2008 [24]. It was shown that by employing a meta-model for describing the scattering, incorporated in an EMSC model, Mie scattering

could be predicted and removed from the absorbance spectra. In the model, a set of Mie extinction curves was calculated from the van de Hulst approximation formula for scattering by non-absorbing, homogeneous spheres. The set of extinction curves was then compressed into a small number of loadings by principal component analysis (PCA). The loadings were included in an EMSC model, resulting in a predicative forward model for Mie scattering correction.

Further development of the model included an attempt to handle the resonant case, i.e. the fact that the real part of the refractive index undergoes a fluctuation in the presence of an absorbance resonance. The formula governing the scattering was however still describing scattering by a non-absorbing sphere. This version of the EMSC model was called resonant Mie scattering (RMieS) EMSC, and the model was implemented in an iterative algorithm. Challenges were however observed with respect to the corrected spectra being strongly biased by the reference spectrum used to initiate the algorithm. In addition, the algorithm was considered time-consuming, due to the Kramers-Kronig transformation that was employed in each iteration step. The Kramers-Kronig relation is used to calculate the real fluctuating part of the refractive index from the imaginary part of the refractive index.

It was later shown that the absorption and scattering can not be treated as independent and additive effects [31], which led to the further improvement of the RMieS-EMSC by employing the van de Hulst approximation formula for absorbing spheres, using the imaginary part of the refractive index. Improvements were also made with respect to speed optimization of the iterative algorithm. This was done by reducing the number of parameters in the Mie meta-model from three to two, as well as replacing the Fourier transform by the Hilbert transform in the calculations of the fluctuating part of the real part of the refractive index. This resulted in what is known as the fast iterative Mie scatter correction algorithm. It was observed that the corrected spectra moves away from the initial reference spectrum, and tends towards the chemical features of the measured spectrum with increasing number of iterations.

After the RMieS-EMSC model was proposed in 2010, a compiled program was published, which allowed the correction of the dispersion effect. This is currently the only available Mie correction code. Konevskikh et al. [28] addressed the challenges related to the 2010-algorithm, however a Mie correction code according to the improved model is not yet available. In order to publish a user-friendly program a number of improvements of the algorithm are needed. The number of loadings used in the Mie meta-model should be set automatically by the program, and the stop criterion should ensure a stable termination of the iterative algorithm. In addition, measures should be made to stabilize the baseline correction, an the initialization parameters should be standardized. Further, a tool for simple quality testing of the corrected spectra is desired, in order to perform a preliminary sorting of whether the correction can be considered successful or not.

The overall aim of this thesis was to continue the work on the Mie scatter correction algorithm, by further development and validation of the algorithm and by development of a user-friendly matlab code that can be share with the infrared biomedical community. The specific aims of the thesis were: (1) To implement a functionality in the algorithm to automatically set the number of components used in the Mie meta-model. (2) To revise the stop criterion and to ensure a stable correction based on the convergence of the forward model. (3) To standardize the initialization parameter ranges used in the EMSC model. (4) To obtain a stable algorithm for correction allowing the automated correction of a large number of spectra. (5) To introduce a simple quality test to validated if a correction is successful or not. (6) To validate the algorithm with respect to its ability to retrieve the true pure absorbance spectrum. (7) To assess critical features of the algorithm such as the dependency on the reference spectrum, the ability to retrieve the amide I peak position and its sensitivity towards initialization parameters. (8) To publish a user-friendly Matlab code for Mie correction, and a manual for the use of the program.

The thesis is structured as follows. In the theory chapter of this thesis, the scattering and absorption if infrared radiation by biological materials is described from a physical point of view. This is followed by the derivation of a multivariate EMSC model which is modelling the electromagnetic theory of scattering and absorption of infrared radiation by cells and tissues. The iterative Mie scatter correction algorithm is explained in detail in the method chapter. In addition, a method for simulating apparent absorbance spectra with known underlying pure absorbance spectrum is presented in the method section. The simulated spectra are used for validation. Improvements and validation of the algorithm are described in the result and discussion section.

The dependency on the reference spectrum used for initiating the algorithm, and the ability to retrieve the true amide I peak position is is further reviewed. Subsequently, the sensitivity towards the initialization parameters is assessed through examples which illustrates the the effect of parameter choice and adjustment. This section serves as a user guide for the Mie correction code which is published at https://bitbucket.org/biospecnorway/mie-emsc-code.

Chapter 2

Theory

2.1 Infrared spectroscopy

Spectroscopy refers to the study of how electromagnetic waves interact with matter. This includes measurements and interpretation of absorption, emission and scattering by different materials. Such measurements can be used to identify substances, or to obtain information on physical and chemical properties of different materials. When analyzing biological materials by chemical and biochemical analysis techniques, components are extracted from the material and the structure of the material is destroyed. Spectroscopic techniques however, allow to investigate biological material non-destructively.

As the name implies, infrared (IR) spectroscopy refers to the study of interaction between infrared radiation with material. The infrared range of the electromagnetic spectrum comprises in general the near, mid and far infrared range. In this thesis, we refer to infrared spectroscopy as the study of interaction of mid-infrared radiation with material. The mid-infrared region of the electromagnetic spectrum refers to wavelengths in the region between approximately $2.5\,\mu\mathrm{m}$ to $25\,\mu\mathrm{m}$. Nature arranged it in a way, that fundamental chemical vibrations are located in the mid-infrared range.

Fourier transform infrared (FTIR) spectroscopy is today the most used instrumental technique for infrared spectroscopy. The name Fourier transform infrared spectroscopy refers to the fact that in this technique, a mathematical process is needed for converting the measured signals into the IR spectrum. Most common FTIR spectrometers use a broadband light source radiating in the mid- and near-infrared region of the electromagnetic spectrum. IR spectra can then be produced for a wide spectral range, for wavelengths between $2.5\,\mu\mathrm{m}$, $20\,\mu\mathrm{m}$. New techniques for measuring infrared spectra are developed, such as the use of so-called quantum cascade lasers [7] and ATR infrared spectroscopy [14, 13].

IR absorbance spectra are recorded by passing IR radiation through a sample, and measuring the spectral attenuation. Characteristic absorption signatures are obtained when the incident radiation is attenuated by molecular absorption: Photons in the infrared region of the electromagnetic spectrum do not have enough energy to excite individual electrons in the atoms. However, the energy can be sufficient to excite vibrational states in molecules from the ground state, by stretching and bending covalent bonds. When IR radiation in the mid-infrared region hits a molecule, energy at wavelengths corresponding to a specific vibrational transition from the ground state to the first excited state are absorbed. Radiation in the near-infrared region has higher energy, and allows for transitions from the ground state to overtones. The energy needed for the spectral absorption is characteristic for each molecule. Therefore, absorbance spectra can be used to determine the chemical content of a sample.

Figure 2.1 illustrates how absorbance spectra are recorded. IR radiation from the source, given by the intensity $\tilde{I}_0$, is first passed through the optical setup with an empty slide. The intensity I_0 transmitted through the empty optical setup is recorded at the detector. The measured intensity I_0 is often referred to as the background intensity. Then the sample is mounted

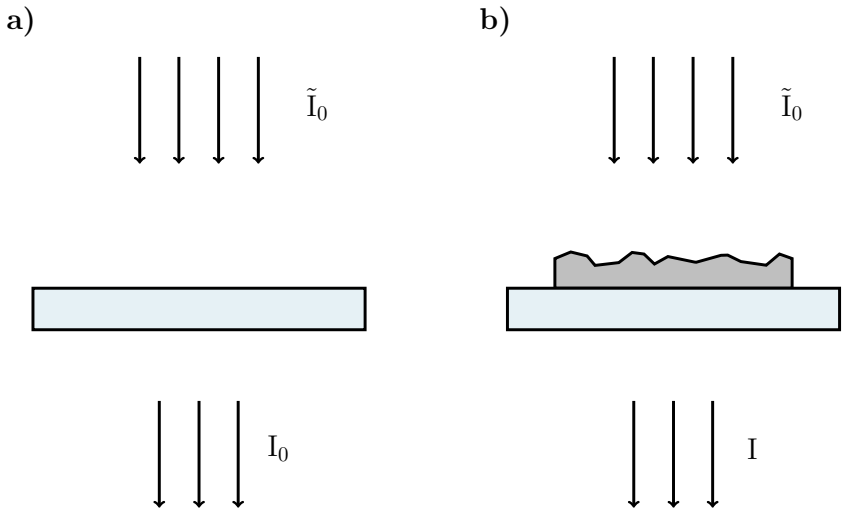


FIGURE 2.1: Recording of IR absorbance spectra. a) Firstly, the reference intensity is measured by probing the beam with an empty sample holder. b) The transmitted intensity through the sample is thereafter measured with the sample mounted on the slide.

on the slide, and the transmitted intensity I, is recorded. The transmittance T is then calculated as

$$T(\tilde{\nu}) = \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} \tag{2.1}$$

in terms of wavenumbers $\tilde{\nu}$. The absorbance Z is obtained as

$$Z(\tilde{\nu}) = -\log_{10} T(\tilde{\nu}) \tag{2.2}$$

In the following, we neglect the optical setup and measurement surroundings. The situation will thus be presented as if I_0 is the intensity of the incident light, as this is the radiation which is affected by the sample only. In IR spectroscopy it is common practice to use the unit of wavenumbers $\tilde{\nu}$. In vacuum, the wavenumber relates to the frequency by $f = c \cdot \tilde{\nu}$, which leads to wavenumbers and frequency being used interchangeably in this thesis.

The main building blocks of biological material, such as cells and tissues, are proteins, lipids, and carbohydrates. Therefore, IR absorbance spectra obtained from such samples share similar basic features. An example of a typical spectrum is presented in Fig. 2.2. The spectrum in Fig. 2.2 is an absorbance spectrum obtained from an extracellular matrix called Matrigel [5], which is is an artificially produced basement membrane used as a substrate for growing cells. Matrigel consists of a gelatinous protein mixture, where the main content is structural proteins such as laminin, collagen IV and entactin.

Absorbance spectra are often divided into two main regions, the functional group region and the fingerprint region. At wavenumbers lower than $\sim$ 1,500 cm^{-1} we find the fingerprint region. This region provides information about the molecule as a whole. At wavenumbers above this we find the functional group region. In addition, the region between 1,800 cm^{-1} and 2,800 cm^{-1} is sometimes referred to as the silent region, as this region is usually chemically inactive.

It is important to note that absorbance and absorption refer to two different issues. In electromagnetic theory, **absorption** refers to the physical phenomenon by which a part of the energy of the radiation is absorbed by the medium through which the wave propagates. **Absorbance** on the other hand, is a measure of attenuation of an electromagnetic wave, defined

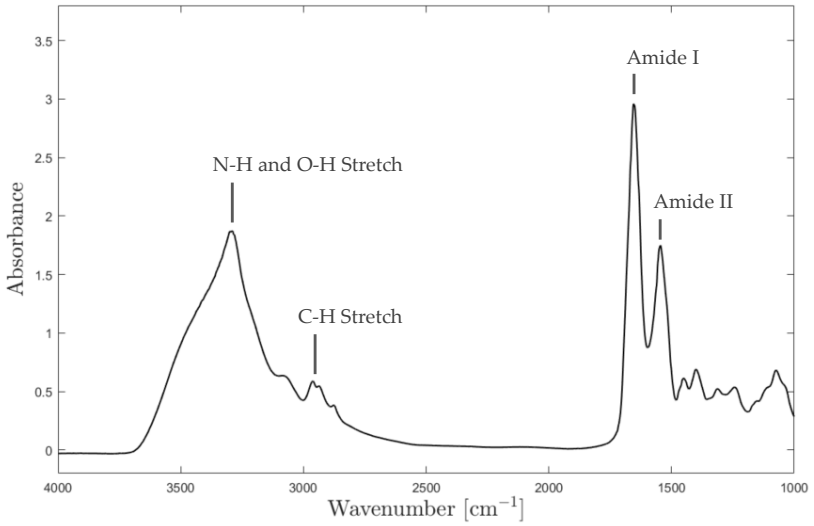


FIGURE 2.2: The Matrigel absorbance spectrum [5]. Matrigel is an artificially produced basement membrane consisting of mainly proteins.

by Eq. 2.2. Ideally, absorbance spectra show only characteristics due to chemical absorption, but in practice absorbance spectra also contain attenuation due to scattering. Typical examples of IR absorbance spectra that contain considerable attenuation due to scattering are absorbance spectra obtained from biological cells and tissues. Biological cells and tissues are highly scattering samples in the infrared region of the electromagnetic spectrum, since their morphology changes on the same scale as the size of the wavelength of the infrared radiation, i.e. on micrometer scale [38, 5]. In order to account for the fact that absorbance spectra may contain strong contributions from attenuation due to scattering, the expression apparent absorbance was introduced. **Apparent absorbance** refers to absorbance spectra that, in contrast to pure absorbance spectra, contain scattering features. The apparent absorbance spectrum is the measured spectrum, describing attenuation caused by both chemical and physical properties, while the **pure absorbance** expresses only chemical attenuation, i.e. absorption. In this thesis, pure absorbance spectra are denoted Z_{pure} while apparent absorbance spectra are denoted Z_{app}.

2.1.1 Definition of the pure absorbance spectrum

In the ideal case of IR absorption spectroscopy, the absorbance is caused by molecular absorption only. In order to refer to this ideal case, spectroscopist use the term pure absorbance spectrum. To define the pure absorbance spectrum, we consider a sample which is shaped as a thin film of thickness d. Figure 2.3 shows an incident plane electromagnetic wave travelling in x-direction (downwards) and with intensity I_0. The radiation propagates through the sample, and a portion of the radiation is absorbed. At the detector, the intensity I is measured. The intensity of the incoming radiation can be expressed in terms of the incoming electric field amplitude E_0

$$I_0 = \frac{c}{2}\epsilon_0|E_0|^2 \tag{2.3}$$

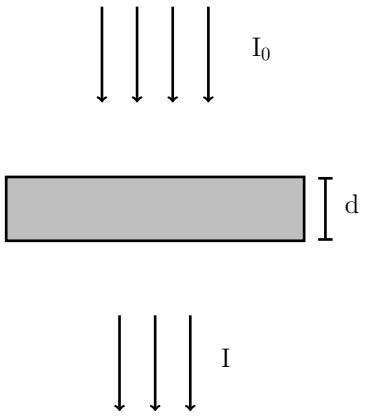


FIGURE 2.3: The ideal case of IR absorption spectroscopy, the sample can be considered as a thin, scatter free film of thickness d.

where c is the speed of light and ϵ_0 is the vacuum permittivity. In an ideal case, without backscattering from the front or back surface of the sample film, the attenuation of the incoming radiation is only caused by absorption. By this assumption, the electric field E inside the sample can be written as

$$E(x) = E_0 e^{i\tilde{k}x} \tag{2.4}$$

where x is the depth into the sample and $\tilde{k}$ is the complex angular wavenumber. The complex angular wavenumber $\tilde{k}$ can be written as

$$\tilde{k} = k + i \cdot \kappa = k_0 \cdot m \tag{2.5}$$

where k and κ are the real and imaginary parts of $\tilde{k}$, respectively, and k_0 is the angular wavenumber in vacuum and m is the complex refractive index. When the electromagnetic wave arrives at the detector, it has propagated through the sample over a distance of $x = d$ and the intensity measured at the detector can be written as

$$I = \frac{c}{2}\epsilon_0 |E|^2 = \frac{c}{2}\epsilon_0 |E_0|^2 e^{-2\kappa d} \tag{2.6}$$

According to Eq. 2.3 and 2.6, the transmittance can now be written as

$$T = e^{-2\kappa d} = e^{-4\pi n' d\tilde{\nu}} \tag{2.7}$$

by inserting $\kappa = k_0 \cdot n'$, where n' is the imaginary part of the refractive index, and $k_0 = 2\pi\tilde{\nu}$. The absorbance can now be expressed by

$$Z_{pure} = -\log_{10} T = \frac{4\pi n' d\tilde{\nu}}{\ln(10)} \tag{2.8}$$

The absorbance is here denoted Z_{pure}, as is derived from a situation without any scattering. Thus, Eq. 2.8 defines the pure absorbance, i.e. absorbance corresponding to absorption only. As the sample thickness d may vary, a scaling effect can be seen between different samples.

From Eq. 2.8 we can derive Beer-Lambert's law

$$Z(\tilde{\nu}) = a(\tilde{\nu}) \cdot d \cdot c \tag{2.9}$$

where a is the characteristic absorptivity corresponding to a chemical vibrational bond, d the sample thickness and c the concentration of the compound corresponding to this chemical bond. Hence, $\frac{4\pi n' \tilde{\nu}}{\ln(10)}$ can be interpreted as the absorptivity of the sample. The Beer-Lambert's

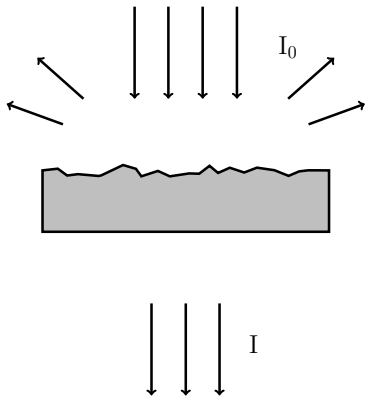


FIGURE 2.4: Scattering at a rough sample surface causes loss of incident radiation.

law is the starting point for setting up the multiplicative signal correction model later, and therefore an important connection point between the multivariate modelling of the scattering and absorption and the electromagnetic theory. This connection is hereby established.

Only in the ideal case, a sample in an infrared spectroscopic experiment can be considered a thin and scatter free film. For all practical situations this is only an approximation, and in most cases this approximation is not applicable. A considerable part of physical effects are involved and contribute to other features than absorption features to the spectrum. In practice, the surface of the sample is rough, resulting in diffuse scattering of the incident radiation. This situation is illustrated in Fig. 2.4. As a result of diffuse scattering, a portion of the background intensity I_0 does not penetrate into the sample. Therefore, when diffuse scattering is present, the attenuation of the electromagnetic radiation is caused by a combination of diffuse scattering and absorption. Assuming that the portion of electromagnetic radiation that is lost by diffuse scattering is constant with respect to the wavenumber [25], the absorbance can be expressed by

$$\begin{aligned} Z_{app}(\tilde{\nu}) &= -\log \frac{I(\tilde{\nu})}{\alpha I_0(\tilde{\nu})} \\ &= -\log \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} + \log \alpha \end{aligned} \tag{2.10}$$

where α is the portion of the incident radiation which is not scattered by the sample. By substituting the first term in the second line of Eq. 2.10 with the pure absorbance given in Eq. 2.8, we obtain the relation between the pure and apparent absorbance spectra

$$Z_{app}(\tilde{\nu}) = \frac{4\pi n' d\tilde{\nu}}{\ln(10)} + \log \alpha \tag{2.11}$$

where we recognize the first term on the right hand side as the pure absorbance. From this equation, it is evident that scattering of the incident wave may introduce a constant baseline shift, given by $\log \alpha$, with respect to the pure absorbance spectrum. Baseline shifts also occurs if the intensity of the source varies between measurements.

In this section, we have defined the pure absorbance spectrum and seen how spectra may deviate from this in an experimental situation. This is due to physical parameters affecting the experiment, such as diffuse scattering and the thickness of the sample. The diffuse scattering leads to constant baseline shifts $\log \alpha$, and the variability in sample thickness b to a multiplicative effect. We will in the following show how this variability can be greatly reduced by a

chemometrics method called multiplicative signal correction [32, 18].

2.1.2 Multiplicative signal correction and extended multiplicative signal correction

In order to analyze chemical differences between the absorbances in a set of measured spectra, it is desirable to pre-process the data in order to obtain spectra that are approximately pure absorbance spectra. When spectra are pre-processed and normalization of the spectra is achieved, peak heights are directly comparable. In addition, pre-processing leads to multivariate calibration and classification models with lower complexity [33]. A constant baseline and multiplicative effects can be removed by the model-based pre-processing method called multiplicative signal correction (MSC), proposed by Martens et al. in 1983 [32]. In this method, chemical and physical contributions to the apparent absorbance spectrum is separated in accordance with the electromagnetic model described in the previous section. In MSC, the apparent absorbance spectra are normalized to a given baseline and standard effective optical path length through a reference spectrum Z_{ref}

$$Z_{ref}(\tilde{\nu}) = \frac{4\pi d_s n' \tilde{\nu}}{\ln(10)} \tag{2.12}$$

where d_s is the standard thickness. According to Eq. 2.11, the apparent absorbance spectra can be described by

$$Z_{app}(\tilde{\nu}) = c + bZ_{ref}(\tilde{\nu}) + \epsilon(\tilde{\nu}) \tag{2.13}$$

where ϵ is the residuals, and

$$c = \log \alpha \quad \text{and} \quad d = b \cdot d_s \tag{2.14}$$

where c and b refer to the constant baseline shift and the scaling parameter respectively. The parameters ϵ, c and b are usually determined by least squares regression. In Eq. 2.13, the residuals express the unmodelled part of the apparent absorbance spectrum, and contain e.g. noise and chemical differences between the pure absorbance spectrum and the reference spectrum.

An example which illustrated the MSC model is shown in Fig. 2.5. Measured apparent absorbance spectra obtained from samples of Listeria monocytogenesis are shown in Fig. 2.5 a), where the scaling effect is evident. The spectra are obtained from high-throughput FTIR measurements of thin dried films of bacterial cells. Details of the experimental procedure and the measurement setup can be find in [42]. By using a mean of all spectra as a reference spectrum Z_{ref}, the scaling effect is reduced by MSC, as shown in Fig. 2.5 b). The reference spectrum is shown in black, and the corrected spectra are shown in grey.

The MSC method is based on a physical model, derived from assumptions and *a priori* knowledge, while the parameters are estimated from the data. In this way, MSC can be seen as an intermediate between the so-called "hard modelling" and "soft modelling", where the former is based on causality, assumptions and knowledge about the system, and the latter are data-driven, mathematical models [35, 33].

The fact that all apparent absorbance spectra are modelled around a reference spectrum offers a great advantage with respect to stability. As the difference between the reference spectrum and the true pure absorbance spectrum is relatively small, the main share of the pure absorbance is described through the reference spectrum. There will be no competition between the reference spectrum an the remaining terms in the MSC model in the parameter estimation, and therefore the reference spectrum offers a stable basis for modelling around. The parameter b indicates whether there is any relevant chemical information in the spectrum or not.

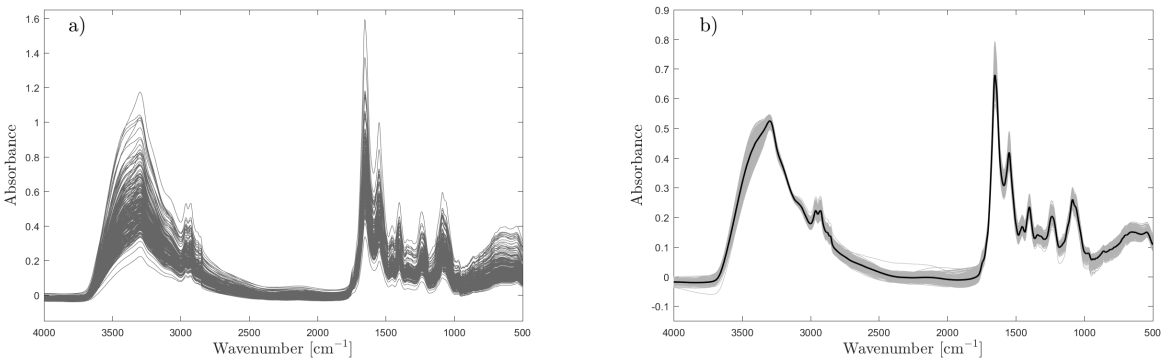


FIGURE 2.5: Multiplicative signal correction of a) measured absorbance spectra from Listeria monocytogenesis samples. b) The corrected spectra (grey) adapt to the baseline and scaling of the reference spectrum (black). The mean of all spectra are used as reference spectrum.

The electromagnetic model described in section 2.1.1 is a coarse approximation to most practical problems. As demonstrated before, the measured absorbance spectrum is a result of morphological and optical properties of the sample. In many cases, the share of the incident radiation that is lost by scattering is not constant with respect to the wavenumber. Therefore, non-constant baseline variations are frequent.

An example of how optical properties of a sample can introduce non-constant baseline variations is shown in Fig. 2.6 a). Figure 2.6 a) shows FTIR absorbance spectra obtained from samples of phenantherene in alkali halide (KBr) pellets [60]. The samples have been exposed to heat treatment at different degrees, a process which effects the optical properties of the whole pellet, such as particle size and packing. In order to analyze the chemical effect of heat treating the samples, a pre-processing of the spectra is necessary before comparison.

To handle situations where the scattering is wavenumber dependent, but still can be described by polynomials, extended multiplicative signal correction (EMSC) was proposed in 1991 by Martens et al. [35, 34]. In basic EMSC, a linear and quadratic wavenumber dependent term is added to Eq. 2.13 resulting in the following model

$$Z_{app}(\tilde{\nu}) = c + bZ_{ref}(\tilde{\nu}) + d\tilde{\nu} + e\tilde{\nu}^2 + \epsilon(\tilde{\nu}) \tag{2.15}$$

where c refers to the constant baseline shift and b to the multiplicative effect. Parameters d and e refers to a linear and quadratic baseline shift, respectively. EMSC allows to add further higher order terms, as well as chemical constituent spectra [1, 53, 57]. Sharper chemical features allows higher polynomial terms to be added without risk of modelling and removing chemical features. An example of spectra with sharp peaks where higher order polynomials usually are added in the EMSC model is Raman spectra [1].

EMSC has been favoured for its flexibility to handle a range of different problems. As an example, EMSC has been used to successfully suppress water vapour in absorbance spectra [9] and water variations in ATR spectra [36]. It can be used to remove variations due to paraffin in paraffin-embedded tissue samples [53, 57] and to correct shifts in Raman spectra [30, 48].

Figure 2.6 illustrates the difference between MSC and EMSC. In Fig. 2.6 c), the spectra from the KBr-pellet samples have been corrected with MSC, with the spectrum shown in Fig. 2.6 b) as reference spectrum. It is evident that the corrected spectra still contain scattering features. By employing a basic EMSC model as in Eq. 2.15, these baseline variations are corrected, as illustrated in Fig. 2.6 d). Note that if the corrected spectra would be used for determining relative concentration of chemical compounds by comparing peak heights, the reference spectrum in Fig. 2.6 b) should be baseline corrected prior to the EMSC correction. Alternatively, by mean

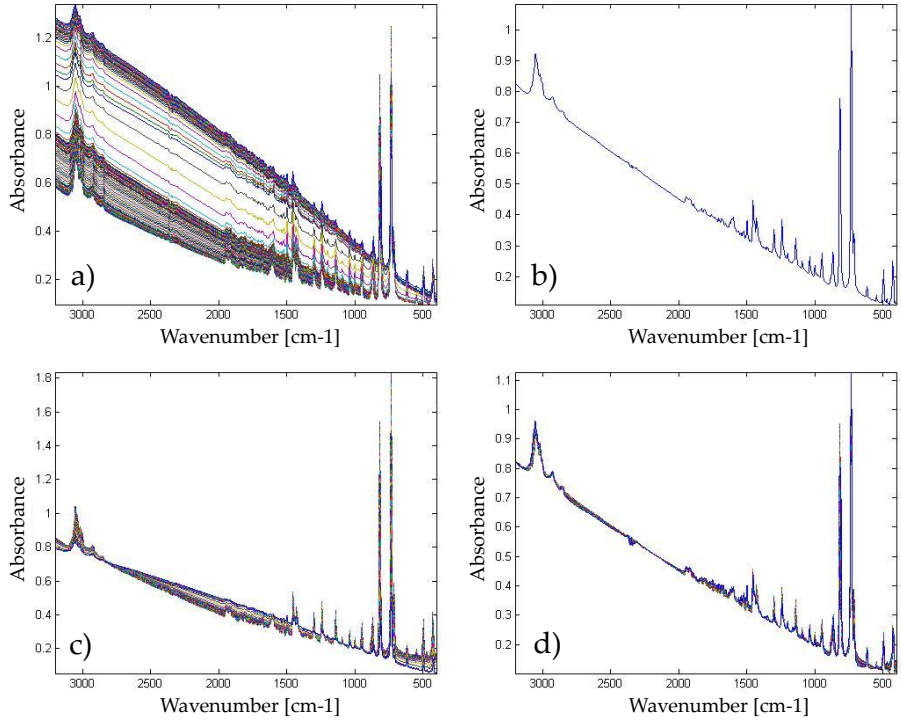


FIGURE 2.6: FTIR absorbance spectra obtained from samples of phenantherene in alkali halide (KBr) pellets, where a) shows the measured raw spectra. b) Reference spectrum used for c) correction with MSC, and d) correction with basic EMSC. By courtesy of Boris Zimmermann, Department of Science and Technology, NMBU.

centering the spectra, chemical differences can be explored without correcting the baseline of the reference spectrum [1].

One of the great advantages of model based pre-processing such as MSC and EMSC, is that valuable physical information is not lost in when spectra are corrected. MSC and EMSC are the only methods that takes into account that the appearance of the baseline is directly linked to the optical properties of the sample, and where the scatter information is parameterized and accessible for interpretation. It is well known that physical properties can give valuable information in biological and chemical applications. The parameters from Eq. 2.13 and 2.15 which are retrieved from the pre-processing can be used to gain additional information about the sample, related to sample size [21], density of the sample [25], temperature [25], denaturation [25] or phase transitions [60]. In the example of the absorbance spectra from heat treated phenantherene samples in Fig. 2.6 a), parameters from the EMSC correction were used to determine the phase transition temperatures [60]. In a study of allergenic pollen, it was shown that using the EMSC parameters as a part of the classification model resulted in an overall better classification [62].

Further, the residuals play a key role in the MSC and EMSC-modelling as they carry valuable chemical information. With both MSC and EMSC, an underlying assumption is that for all spectra in the data set, the pure absorbance spectra share the same basic features. This is the logic behind modelling around a reference spectrum. Individual deviations from the reference spectrum is contained in the residuals, and so the residuals can be used to express chemical differences within the data set. A better estimation of the pure absorbance than Z_{ref} can therefore be obtained by updating the reference spectrum with the residuals, a technique

which will be explained in detail later. In situations where the scattering and absorption can not be treated as independent, a good estimation of the pure absorbance spectrum is crucial.

As with all types of model based pre-processing methods, some knowledge is required about the sample and measurement technique *a priori*. EMSC should not be used for pre-processing absorbance spectra where the underlying assumptions of the electromagnetic model is not applicable.

In this section we described how multiplicative effects and constant baseline shifts in infrared absorbance spectra can be handled with the MSC model. Polynomial terms can be added to the MSC model, in order to handle situations where no analytic solution to the electromagnetic model can be used to describe the scattering. The model is then called EMSC. In some situations, we can expect the scattering to take on specific forms. For samples shaped like approximately perfect films or spheres, the absorbance spectra will be affected by interference fringes and Mie scattering respectively. These situations are two of the few cases where scattering in spectroscopy can be described by analytical solutions of electromagnetic models.

2.1.3 The complex index of refraction and the Kramers-Kronig relation

If the sample is shaped like a perfect film, back scattering from the surface on both sides of the film may cause baseline distortions due to interference. An example where this phenomenon is observed, is in live cell measurements of algae in water films [19]. The probability for the transmission and backscattering is defined by the reflection and transmission coefficients. A wave propagating through the film may be backscattered multiple times by the surfaces, and thus the electric field in the sample can not be described by Eq. 2.4. In short, in the formula for the transmittance, given by Eq. 2.7, the real part of the refractive index is not cancelled out, and the transmittance depends on both the real and imaginary part of the refractive index. In general, this case leads to interference effects, which occur due to differences in path lengths of the transmitted and internally reflected waves. These interference effects cause sine and cosine fluctuations in the transmittance and absorbance, which are also called interference fringes. In the literature, several methods for correcting interference fringes in infrared spectroscopy are described [11, 17]. Analytic solutions to scattering at perfect films has been incorporated in an EMSC model in order to correct fringes in absorbance spectra [27].

Similar but more complex scattering phenomena occur when the scatterer has a spherical shape. The scattering phenomena occurring at spherical shaped scatterers are in general strongly dependent on the refractive index. In general, a complex index of refraction is needed to describe the wave propagation dynamics and absorption. In the following, a brief overview of the complex refractive index is given.

The complex index of refraction is one of the fundamental material parameters related to electromagnetic wave propagation in matter. In general, it is frequency dependent and can be expressed as

$$m(\tilde{\nu}) = n(\tilde{\nu}) + i \cdot n'(\tilde{\nu}) \tag{2.16}$$

where n and n' is the real and imaginary part, respectively. The imaginary part of the refractive index is related to dissipation of energy due to absorption, while the real part is related to wave propagation dynamics. When electromagnetic waves propagate through matter, their speed and wavelength are decreased. Their speed and wavelength is inversely proportional to the real part of the refractive index.

Let c be the speed of the wave in vacuum, and v the phase velocity of the wave when it propagates in the medium. The real part of the refractive index gives the relation between c and v

$$n(\tilde{\nu}) = \frac{c}{v(\tilde{\nu})} \tag{2.17}$$

Since n is a measure for how much the plane electromagnetic wave is bent when entering the material, n was named the "refractive index".

It follows from the principle of causality that there must be a relationship between n and n'. This causality relation can be demonstrated by a simple thought experiment. Consider a filter media absorbing all radiation at the frequency $\tilde{\nu}_0$, probed with an electromagnetic pulse at the time t = 0. The pulse consists of different frequencies, say a Gaussian distribution around $\tilde{\nu}_0$. Accordingly, the pulse has some extension around $t = 0$ in the time domain. If there is no relation between n and n', the result of removing $\tilde{\nu}_0$ from the pulse in the frequency domain, is a pulse which extends into $t < 0$ as well as $t > 0$. It appears that the absorption of $\tilde{\nu}_0$ results in an effect prior to its cause, which clearly breaks with causality. To prevent the effect to occur before the pulse hits the filter, a phase shift must be introduced at frequencies around $\tilde{\nu}_0$, which cancels out the effect at $t < 0$.

Accordingly, it is evident that the principle of causality has an effect on the optical properties of matter. We take a closer look at the electric susceptibility χ, which relates to the refractive index by

$$\chi(\tilde{\nu}) = m(\tilde{\nu})^2 - 1 \tag{2.18}$$

The electric susceptibility is a fundamental property of dielectric matter, which describes the degree of polarization in response to an applied electric field. In order to avoid that the effect of absorption in the frequency domain is prior to its cause in the time domain, a phase lag needs to be introduced in the frequency-dependent susceptibility function. Accordingly, the susceptibility and the refractive index m are complex functions. The complex susceptibility χ can be written as

$$\chi(\tilde{\nu}) = \chi'(\tilde{\nu}) + i \cdot \chi''(\tilde{\nu}) \tag{2.19}$$

where χ' and χ'' is the real and imaginary part of χ, respectively. As will be illustrated in the following, the principle of causality implies that χ' and χ'' are dependent on each other, leading to the fact that dispersion and absorption always appear together.

In the presence of a moderate external electric field E, the polarization P of dielectric matter at time t_0 is described by the linear impulse response function $\chi(t_0 - t)$ through

$$P(x, t_0) = \int_{-\infty}^{t_0} \chi(t_0 - t) E(x, t) \mathrm{d}t \tag{2.20}$$

which is known as a convolution integral [58, 52]. Response functions, such as $\chi(t_0 - t)$, describe how a dynamic system responds to an external impulse. The applied pulse E is the input to the system, while P describes the output. In this thesis, we will not go into details of impulse response functions. However Eq. 2.20 motivates why we should look closer at the susceptibility function. The convolution integral given in Eq. 2.20 states that the system response to a complex pulse is a superposition of the responses to all constituent pulses. This motivates why we should look closer at the associated Fourier transform pairs of the electric susceptibility which are given by

$$\hat{\chi}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \chi(\omega) e^{-i\omega t} \mathrm{d}\omega \tag{2.21}$$

$$\chi(\omega) = \int_{-\infty}^{\infty} \hat{\chi}(t) e^{i\omega t} \mathrm{d}t \tag{2.22}$$

When requiring that the time-dependent susceptibility function is a real and causal function, it can be shown that the real part of $\chi(\tilde{\nu})$ must be an even function, while the imaginary part must be an odd function. This requirement can be written as

$$\begin{aligned} \chi'(\tilde{\nu}) &= \chi'(-\tilde{\nu}) \\ \chi''(\tilde{\nu}) &= -\chi''(\tilde{\nu}) \end{aligned} \tag{2.23}$$

Based on Eq. 2.22 and 2.23, it can be further shown that the frequency-dependent susceptibility can be expressed as

$$\chi(\tilde{\nu}) = \frac{i}{\pi} P \int_{-\infty}^{\infty} \frac{\chi(s)}{\tilde{\nu} - s} \mathrm{d}s \tag{2.24}$$

where P is the Cauchy principal value. From the requirements posed in Eq. 2.23, the real and imaginary parts of χ can now be expressed as

$$\chi'(\tilde{\nu}) = -\frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\chi''(s)}{\tilde{\nu} - s} \mathrm{d}s \tag{2.25}$$

$$\chi''(\tilde{\nu}) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\chi'(s)}{\tilde{\nu} - s} \mathrm{d}s \tag{2.26}$$

which is known as the Kramers-Kronig relations. From these relations it is evident that the dispersive and absorptive properties of the medium are in fact not independent.

For the real part of the refractive index, which is an even function, the Kramers-Kronig relations can be written as

$$n(\tilde{\nu}) = n_0 + \frac{2}{\pi} P \int_{0}^{\infty} \frac{s n'(\tilde{\nu})}{s^2 - \tilde{\nu}^2} \mathrm{d}s \tag{2.27}$$

where n_0 is the constant part of n, and the second term on the right hand side accounts for the dispersive part, i.e. the frequency dependent part. Likewise, n' can be expressed in terms of n in a similar expression.

The Kramers-Kronig relation can be illustrated by the Lorentz-model, a semi-classical model for describing absorption in dielectric media. It derives from considering the medium as consisting of diatomic molecules, and solving the associated equation of motion. The Lorentz-model allows the calculation of the complex refractive index for absorptive dielectric media, while it simultaneously fulfills the Kramers-Kronig relation. Figure 2.7 illustrates the relation between n and n', and shows that the real part of the refractive index fluctuates when passing through an absorption resonance. In this example n_0 is set to 0 for simplicity. At the right side of the absorption resonance, the real part of the refractive index approaches n asymptotically. Absorption resonances outside the IR region of the electromagnetic spectrum give long range contributions to the real part of the refractive index in the infrared region. These long-range contributions cause its constant background.

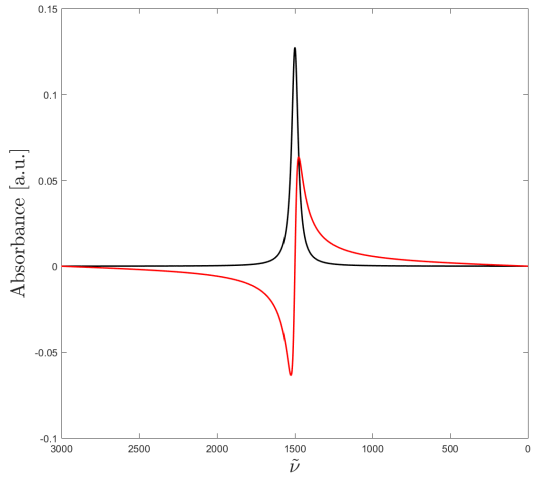


FIGURE 2.7: The Lorentz model: illustration of the relation between the fluctuating part of the real part of the refractive index, shown in red, and the imaginary part of the refractive index, shown in black. At absorption resonances, the real part of the refractive index fluctuates.

2.1.4 Approximately spherical scatterers or scatterers with surfaces that change their morphology on micrometer scale

Single cells and tissues have been observed to be effective scatterers in the infrared region of the electromagnetic spectrum. This type of scattering has been interpreted as Mie type scattering [38], i.e. scattering by dielectric spheres. Mie scattering in FTIR spectroscopy introduces gross baseline variations in the absorbance spectra, as seen in figure 2.8, and a more advanced version of EMSC is therefore needed for pre-processing these spectra. A model-based pre-processing method to handle this case is obtained by establishing an EMSC meta model. This is achieved by incorporating a data model based on Mie theory into an EMSC model, as will be explained in section2.1.6. Before presenting a brief overview of Mie theory, it is expedient to define the concept of scattering.

Scattering of an electromagnetic wave is defined as deviations from rectilinear wave propagation, which occur due to interaction with an object. When the wavelength of the incident radiation is small compared to the changes of the surface of the object, the electromagnetic wave can be approximated by rays, and the phenomenon of scattering can be described by reflection and refraction. However, when the wavelength of the incident radiation is comparable to the scale at which the surface of the object changes, the principles of geometrical optics fail in describing the resulting electromagnetic field. This type of scattering is related to interference phenomena, diffraction and resonances. To solve scatter problems involving diffraction, such as scattering at spheres of sizes comparable to the wavelength of the incident radiation, solutions to Maxwell's equations are required.

The physical process behind diffraction is the same as for reflection and refraction: The varying electric field in the incident radiation accelerates charged particles in the object, causing them to oscillate. When charged particles oscillates, they send out electromagnetic radiation in all directions. In this way, the object can be thought of as consisting of numerous point sources of radiation. In the case of a nonabsorbing object, scattering removes a portion of the energy in the incident radiation in forward direction, while the rest of the radiation is transmitted. The field around the scatterer can be viewed as a superposition of the transmitted radiation, and the radiation emitted from each point source. For absorbing materials, the

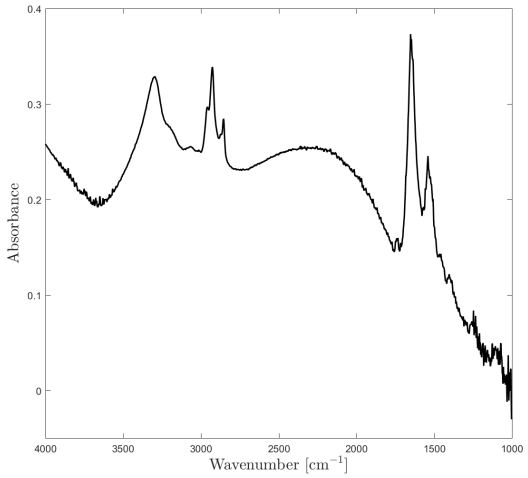


FIGURE 2.8: Apparent absorbance spectrum obtained from a single lung cancer cell [24]. Mie scattering introduces gross baseline variations in the measured spectrum.

scattering and absorption is mutually dependent on each other. As mentioned in the previous section, absorption resonances leads to fluctuations in the real part of the refractive index, often referred to as the dispersion effect. Therefore, absorption affects the wave propagation dynamics. Further, the morphology and optical properties of the scatterer affects the internal electric field. At wavenumbers corresponding to so-called shape resonances, the internal electic field is enhanced, leading to an increased absorption.

Scattering of a plane wave by a homogeneous sphere is one of the situations where an analytical solution to Maxwell's equations exist. The solution was first described by the Danish physicist Ludvig Lorenz in 1890. However, the theory went unrecognized until eighteen years later when the German physicist Gustav Mie published what is now known as "Mie theory" [58]. In the following section a brief overview of Mie theory will be given, followed by Mie scattering in absorbance spectra.

2.1.5 Mie theory

The Mie theory provides an exact analytical solution for the scattering of a plane wave at a homogeneous sphere. The exact Mie solutions are as intriguing as they are intricate. Therefore, for numerical computations, approximate formulas such as the Van De Hulst formula are often used. In this section the exact solutions will be presented, followed by a presentation of the van de Hulst approximation.

Efficiency factors: In Fig. 2.9, we consider a plane wave propagating in the x-direction. The plane wave is impinging on an object and propagating through the object. The transmitted radiation is measured by the detector at the bottom in Fig. 2.9.

As illustrated in Fig. 2.9, part of the incident radiation is extinguished by the object. Hereby we mean that only a part of the incoming plane wave is transmitting through the object in forward direction and reaching the detector. A substantial part of the plane wave is extinguished by either scattering or absorption. How effective the object is in terms of extinction is described by the extinction efficiency factor. The extinction efficiency Q_{ext} is given as

$$Q_{ext} = Q_{abs} + Q_{sca} \tag{2.28}$$

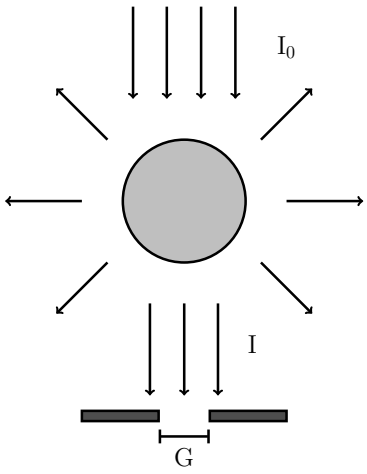


FIGURE 2.9: Illustration of Mie scattering in IR spectroscopy. A portion of the incident radiation is scattered by the spherical sample, and less radiation makes its way to the detector. The geometrical cross section of the detector is denoted G.

where Q_{abs} and Q_{sca} is the efficiency factors corresponding to absorption and scattering, respectively. The efficiency factors are dimensionless constants describing which part of the incident radiation is absorbed and scattered, i.e. extinguished in total. The extinction efficiency relates to the extinction cross section C_{ext} by

$$Q_{ext} = \frac{C_{ext}}{g} \tag{2.29}$$

where g is the geometrical cross section of the interfering object. For a sphere of radius a, the geometrical cross section is $g = \pi a^2$. The extinction cross section is the part of incident radiation flux area, which is lacking from the total incident radiation flux after interaction with the scatterer.

When solving Maxwell's equations for a plane wave that is incident on a spherical homogeneous scatterer, the extinction efficiency can be written as

$$Q_{ext}(\tilde{\nu}) = \frac{2}{x(\tilde{\nu})^2} \sum_{n=1}^{\infty} (2n+1)\,\mathrm{Re}\{a_n + b_n\} \tag{2.30}$$

where a_n and b_n are scattering coefficients corresponding to resonant electric and magnetic modes, respectively, given by

$$\begin{aligned} a_n &= \frac{\psi_n'(y)\psi_n(x) - m\psi_n(y)\psi_n'(x)}{\psi_n'(y)\zeta_n(x) - m\psi_n(y)\zeta_n'(x)} \\ b_n &= \frac{m\psi_n'(y)\psi_n(x) - \psi_n(y)\psi_n'(x)}{m\psi_n'(y)\zeta_n(x) - \psi_n(y)\zeta_n'(x)} \end{aligned} \tag{2.31}$$

where ψ_n is the incoming wave function, ζ_n is the scattered wave function [20] and $y = m \cdot x$. The scattering coefficients are dependent on the complex refractive index m. The size factor x is the ratio of the circumference of the sphere to the wavelength of the incident radiation given as,

$$x(\tilde{\nu}) = \frac{2\pi a}{\lambda} = 2\pi a \tilde{\nu} \tag{2.32}$$

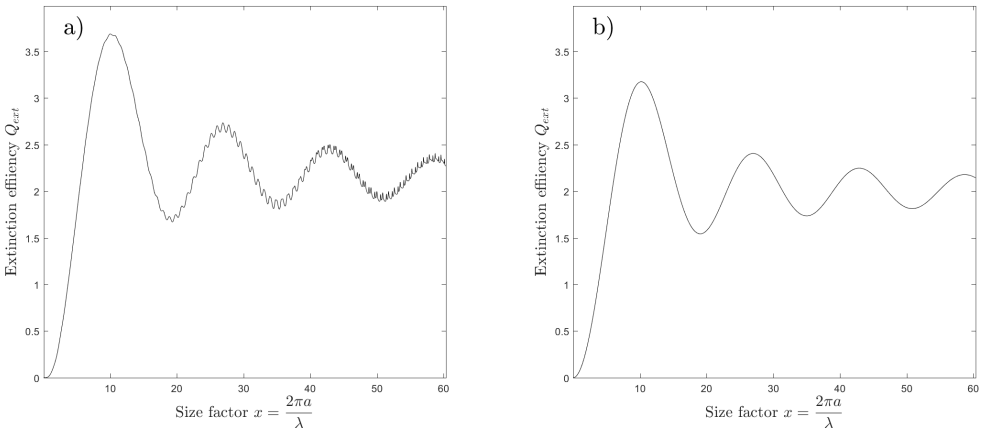


FIGURE 2.10: Extinction efficiency by spherical scatterer with refractive index of $m = 1.2$ given by a) exact Mie theory, and b) the van de Hulst approximation formula. The approximation formula expresses only the broad wiggles, and not the sharp ripples.

where a is the radius of the sphere and λ is the wavelength. Figure 2.10 a) shows how the extinction efficiency varies with the size factor, for a nonabsorbing sphere with refractive index $m = 1.2$. The broad oscillations are called wiggles, while the sharper ones are called ripples. The wiggles are caused by the superposition of the undisturbed transmitted field and the scatter field emitted from each point source. The ripples occur due to shape resonances, which occur at specific wavenumbers as mentioned above. For spherical geometries, shape resonances are often referred to as whispering gallery modes [8, 43].

As Eq. 2.30 is rather computationally expensive, and since it contains features such as ripples, which are commonly not observed in IR microscopy spectra of cells, it has not in practice been used for pre-processing of IR spectra spectra of single cells and tissues. One of the advantages of infrared spectroscopy and imaging is the ability to take measurements in a rapid manner. Currently, the data processing subsequent to the measurements is a bottle neck. While quantum cascade lasers can be employed to acquire large amount of imaging spectra from tissues, the modelling of the scattering is time-consuming by various reasons. A far less computationally expensive approximation formula derived by van de Hulst [20] is therefore used in place of Eq. 2.30 in pre-processing of IR spectra. The van de Hulst approxiamtion formula is given by

$$Q_{ext}(\tilde{\nu}) = 2 - 4e^{-\rho\tan\beta}\frac{\cos\beta}{\rho}\sin(\rho-\beta) - 4e^{-\rho\tan\beta}\Big(\frac{\cos\beta}{\rho}\Big)^2\cos(\rho-2\beta) + 4\Big(\frac{\cos\beta}{\rho}\Big)^2\cos(2\beta) \tag{2.33}$$

where

$$\rho = 4\pi a\tilde{\nu}(n-1) \quad \text{and} \quad \tan\beta = \frac{n'}{n-1} \tag{2.34}$$

where n and n' is the real and imaginary part of the refractive index, respectively, and a is the refractive index [20].

The approximation formula given by van de Hulst is plotted in Fig. 2.10 b) for a sphere with refractive index $m = 1.2$. Comparing with Fig. 2.10 a), we see that the Van de Hulst formula expresses only wiggles. As mentioned, ripples are usually not seen in IR spectra of biological samples, for reasons that are still under investigation. A possible explanation is the fact that biological samples are rarely to be viewed as perfectly spherical.

Equation 2.33 provides the total extinction in forward direction for an absorbing, homogeneous spherical scatterer. It takes into account both chemical absorption and scattering. By considering conservation of energy and the detector aperture size, it can be shown that the apparent absorbance spectrum relates to the extinction efficiency by

$$A_{app}(\tilde{\nu}) = -\log_{10}\left[1 - \frac{\pi a^2}{G\ln(10)} Q_{ext}(\tilde{\nu})\right] \tag{2.35}$$

where G is the geometrical cross section of the detector aperture. By expanding the logarithm and considering only the first term, Eq. 2.35 can be reduced to

$$A_{app}(\tilde{\nu}) = -\log_{10}\frac{\pi a^2}{G\ln(10)} Q_{ext}(\tilde{\nu}) \tag{2.36}$$

Eq. 2.36 is of importance, as we later will assume A_{app} to be proportional to Q_{ext}. In the following, we demonstrate how the absorption and scattering are mutually dependent on each other.

Mutual dependence of scattering and absorption

From Eq. 2.28 it may seem like absorption and scattering are additive effects, such that the absorption efficiency is proportional to the pure absorbance, and the scattering efficiency could simply be calculated from wave propagation dynamics. If this was the case, scattering components could be included in the EMSC model as additive terms, as done with the polynomials in the basic EMSC. However, the effects of scattering and absorption are mutually dependent on each other, and should therefore not be treated as additive. Both scattering and absorption depend on the optical properties and morphology of the scatterer, which together determine the internal electric field. This mutual dependence of scattering and absorbance is mirrored by the existence and interaction of shape resonances and absorption resonances: while the shape resonances are due to scattering and affect the absorption properties, the absorption resonance are due to chemical absorption and affect the scattering properties.

At wavenumbers corresponding to shape resonances, the internal electric field is enhanced due to a constructive interference wave pattern, leading to standing waves inside the scatterer. For spherical scatterers, this effect is commonly referred to as whispering gallery modes, after a similar effect observed in the dome of St. Paul's cathedral in London [8, 43]. It was observed that by whispering along the wall of the dome, the sound could be heard clearly along the wall, all around the gallery. This phenomenon occurs due to the sound waves travelling as standing waves along the concave wall of the dome, by total internal reflection. At specific wavelengths, constructive wave interference is achieved. Since these constructive interference patterns are resonance effects and depend on the shape of the object, we also refer to them as shape resonances. The same phenomena are observed in electromagnetic wave propagation in dielectric spheres, which is illustrated in Fig. 2.11 [8]. The figure illustrates a shape resonance in a dielectric, nonabsorbing sphere with refractive index $n_0 = 1.8$. A plane wave is incident on the sphere from the left, and the wavelength and size of the sphere corresponds to a size factor of 8.5. The enhanced electric field is seen as bright spots at the inner edge of the sphere. An enhanced electric field in an absorptive medium leads to enhanced absorption, and in the absorbance spectra, shape resonances appear as the ripples. As mentioned, ripples are usually not observed in IR absorbance spectra from biological samples. However, it is obvious that the internal electric field in general has an effect on the absorption efficiency.

How the real part of the refractive index depends on the absorptivity of the medium is already described in section 2.1.3. At absorption resonances, the real part of the refractive index fluctuates as illustrated by the Lorentz-model. It is evident that the wave propagation dynamics are affected by absorption. However from Eq. 2.30 and 2.33, it is not intuitive in which way.

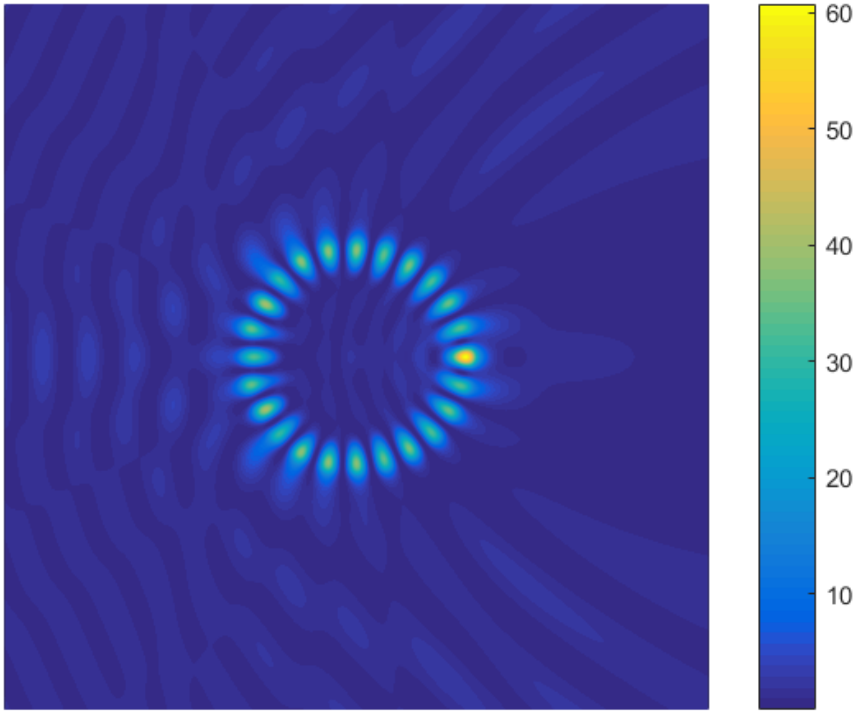


FIGURE 2.11: Illustration of whispering gallery modes in a dielectric sphere of refractive index 1.8. A plane wave is incident on the sphere from the left at a wavelength corresponding to a shape resonance, resulting in an enhanced internal electric field. The wavelength and size of the sphere corresponds to a size factor of 8.5. The color bar represents field strength in arbitrary units. By courtesy of Maren Anna Brandsrud, Department of Science and Technology, NMBU.

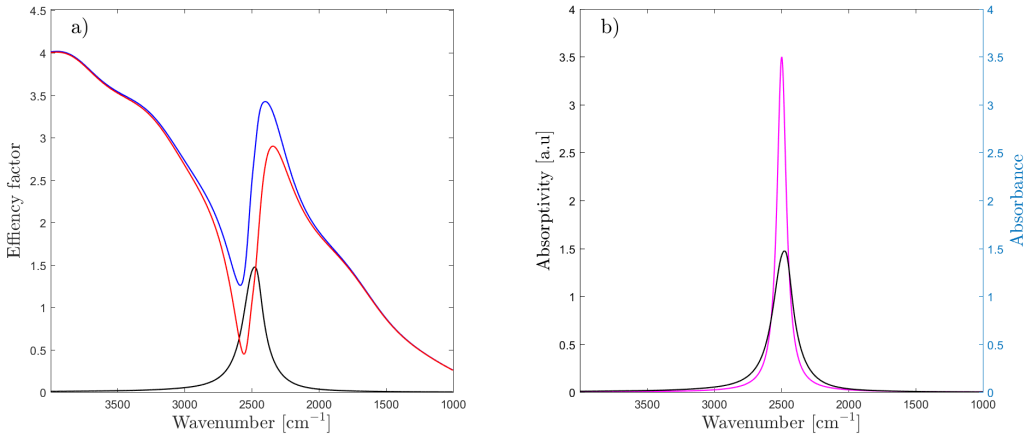


FIGURE 2.12: Illustration of the dispersion effect. a) Extinction efficiency in blue, by a sphere with constant refractive index of 1.4, and absorbance corresponding to a Lorentzian line with the peak position at 2,498 cm^{-1}. Scattering efficiency and absorption efficiency are plotted in red and black, respectively. b) The absorption efficiency in black, is not proportional to the absorbance shown in purple.

The effect can be illustrated by an example as shown in Fig. 2.12. In Fig. 2.12 a), the extinction efficiency of a simulated sphere of radius $12\,\mu\mathrm{m}$ is shown. The absorption and scattering efficiency is plotted in black and red, respectively. The extinction efficiency is plotted in blue. The imaginary part n' of the refractive index of the sphere was constructed by simulating a single absorption band, a Lorentzian line with the peak position at 2,498 cm^{-1}. The real part was then determined by setting the constant part n_0 to 1.4, and calculating the fluctuating part n_{kk} from n'. From Fig. 2.12 a), it is evident that the absorption resonance at 2,498 cm^{-1} introduces a peak in the extinction efficiency, at 2,400 cm^{-1}. We observe, that the peak at 2,498 cm^{-1} in the extinction efficiency is shifted towards lower wavenumbers. This occurs due to scattering of the incident radiation which is illustrated by the scattering efficiency in red. Scattering of the incident radiation also affects the absorption efficiency, which is illustrated in Fig. 2.12 b). It is evident that the absorption efficiency in black is not proportional to the absorptivity of the sphere, shown in purple. The absorption efficiency can thus not be considered proportional to the pure absorbance. The effect of absorption resonances on the extinction efficiency is sometimes referred to as the "dispersion effect", as it is a result of the wavenumber dependent, i.e. dispersive, refractive index.

The dispersion effect is observed in IR absorbance spectra of approximately spherical biological samples. In particular, the amide I absorbance band is affected by the dispersion effect, causing the peak position to shift towards lower wavenumbers. Without successfully correcting this effect by pre-processing methods, a shift in amide I can be erroneously interpreted as a change in the secondary structure of proteins. The amide I band in the absorbance spectrum is frequently used for protein structural analysis [4]. As absorption bands consist of multiple overlapping bands, both peak position and band shape are crucial for the chemical analysis. A shift in the amide I peak position towards lower wavenumbers can be interpreted as a a decrease in the concentration of α-helix and an increase in the concentration of β-sheet. In some cases, the concentration of α-helix and β-sheet is used to distinguish between healthy and deceased tissue [39].

2.1.6 Resonant Mie scatter EMSC

In order to separate attenuation caused by physical and chemical properties in IR spectra affected by Mie scattering, Kohler et al. proposed in 2008 to incorporate the van de Hulst approximation formula in a predicative EMSC model [24]. In this first model, the approximation formula for a non-absorbing sphere was employed, treating the scattering and absorption as independent. By setting the imaginary part and fluctuating real part of the refractive index to zero, Eq. 2.33 reduces to

$$Q_{ext}(\tilde{\nu}) \approx 2 - (\frac{4}{\rho}) \sin(\rho) + (\frac{4}{\rho^2})[1 - \cos(\rho)] \tag{2.37}$$

where

$$\rho = \alpha\tilde{\nu} \quad \text{and} \quad \alpha = 4\pi a(n_0 - 1) \tag{2.38}$$

It is evident from Eq. 2.37 that in order to predict the scattering by the sample, knowledge about the radius and the refractive index is required. However, in most practical cases, these parameters are unknown. Therefore, the parameters need to be estimated in the modelling process, and a model that allows the use of a broad range of radius and refractive index is needed. In order to achieve this, the Mie theory was implemented as a meta-model consisting of a compressed form of a set of extinction curves, that were calculated from a range of relevant refractive indices and radii. To summarize the extinction curves in an efficient way, the set of extinction curves were compressed by a principal component analysis (PCA) model into a few loadings, before they were included in the EMSC model. The set of loadings used in the EMSC model allowed in addition that the samples are not modelled only as perfectly homogeneous spheres, but also as a superposition of spheres of different radii and refractive indices. The obtained EMSC model represents thus a meta-model of the Mie scattering described by the Van de Hulst formalism, and is therefore based on electromagnetic theory. In this thesis we refer to this model as the Mie EMSC model.

In the following, a brief overview of PCA is given. While PCA is a commonly used tool for multivariate data analysis of spectroscopic data, it is also a central part of the Mie EMSC model. Later in this thesis, PCA will be used for investigating clustering of data by means of both chemical and physical properties. However, PCA as a tool for explorative data analysis is well described in literature, and therefore the following description will be related to how the PCA is used to construct the meta-model.
Let $\boldsymbol{Q_{ext}}$ be a $N \times K$-matrix which consists of N extinction curves, expressed by K variables. In spectroscopic data sets, the K variables corresponds to wavenumbers. Each extinction curve can be represented as a point in a K-dimensional space. For spectra represented by several thousand wavenumbers, this is impossible to illustrate graphically. We therefore illustrate the problem with an example data set described by three parameters. The example data set is the famous Iris-data set from Fisher, where 150 iris flowers are described by four parameters [55]. These parameters are petal and sepal length and width. The 150 iris flowers samples are plotted for the two parameters petal length and width, in Fig. 2.13.

In data sets with several variables, such as spectroscopic data, a high degree of co-variation between the variables is often present. In order to find these covariation patterns, we search for the direction in the variable space, which expresses the strongest variance. In other words we seek the direction at which samples shows the strongest spread when projected onto it. In Fig. 2.13, this direction is illustrated by the red arrow. This direction represents the first principal component and it is mathematically described by the loading vector. The first loading vector is represented by unit vector $\boldsymbol{p_1}$. The direction at which the second most variance is expressed,

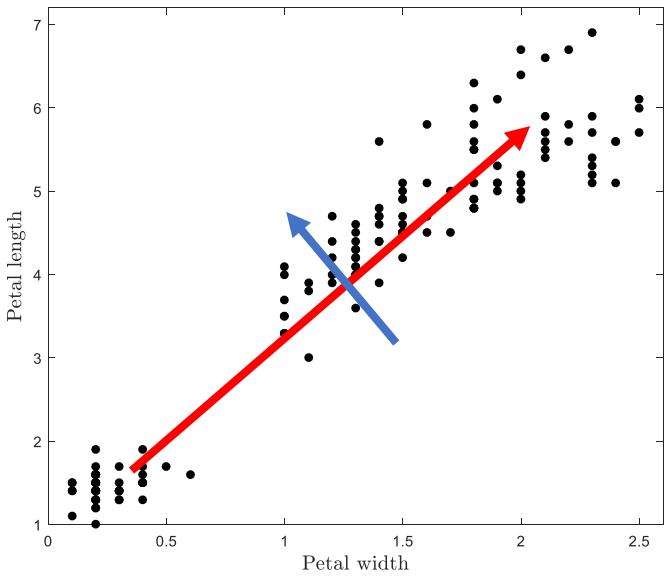


FIGURE 2.13: The Iris-data set by Fisher [55], represented by the two variables petal lengths and width. The red and blue arrow illustrates the first and second principal component, respectively.

the second principal component, is illustrated by the blue arrow in Fig. 2.13. The corresponding unit vector is denoted $\boldsymbol{p_2}$, and is orthogonal to $\boldsymbol{p_1}$. The third principal component (not shown in Fig. Fig. 2.13) represents the third most variance, and so on.

Mathematically, PCA uses an orthogonal linear transformation, which maps the data to a coordinate system of A latent variables. This set of latent variables expresses the main covariance pattern in the original data set. The PCA model can be written as

$$\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{T_A}\boldsymbol{P'_A} + \boldsymbol{E_A} \tag{2.39}$$

where $\bar{\boldsymbol{X}}$ is a matrix which contains the mean of all samples in each row, i.e. the sample mean is repeated in each row. The matrix $\boldsymbol{T_A}$ is the score matrix of size $N \times A$ and $\boldsymbol{P_A}$ is the loading matrix of size $K \times A$. The matrix $\boldsymbol{E_A}$ is the residual matrix, which describes the reconstruction error, i.e. the variance which is not expressed by $\boldsymbol{T_A}\boldsymbol{P'_A}$. The above mentioned loading vectors $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ constitute the first two columns of $\boldsymbol{P_A}$. The loading matrix $\boldsymbol{P_A}$ is organized such that the loading which describes the most variance is placed in the first column, and the loadings follow in decreasing order. The score matrix $\boldsymbol{T_A}$ is organized in the same way. When the data set $\boldsymbol{X}$ contains a high degree of covariation, only a few principal components $\boldsymbol{A}$ are sufficient to describe most of $\boldsymbol{X}$, and the residual $\boldsymbol{E_A}$ is small.

We return now to the matrix of extinction curves $\boldsymbol{Q_{ext}}$. It has been shown, that the matrix $\boldsymbol{Q_{ext}}$ can be represented by few principal components without loosing much information [24]. The reconstructed data set can be expressed as

$$\boldsymbol{Q_{ext}} = \bar{\boldsymbol{Q}}_{\boldsymbol{ext}} + \boldsymbol{T_A}\boldsymbol{P'_A} + \boldsymbol{E_A} \tag{2.40}$$

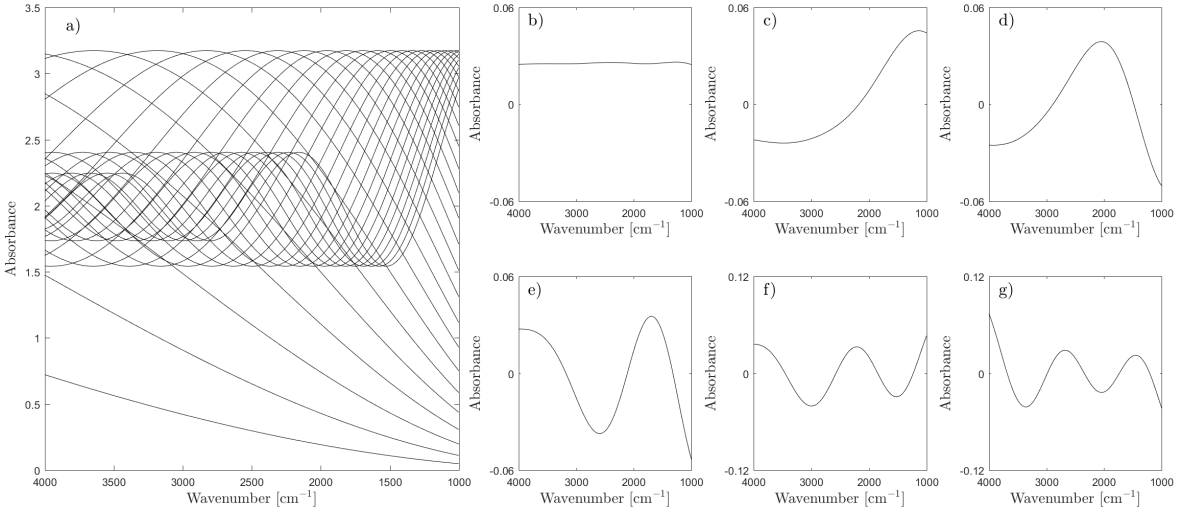


FIGURE 2.14: a) A set of extinction curves calculated by Eq. 2.37 with the parameter range of 30 equidistantly spaced values of $\alpha \in [2.8, 12.3] \times 10^{-6}$. b-g) The six first loadings from a PCA on the extinction curves shown in a).

The matrix of extinction efficiencies $\boldsymbol{Q_{ext}}$ can be represented by a few components p_i in an EMSC model according to

$$Z_{app}(\tilde{\nu}) = c + bZ_{ref}(\tilde{\nu}) + \sum_{i=1}^{A} g_i p_i(\tilde{\nu}) + \epsilon(\tilde{\nu}) \tag{2.41}$$

where Z_{ref} is the reference spectrum representing the pure absorbance spectrum. The parameter A is the number of principal components used in the Mie meta-model. As a starting point, we consider an extinction efficiency $\boldsymbol{Q_{ext}}$ obtained by the van de Hulst approximation for a nonabsorbing sphere. In addition, we treat the absorption and scattering as independent. We further use the approximation that the extinction efficiency Q_{ext} is proportional to Z_{app}.

Representing $\boldsymbol{Q_{ext}}$ by a few components p_i has several advantages. It allows to represent a range of radii and refractive indices without running into stability problems in the parameter estimation. In Fig. 2.14 a) a set of extinction curves calculated from Eq. 2.37 for a range of 30 equidistantly spaced values of $\alpha \in [2.8, 12.3] \times 10^{-6}$ is shown. The six first loadings from a PCA on the set of extinction curves in Fig. 2.14 a) are shown in b-g).

In ref. [24], it was shown that the broad Mie oscillations could be corrected with the Mie EMSC model of Eq. 2.41. However, the model revealed challenges related to the dispersive effect, as expected when treating scattering and absorption as independent. The Mie EMSC model was further developed in 2010 by Bassan et al. [5] to handle the effect of absorption resonances on the extinction efficiency. The extinction efficiency was still calculated from Eq. 2.37, i.e. considering a nonabsorbing sphere, except taking into account a fluctuating real part of the refractive index. An estimation of the imaginary part of the refractive index was obtained by assuming proportionality between n' and the reference spectrum, which represents the pure absorbance, as expressed in Eq. 2.8. By the Kramers-Kronig relation, an estimation of n_{kk} was obtained. The Mie meta-model was set by three parameters, i.e. the radius of the sphere, the constant real part of the refractive index, and the proportionality relation between the imaginary part of the refractive index and the reference spectrum. The model was named resonant Mie scattering EMSC, as the dispersion effect sometimes is referred to as the resonant case. It was shown that the dispersive dip close to the amide I, and the associated shift in the

amide I peak position, could be corrected with this approach [5]. In this thesis we in general refer to the resonant Mie scattering EMSC as Mie EMSC for simplicity. When it is relevant to make a difference between the resonant case and the non-resonant case, we state this explicitly. Further, the model was implemented in an iterative algorithm, of which the concept will be explained in section 3.1. Challenges were however observed with respect to the corrected spectra being strongly biased by the reference spectrum, causing the corrected spectra to adapt features from the reference spectrum.

The latest improvements of the Mie EMSC model were proposed by Konevskikh et al. in 2016 [28]. By the use of the relation described in Eq. 2.8, the imaginary part of the refractive index could be calculated from the reference spectrum according to

$$n'(\tilde{\nu}) = \frac{Z_{ref}(\tilde{\nu}) \ln 10}{4\pi d_{eff} \tilde{\nu}} \tag{2.42}$$

where the standard sample thickness d_s is replaced by the effective diameter d_{eff}, as the thickness of a sphere is not constant. The effective diameter could in principle be calculated directly from the radius of the sphere according to

$$d_{eff} = \frac{\pi}{2} a \tag{2.43}$$

However, as biological samples are not to be considered as perfect spheres, a scaling parameter h is introduced such that

$$d_{eff} = \frac{\pi}{2} \frac{a}{h} \tag{2.44}$$

Further, Konevskikh et al. [28] showed that only two independent model parameters were present and reduced the model parameters from three parameters to two parameters. The two-parameter model was obtained by expressing the imaginary and fluctuating real part of the refractive index as scaled parameters, n'_s and $n_{kk,s}$ respectively, according to

$$n'_s(\tilde{\nu}) = \frac{n'(\tilde{\nu})}{f} \quad \text{and} \quad n_{kk,s}(\tilde{\nu}) = \frac{n_{kk}(\tilde{\nu})}{f} \tag{2.45}$$

where the scaling factor f is given by

$$f = h \frac{\ln(10)}{2\pi^2 a} \tag{2.46}$$

Further it could be shown that ρ and $\tan\beta$ from Eq. 2.34 can be expressed by

$$\rho = \alpha_0 (1 + \gamma n_{kk,s}) \tilde{\nu} \quad \text{and} \quad \tan\beta = \frac{n'_s}{\frac{1}{\gamma} + n_{kk,s}} \tag{2.47}$$

where

$$\alpha_0 = 4\pi a (n_0 - 1) \quad \text{and} \quad \gamma = h \frac{2 \ln(10)}{\pi \alpha_0} \tag{2.48}$$

The ranges and parameter distribution of α_0 and γ determines how the Mie extinction curves are expressed through the loadings from the PCA. Details on this relationship will be assessed elsewhere. When using the formalism governing an absorptive sphere, the Mie extinction curves contain chemical features as well as scattering signatures. This is illustrated in Fig. 2.15 a) where 100 extinction curves calculated by Eq. 2.33 are plotted, for the parameter ranges $\alpha_0 \in [2.5, 35.7] \times 10^{-5}$ and $\gamma \in [1.2, 14.6] \times 10^4$. The scaled imaginary part of the refractive index in Eq. 2.45 was calculated by using the Matrigel spectrum as the reference spectrum in Eq. 2.42.

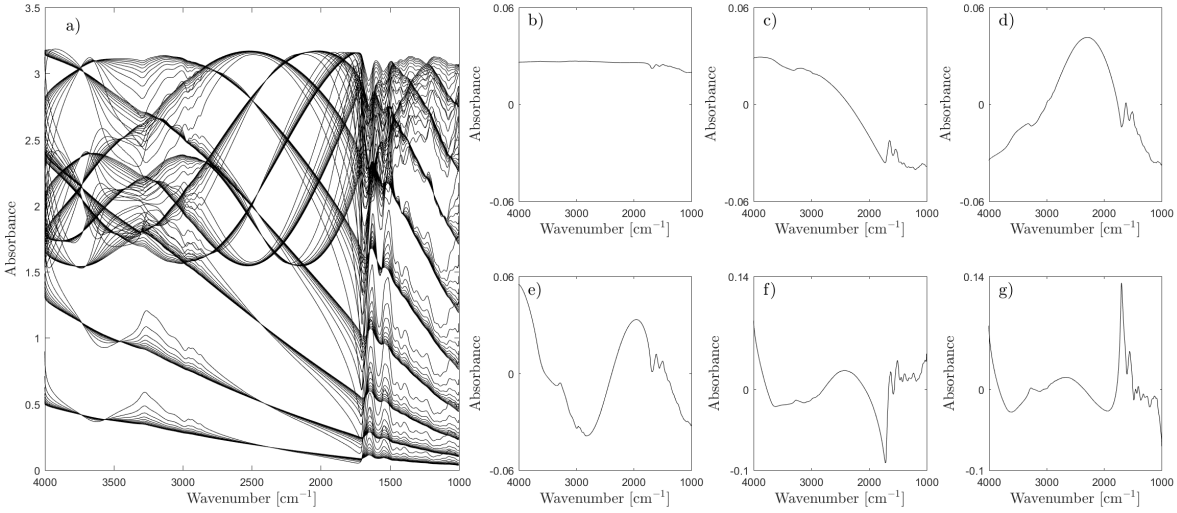


FIGURE 2.15: a) A set of extinction curves calculated by Eq. 2.33 with the parameter ranges $\alpha_0 \in [2.5, 35.7] \times 10^{-5}$ and $\gamma \in [1.2, 14.6] \times 10^4$. The Matrigel spectrum was used as input for the imaginary part of the refractive index. b-g) The six first loadings from a PCA on the extinction curves shown in a).

It is important to note that the algorithm by Konevskikh at al. [28] employs the van de Hulst formalism for an absorptive sphere. Thus, the extinction curves Q_{ext} express both scattering and absorption efficiency. By repeating the chemical absorbance features in Z_{ref} in addition to the extinction curves Q_{ext}, the different terms in the Mie EMSC model would compete in the parameter estimation. Therefore, the matrix containing the extinction curves $\boldsymbol{Q_{ext}}$ are made orthogonal to the reference spectrum Z_{ref} according to

$$\boldsymbol{\tilde{Q}_{ext}} = \boldsymbol{Q_{ext}} - \boldsymbol{Q_{ext}} \cdot \frac{\boldsymbol{z_{ref}} \cdot \boldsymbol{z'_{ref}}}{|\boldsymbol{z_{ref}}|^2} \tag{2.49}$$

where $\boldsymbol{\tilde{Q}_{ext}}$ is the matrix containing the orthogonal extinction curves and $\boldsymbol{z_{ref}}$ is the $K \times 1$ vector containing the reference spectrum Z_{ref}. The parts of $\boldsymbol{Q_{ext}}$ that can be expressed by Z_{ref} are now extracted from the set of extinction curves, and a stable modelling around the reference spectrum is guaranteed. The Mie EMSC model of 2016 is implemented in the fast iterative Mie scatter correction algorithm [28], which is described in section 3.1. In Fig. 2.16 a) the extinction curves which are orthogonal to Z_{ref} is shown. The Matrigel spectrum is used for calculating the imaginary part of the refractive index, and the ranges for the parameters in 2.48 are set to $\alpha_0 \in [2.5, 35.7] \times 10^{-5}$ and $\gamma \in [1.2, 14.6] \times 10^4$. The six first loadings from a PCA on the orthogonalized extinction curves are shown in Fig. 2.16 b-g).

Scientific models and the Mie meta-model

In the philosophy of sciences, the concept of models has lately attained an increasing degree of interest. The notion of models covers a wide range of concepts, but in reality, models can be categorized based on two fundamentally different functionalities. A model can be either a representational model, or a model of theory [44].

Representational models are models describing a part of the world, which in this sense is interpreted as an observable fact or event related to more or less stable features of the world.

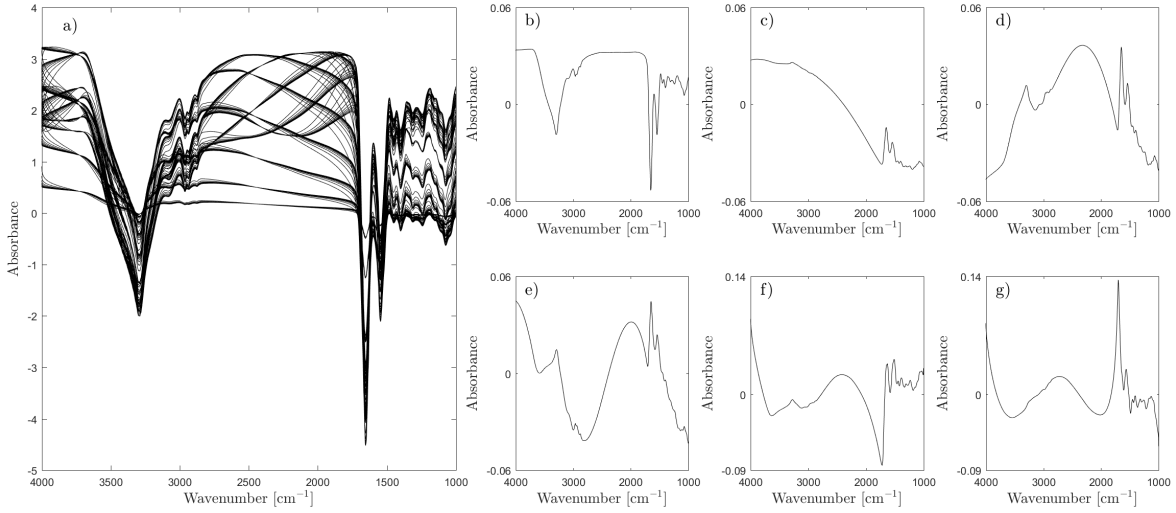


FIGURE 2.16: a) A set of extinction curves which are made orthogonal to the reference spectrum. Parameter ranges of $\alpha_0 \in [2.5, 35.7] \times 10^{-5}$ and $\gamma \in [1.2, 14.6] \times 10^4$ were used as input for Eq. 2.33. b-g) The six first loadings from a PCA on the extinction curves shown in a).

Representational models can be further divided into models of phenomena and models of data.

In *models of phenomena*, we distinguish between different representational styles of models, where the most important includes scale models, idealized models, analogical models and phenomenological models. It is often hard to draw a line between the different styles, as some models exhibits properties from more than one style. Examples of models of phenomena are the billiard ball model of a gas and the Bohr model of the atom.

Models of data cover the representation of processed raw data. They present data in a certain way, e.g. access information that show trends or confirm hypotheses [44, 51]. Data processing is a crucial part of models of data. Data processing may remove noise and outliers in the raw data and at the same time preserve valuable information. The method of data processing must be chosen carefully, and the model may include considerations about experimental method and the phenomenon under investigation. Models of data are sometimes referred to as "soft modelling" [33], as we let the data talk for itself without imposing hard modelling of the data, i.e. models of theory, which are described below.

A **model of theory** is the realization of general laws and axioms governing a system such as the Maxwell theory or the Schrödinger theory. As models of theory are based on causality relations and *a priori* knowledge about the system, the use of these models are sometimes referred to as "hard modelling" [33].

The two categories *representational models* and *models of theory*, are not mutually exclusive. This can be illustrated by Mie scattering in infrared spectroscopy. Mie scattering is a phenomenon which is observed in infrared spectra of single cells, and therefore it describes a part of the world. More specifically, Mie scattering can be classified as a phenomenological model. At the same time, Mie theory is derived from laws and axioms governing the electromagnetic theory by solving Maxwell's equations for scattering at a homogeneous sphere. In this sense, Mie scattering is a model of theory.

In order to model Mie scattering in infrared spectroscopy, soft modelling and hard modelling have been combined in an intermediate model, as done in the Mie meta-model. Neither do we have enough information about the samples to use only the model of theory, nor does biological samples scatter as perfect, homogeneous spheres. The Mie theory is rather used to express possible realizations of the Mie scattering from a sphere of relevant size and optical properties. The possible realization of the hard model is summarized by PCA to a few number of scattering components which constitute the framework for the model of data. The observed Mie scattering is described by soft modelling within these frames.

Chapter 3

Methods

3.1 The fast iterative Mie scattering correction algorithm

As described in section 2.1.5, scattering is dependent on the absorptivity of the scatterer. Therefore, a good estimation of the pure absorbance spectrum, i.e. the reference spectrum in the EMSC model, is needed to predict the Mie scattering precisely. The differences between the measured absorbance spectrum and the predicted absorbance spectrum is expressed in the residuals $\epsilon(\tilde{\nu})$. The underlying idea of the iterative resonant Mie scatter correction algorithm is to update the estimation of the pure absorbance spectrum with the residuals in an iterative process. After each iteration we obtain a gradually better estimation of the true pure absorbance spectrum.

The concept of the iterative algorithm was proposed by Bassan et al. based on the resonant Mie scattering EMSC-model of 2010 [5]. Konevskikh et al. [27, 28] made substantial improvements of the iterative algorithm in 2016 both with respect to the physical model and the performance of the algorithm. In addition to implementing the new improved model, measures were done in order to decrease the computational time. In the following, the latest published version of the algorithm is presented step by step. The algorithm is presented in Fig. 3.1.

Initialization
The algorithm is initialized by selecting a reference spectrum Z_{ref} and parameter ranges for the Mie meta-model, α_0 and γ. The reference spectrum serves as the initial best guess for the pure absorbance spectrum. For the reference spectrum, a standard spectrum such as the Matrigel spectrum can be used, depending on the sample under investigation. Alternatively, a mean spectrum of all spectra in the data set could be used. In cases where the data set contains nearly scatter free spectra, for example in imaging, these spectra may be good candidates for reference spectra. The reference spectrum is in any case preferably baseline free, as all corrected spectra will adapt the baseline of the reference spectrum. In this thesis, the Matrigel spectrum will be used as the initial reference spectrum, if not otherwise stated.

Both the ranges and the grid spacing of the Mie meta-model parameters α_0 and γ need to be specified. To make sure these parameters are within reasonable values, they can be estimated through the radius a, the constant real part of the refractive index n_0 and the scaling parameter h. As default, 10 unevenly spaced values are set for both α_0 and γ. The standard parameter distribution used in this thesis is shown in Fig. 3.2. It can be seen that low values of both α_0 and γ have a tighter spacing than high values. It is observed that lower values of α_0 and γ is beneficial. In order to standardize the parameter ranges, we decided to normalize the reference spectrum with respect to the amide I peak. The reference spectrum needs to be baseline corrected before it is normalized.

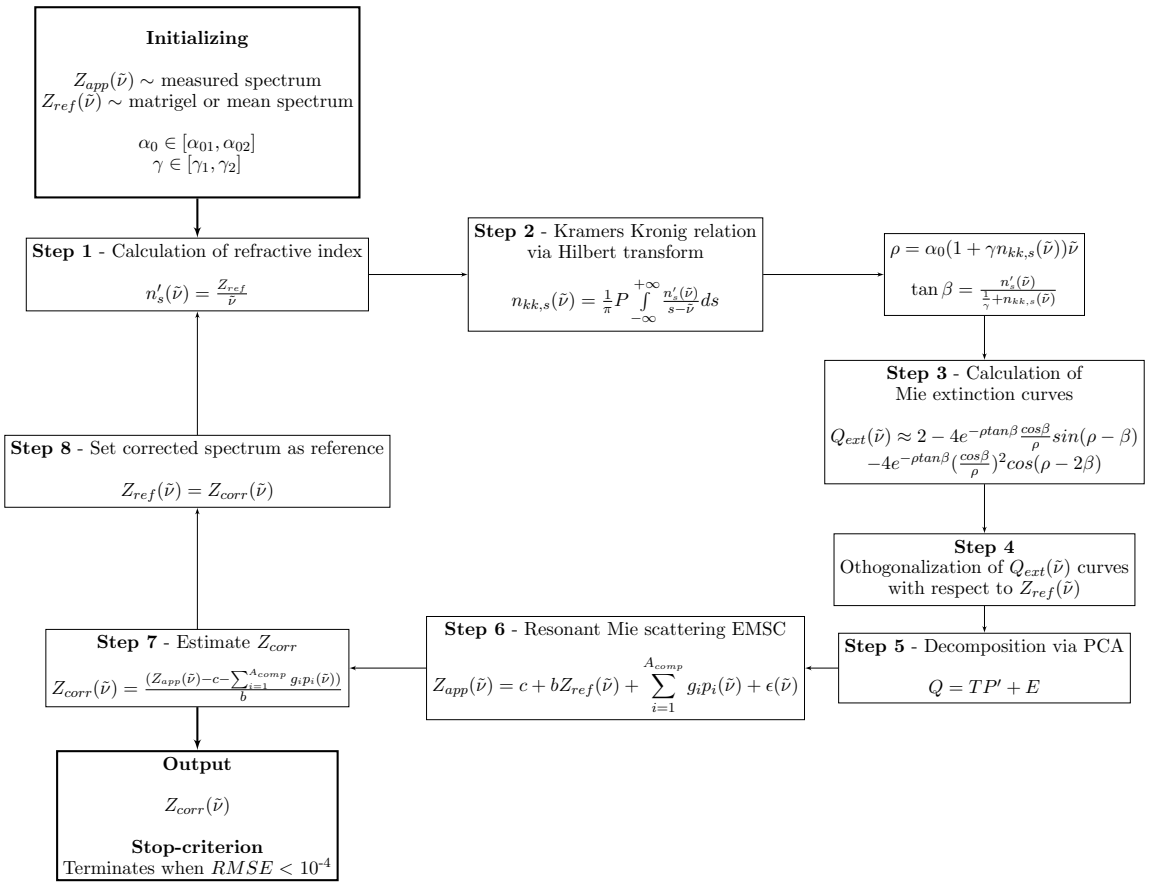


FIGURE 3.1: Schematic representation of the Mie correction algorithm by Konevskikh et al. [26].

Step 1

The scaled imaginary part of the refractive index n'_s is estimated from Eq. 2.45 and the left side of Eq. 2.47. It can be shown that combining these equations gives

$$n'_s = \frac{Z_{ref}}{\tilde{\nu}} \tag{3.1}$$

Step 2

With an estimation of n'_s at hand, the fluctuating real part of the refractive index can be calculated from the Kramers-Kronig relation given in Eq. 2.27. By taking into account symmetry relations, i.e. the fact that the real part of the refractive index is an even function of wavenumbers, and the imaginary part is an odd function, the Kramers-Kronig relation can be calculated via the Hilbert transform given by

$$n_{kk,s} = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{n'_s(s)}{s - \tilde{\nu}} \mathrm{d}s \tag{3.2}$$

Using the Hilbert transform instead of calculating the Kramers-Kronig integral offers a great advantage with respect to computational time. The Hilbert transform is calculated by employing the fast Fourier transform, resulting in a decrease in computational time on a factor of 100 compared to calculating the Kramers-Kronig integral. The Hilbert transform of n'_s with

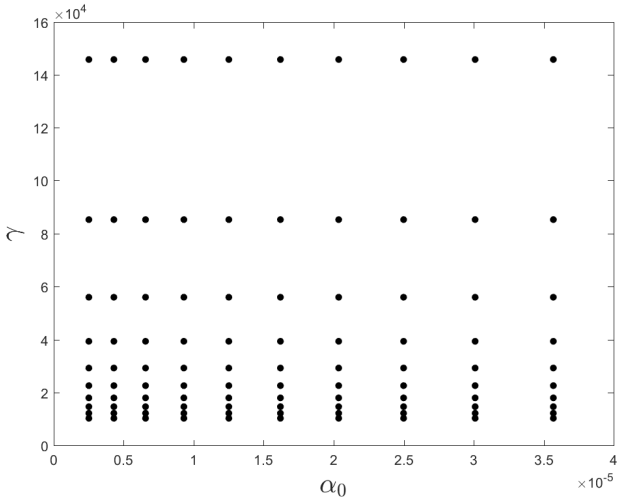


FIGURE 3.2: Default parameter distributions for α_0- and γ-ranges. A non-equidistant spacing is observed to result in a stable correction.

the Matrigel spectrum as Z_{ref} is shown in Fig. 3.3.

Step 3
Mie extinction curves can now be calculated according to the approximation formula for an absorptive sphere, given in Eq. 2.33. With 10 values of both α_0 and γ, we calculate a total of 100 different extinction curves. With the parameter distribution shown in Fig. 3.2 and the Matrigel spectrum as input for the refractive index, the 100 resulting extinction curves are shown in Fig. 2.15 a).

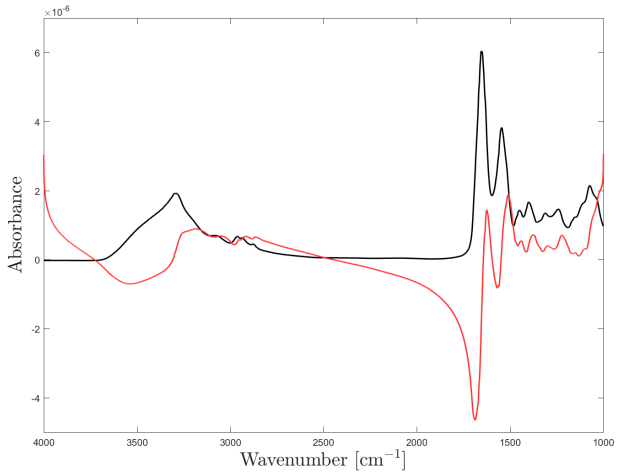


FIGURE 3.3: The scaled imaginary part of the refractive index is shown in black, with the corresponding Hilbert transform shown in red. The red graph is the fluctuating part of the real part of the refractive index.

Step 4
The extinction curves calculated in the previous step are then made orthogonal with respect to the reference spectrum by Eq. 2.49. The reason for why the extinction curves should be orthogonal to the reference spectrum is as mentioned previously due to the fact that the extinction

curves describes both scattering and absorption. The pure absorbance should therefore not be included additionally as the reference spectrum without orthogonalization of the extinction curves with respect to it. The orthogonal extinction curves are shown in Fig. 2.16 a).

Step 5
After orthogonalization, the set of extinction curves are decomposed via PCA, according to Eq. 2.40. The main variance in the set of extinction curves is now expressed by the loadings in $\boldsymbol{P_A}$. In Fig. 2.16 b)-g), the 6 first loadings from the PCA on the set of extinction curves are shown.

Step 6
The loadings in $\boldsymbol{P_A}$, denoted p_i, are included as scatter components in the Mie EMSC. An assessment on the optimal number of loadings A_{comp} is given later. A prediction of the apparent absorbance spectrum is made by estimating the EMSC parameters by least squares regression according to Eq. 2.41. The predicted apparent absorbance spectrum Z_{pred} can be written as

$$Z_{pred} = c + bZ_{ref}(\tilde{\nu}) + \sum_{i=1}^{A} g_i p_i(\tilde{\nu}) \tag{3.3}$$

The residuals ϵ from Eq. 2.41 contains the unmodelled part of the apparent absorbance spectrum, i.e. the error in the prediction. Chemical differences between the reference spectrum and the true pure absorbance spectrum are expressed in the residuals, which is illustrated in Fig. 3.6. Figure 3.6 shows both the measured and predicted apparent absorbance spectrum, in black and red, respectively, and the error of the prediction in blue.

Step 7
Based on the parameters from the Mie EMSC, a corrected spectrum Z_{corr}, can be estimated according to

$$\begin{aligned} Z_{corr}(\tilde{\nu}) &= \frac{Z_{app}(\tilde{\nu}) - c - \sum_{i=1}^{A_{comp}} g_i p_i(\tilde{\nu})}{b} \\ &= Z_{ref}(\tilde{\nu}) + \frac{\epsilon(\tilde{\nu})}{b} \end{aligned} \tag{3.4}$$

where the reference spectrum is updated with the residuals. The corrected spectrum represents in general a slightly better estimation of the pure absorbance spectrum than the employed reference spectrum Z_{ref}. An example of a corrected spectrum is shown in Fig. 3.4. The chemical differences between Z_{ref} and Z_{pure} are expressed in ϵ, which is used to estimate Z_{corr}.

Step 8
The corrected spectrum is now used as the updated reference spectrum, and the correction process is repeated. With a better estimation for the pure absorbance, the apparent absorbance is predicted with a higher precision. We gradually obtain a more precise prediction of the measured spectrum, and accordingly, the residuals decrease in each iteration.

Stop criterion
After a number of iterations, the chemical features of the measured apparent absorbance spectrum are restored in the corrected spectrum. When the correction converges, we have in general obtained a good estimate of the true pure absorbance spectrum. Konevskikh et al. [28]

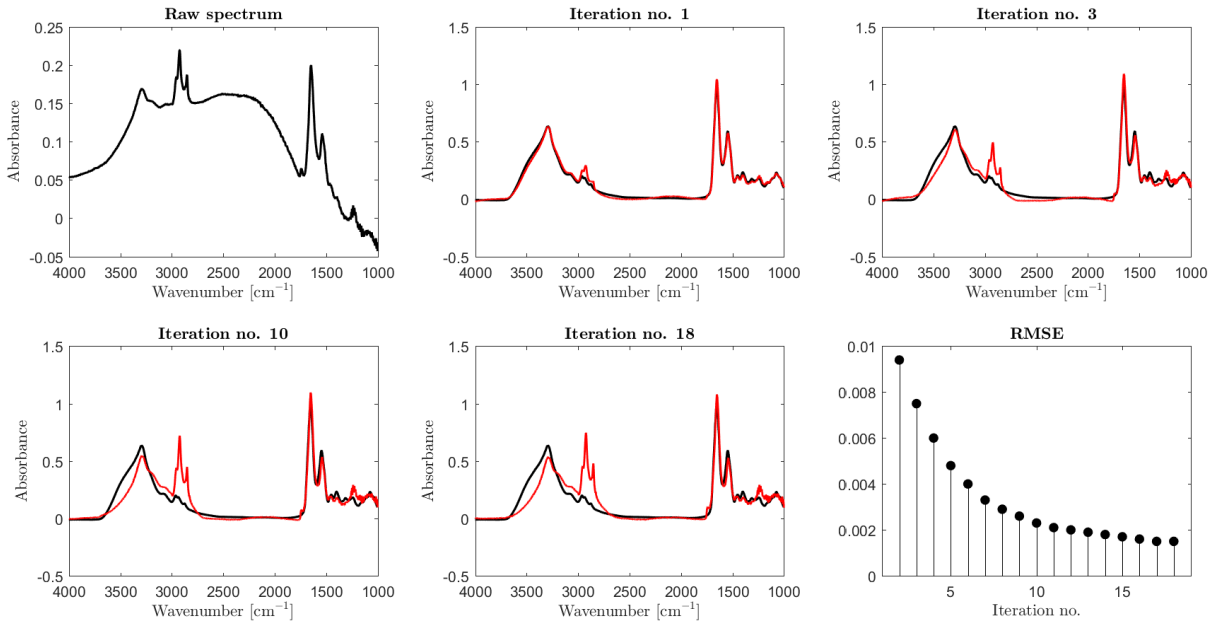


FIGURE 3.4: Illustration of how the corrected spectrum is gradually retrieved. The measured spectrum from a single lung cancer cell [24] (upper left), is corrected with the Matrigel spectrum as the initial reference spectrum, shown in black in the following figures. The corrected spectrum (red) gradually retrieves chemical features of the measured spectrum. In each iteration, the $RMSE$ value decreases (lower right).

proposed to terminate the algorithm when the error of the forward model reached a limit of 10^{-4}. The error is calculated by

$$RMSE = \sqrt{\frac{1}{N}(Z_{app} - Z_{pred})^2} \tag{3.5}$$

where the $RMSE$ is the root mean square error of the forward model and N is the number of wavenumber channels.

Figure 3.4 shows how the corrected spectrum gradually develops from one iteration to the other. In the 1st iteration, the corrected spectrum is dominated by chemical features of the initial reference spectrum. After a few iterations, it is evident that the corrected spectrum retrieves chemical features that are observed in the measured spectrum. Figure 3.6 shows the effect of updating the reference spectrum on the forward model. With a better estimate of the pure absorbance spectrum, both chemical features and scattering features are more precisely predicted.

After a number of iterations, the correction converges. This occurs when the deviation between the predicted apparent absorbance spectrum and the measured apparent absorbance spectrum is small. This is illustrated in Fig. 3.4.

The Mie EMSC model is a forward model

The Mie EMSC model is a forward model, in the sense of seeking to predict the apparent absorbance spectrum, based on an estimation of the pure absorbance spectrum. The pure absorbance spectrum is thus not calculated directly from the measured apparent absorbance spectrum. This differs from inverse modelling, where backward estimations are made based on observations. The forward model is chosen due to the complexity of the Mie formalism:

a direct calculation of the pure absorbance based on the measured extinction efficiency is in general not possible. The concept of the forward model is presented i Fig. 3.5.

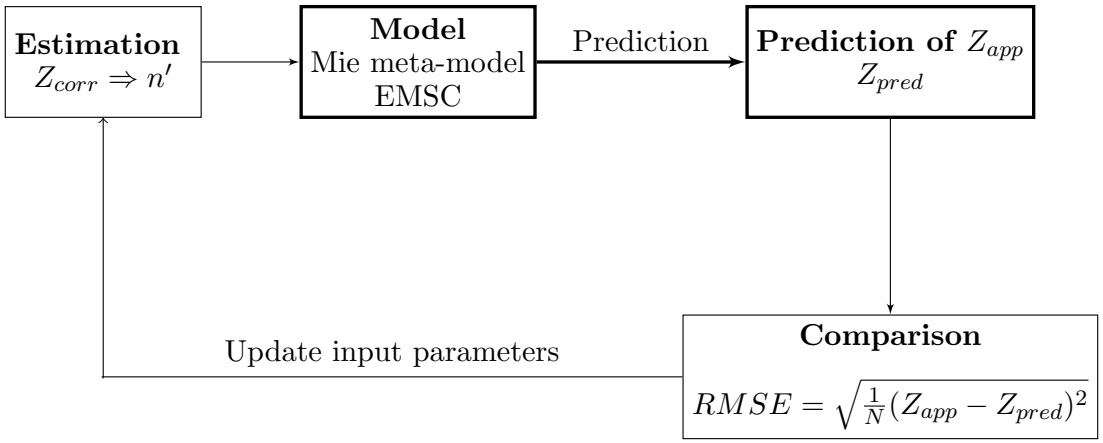


FIGURE 3.5: Schematic representation of the forward model. Based on an estimation of the pure absorbance spectrum, the measured apparent absorbance spectrum is predicted. The error of the model is used to update the estimation for the pure absorbance spectrum.

Estimation: Based on the reference spectrum, i.e. our initial best guess of the pure absorbance, the imaginary part of the refractive index is estimated by Eq. 3.1. With the parameter ranges given for α_0 and γ from Eq. 2.48, the Mie meta-model is established.
Model: By incorporating the Mie meta-model in an EMSC model, the scattering effects are estimated, and the apparent absorbance spectrum is **predicted**.
Comparison: The difference between the predicted apparent absorbance and the measured apparent absorbance, i.e. the $RMSE$, is a measure of how well the model predicts the observation. Based on this deviation, the input parameters for the model are updated. This gives us a new estimate of the sample properties, and a new prediction is made. This iterative process is illustrated in Fig. 3.6, where it is evident that the error of the forward model is decreasing after each iteration, as the apparent absorbance spectrum is more accurately predicted.

3.2 Simulations of apparent absorbance spectra

In order to validate the Mie correction algorithm with the improvements proposed in this thesis, a set of apparent absorbance spectra was simulated. Both the 2010-algorithm [5] and the 2016-algorithm [28] have been validated by use of a simulated data set. In both cases, approximate Mie theory was used to simulate the apparent absorbance spectra: In [5] the Van de Hulst approximation formula [20] employing a real refractive index was used, while in [28] the exact Mie theory employing a complex refractive index was used. In ref. [27], it was shown that the algorithm could not correct the ripples introduced by the exact Mie formalism. This is expected when using the van de Hulst approximation formula which shows no ripples.

In this thesis, a new method for simulating apparent absorbance spectra was developed. The method differs from previous approaches by taking into account that biological cells and tissues do not scatter as perfect and homogeneous spheres. The simulations are subject to the following two requirements:

- For validating the quality of the correction it is required that the underlying pure absorbance spectra are known.

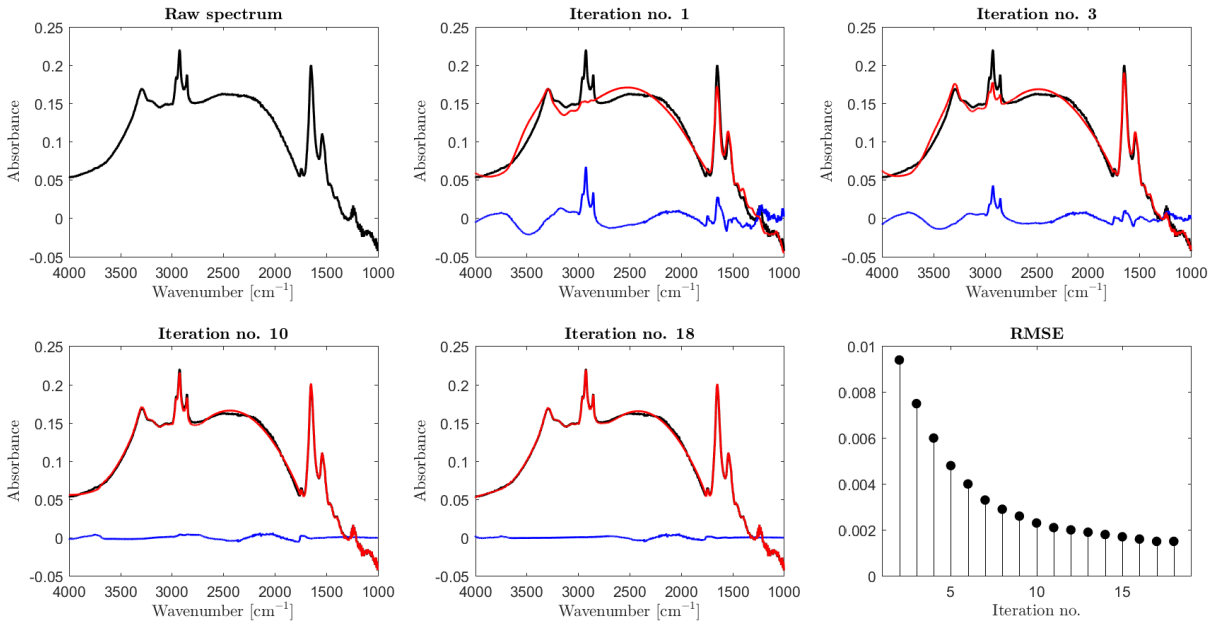


FIGURE 3.6: Illustration of how a more precise prediction of the measured apparent absorbance spectrum is obtained after each iteration. The measured spectrum is obtained from a single lung cancer cell [24]. The predicted spectrum (in red) gradually approaches the measured spectrum (in black). The $RMSE$ is observed to decrease in each iteration.

- For ensuring relevance of the correction for correcting apparent absorbance spectra of cells and tissues, the simulated apparent absorbance spectra should resemble scattering features of measured apparent absorbance spectra of cells and tissues.

In order to fulfill these two requirements, apparent absorbance spectra were simulated according to

$$Z_{app}(\tilde{\nu}) = c + bZ_{pure}(\tilde{\nu}) + \sum_{i=1}^{A} g_i \tilde{p}_i(\tilde{\nu}) \tag{3.6}$$

where each term is analogue to the terms in the Mie EMSC model. Z_{pure} is a simulated pure absorbance spectrum, and the last term on the right-hand side represents the scattering components. Details on how the terms in Eq. 3.6 are obtained will be explained in the following paragraphs.

The pure absorbance spectra Z_{pure} were simulated by using the Matrigel spectrum as a template. Absorbance peaks were modified in order to simulate chemical variations. The Matrigel spectrum was first decomposed into a set of Lorentzian lines. Peak heights and positions were then changed systematically by $\pm$ 20 % and $\pm$ 1 cm^{-1}, respectively. This resulted in simulated pure absorbance spectra, which resembled the Matrigel spectrum, while corresponding to a slightly different absorptivity. Two sets of pure absorbance spectra were simulated with this method, representing two chemically distinct groups, group A and B. Each group consists of 25 spectra, with some random chemical variation within the groups. The two sets are shown in Fig. 3.7 with the Matrigel spectrum plotted in black. Pure absorbance spectra from group A and B are in the following plotted in red and dark blue, respectively.

The physical features expressed in the simulated absorbance spectra are based on experimentally measured apparent absorbance spectra, including 59 spectra obtained from single

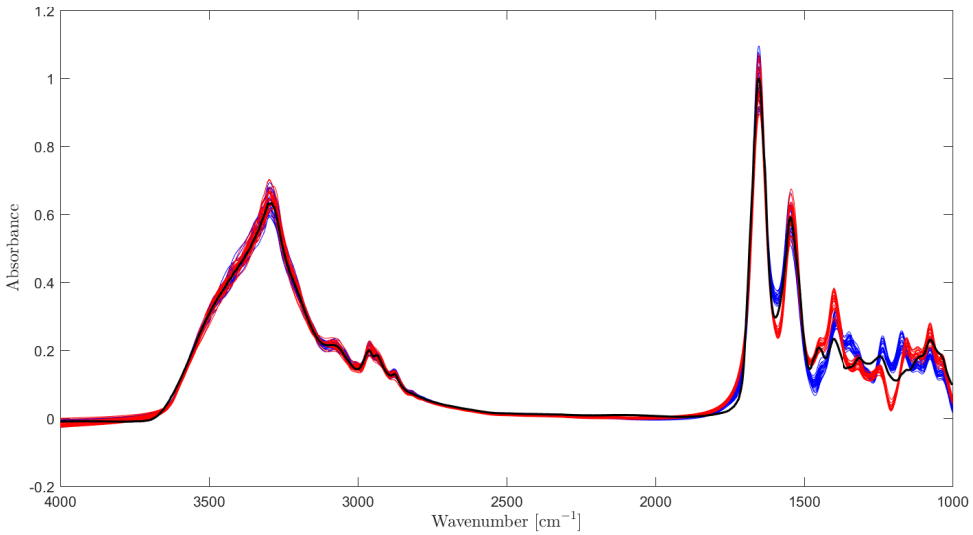


FIGURE 3.7: Simulation of pure absorbance spectra. Two chemically different groups are simulated, group A in red and group B in blue. The Matrigel spectrum (black) was used as a template for the simulations.

lung cancer cells [24]. Chemical and physical features in the experimentally obtained spectra are first parameterized and separated by the Mie correction algorithm presented in this thesis. Parameter ranges for α_0 and γ was set to $[2.5, 35.7]\times10^{-6}$ and $[1.0, 14.6]\times10^{4}$, respectively, with the same distribution as shown in Fig. 3.2, and the number of loadings is set to 7. Physical contributions are now estimated by the parameters c, b and g_i of Eq. 2.41 and inserted directly into Eq. 3.6.

The loadings $\tilde{p}_i$ are obtained by simulating a set of extinction curves with the simulated Z_{pure} as input for the calculations of the complex refractive index. The parameter ranges for α_0 and γ are the same as used in the correction of the measured spectra. The extinction curves are made orthogonal to the simulated pure absorbance spectrum, and subsequently decomposed via PCA. The first seven loadings are included as p_i in Eq. 3.6. In order to simulate apparent spectra, which resemble the experimentally obtained spectra, the Z_{pure} need to be scaled appropriately. The scaling was tuned such that the chemical features were as strongly expressed in the simulated apparent absorbance spectra, as in the experimentally obtained spectra.

Further, notes should be given to Eq. 3.6, emphasizing that this model does not represent a simple additive model where scattering and absorption are treated independently. By calculating the extinction efficiency based on the simulated pure absorbance spectrum, Q_{ext} expresses all types of extinction, as described in section 2.1.5. By orthogonalizing the loadings with respect to Z_{pure}, we make sure that the chemical features of the absorbance spectrum are not included in both p_i and Z_{pure}.

Simulations according to this method result in apparent absorbance spectra, which resemble measured apparent absorbance spectra. In addition, the method allows to define distinct chemical features in underlying pure absorbance spectra, which are known *a priori* and can be used for validating the correction by the Mie EMSC correction. An example of a simulated spectrum can be seen in Fig. 3.8. In Fig. 3.8 a), the underlying pure absorbance spectrum was chosen from group A and is plotted in red, with the Matrigel spectrum in black. Figure 3.8 a) shows the simulated apparent absorbance spectrum in orange. The measured spectrum,

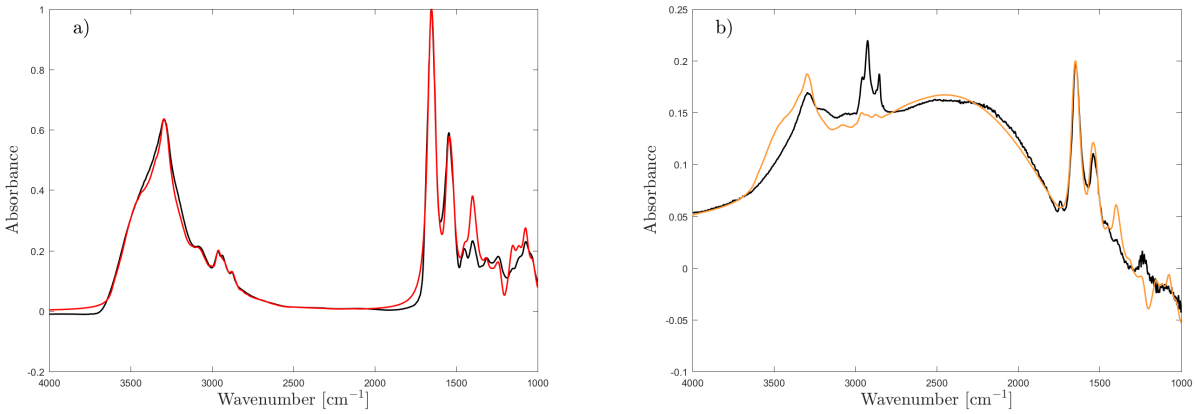


FIGURE 3.8: Simulation of an apparent absorbance spectrum. a) The underlying pure absorbance spectrum (red) was chosen from group A. The Matrigel spectrum (black) was used as a template for simulating pure absorbance spectra. b) The apparent absorbance spectrum (orange) was simulated to mimic scattering features from an experimentally obtained spectrum (black) obtained from a single lung cancer cell [24].

which is used as a template for the scattering contributions, is shown in black. Simulated apparent absorbance spectra with an underlying pure absorbance spectrum from group A are in the following plotted in orange. When a pure absorbance spectrum from group B is used, the apparent absorbance spectrum is plotted in light blue.

Chapter 4

Results and discussion

In this thesis, the Mie correction algorithm of Konevskikh et al. [26] has been validated with a set of simulated apparent absorbance spectra. This resulted in a number of improvements of the algorithm, by further optimization and stabilization. The improvements include:

- Optimization of the number of principal components used in the Mie EMSC model
- Optimization of the stop criterion
- Increased stabilization by weighting the reference spectrum
- Guaranteed positive reference spectrum
- Standardization of initialization parameters

When validating the algorithm with a set of simulated spectra, critical features of the algorithm have been reviewed, including the dependency on the initial reference spectrum, the ability to retrieve the true amide I peak position and the sensitivity towards the initialization parameter settings. As a result of this work, a user-friendly Matlab code for Mie scatter correction will be published based on the algorithm of Konevskikh et al. [26]. The examples given in section 4.5 may serve as a guide for the user to achieve optimal parameter settings for a given set of spectra.

Finally, the use of the algorithm is demonstrated on spectra from an IR imaging data set. This example illustrates how scatter and chemical features can be separated and analyzed. It further demonstrates how the Mie EMSC parameters can be used for morphological analysis of spectra.

4.1 Improvements of the Mie correction algorithm

The proposed improvements to the Mie correction algorithm are illustrated in Fig. 4.1, where the red markings indicates the changes with respect to the algorithm of Konevskikh et al. [26].

Setting the number of loadings

The number of loadings A_{comp} included in the Mie EMSC model affects the precision of the model, its stability and the computational time. Thus, the parameter A_{comp} should be chosen carefully, and such that the Mie oscillations are represented precisely. However, the optimal number of loadings has not been discussed in previous literature. We propose that the number of loadings should be set automatically by the Mie correction program, based on a predefined desired level of explained variance in the set of Mie extinction curves. The number of loadings A_{comp} is set in the first iteration, and is calculated directly from the level of explained variance.

The parameter A_{comp} is dependent on the ranges for α_0 and γ, as well as the grid spacing. The effect of changing the parameter ranges on the level of explained variance is illustrated in

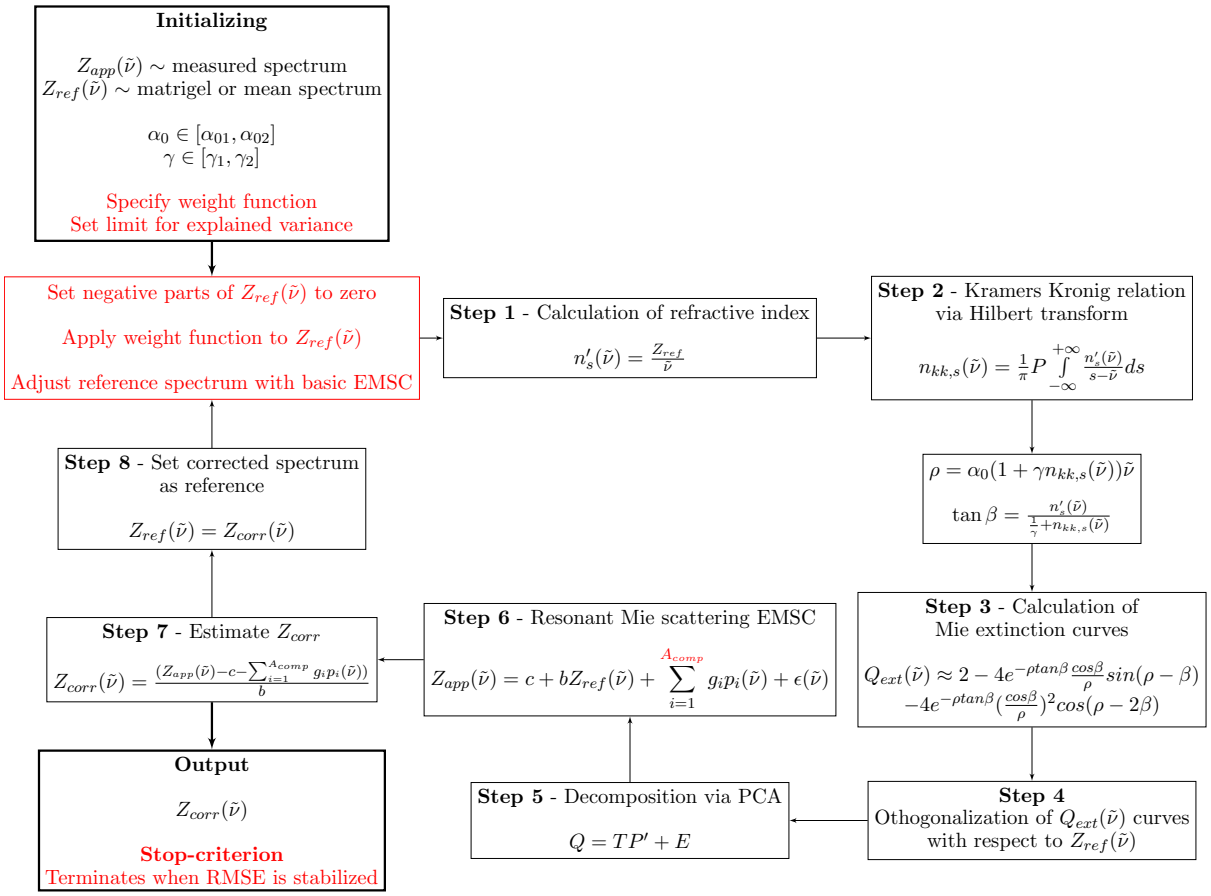


FIGURE 4.1: Schematic representation of the Mie correction algorithm, where the red markings indicates the proposed improvements.

Fig. 4.2. This figure shows how the explained variance in the Mie extinction curves is increased by increasing the number of principal components used in the reconstruction. In the calculation of the Mie correction curves, the Matrigel spectrum was used as input for the imaginary part of the refractive index, and the parameter ranges and distributions were changed. The black graph corresponds to the default parameter ranges. When the parameters are equidistantly spaced, the explained variance does not change considerably, as shown by the red graph. The explained variance is mostly affected by the extension of the parameter ranges, illustrated by the green and blue graph, where green corresponds to decreased parameter ranges and blue corresponds to increased parameter ranges. Details on how the set of Q_{ext} depends on these parameters will be studied elsewhere. Experiences with the data sets at hand for this thesis shows that a level of explained variance at 99.96 % to 99.99 % usually results in precise representations of the Mie oscillations. To obtain this level of explained variance, A_{comp} is typically 7-9.

If the Mie oscillations are not precisely predicted, increasing the number of loadings may lead to a more accurate representation. At one point, adding more components gives a negligible contribution to the model. If the number of loadings is set too high, it may instead result in loadings with strong oscillations leading to instability. Notes is given in section 4.5 on how the user can determine if the level of explained variance is adequate for a given data set. An option to set the number of loadings directly, and not through the explained variance, is added in the Mie correction code, in order to make it convenient for the user to regulate this parameter.

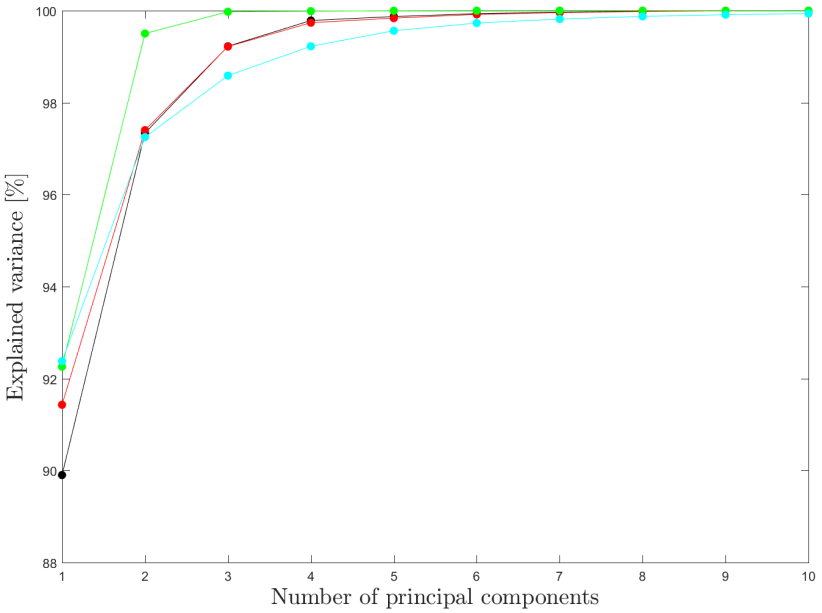


FIGURE 4.2: Explained variance in a set of Mie extinction curves, by number of principal components used in the reconstruction. The graphs correspond to different parameter ranges and distributions for the parameters α_0 and γ in Eq. 2.48. Black: standard ranges and distributions. Red: equidistant distribution. Blue: increased parameter ranges. Green: decreased parameter ranges.

Revised stop criterion

In the latest published version of the Mie correction algorithm, the stop criterion was based on the error of the forward model. The $RMSE$ is expected to decrease for each iteration, and the algorithm was terminated when the $RMSE$ reached a pre-defined limit of 10^{-4}. Nevertheless, in many cases the $RMSE$ does not reach an absolute lower limit. The final $RMSE$ that is reached dependents on the dataset and the initialization parameters. There may be features in the apparent absorbance spectra that are poorly described by the Mie EMSC model. If these features are prominent, the minimum $RMSE$ that can be reached has a relatively high value. Figure 4.4 illustrates how the final $RMSE$ after 45 iterations vary within a data set. The corrected lung cancer cell spectra are shown in Fig. 4.4 a), and the final $RMSE$ for all spectra are shown in Fig. 4.4 b). The lowest obtained $RMSE$, i.e. 6.9×10^{-4}, was obtained for spectrum no. 25.

We propose to use a flexible limit for the $RMSE$, rather than an absolute limit, based on an evaluation on whether the $RMSE$ has converged or not. If the $RMSE$ has converged, or if the $RMSE$ increases from one iteration to the other, the algorithm is terminated. The development of the $RMSE$ after each iteration is shown in Fig. 4.4 c) for three spectra from the data set. The dots mark at which iteration the correction is terminated when the new stop criterion is applied. The spectra corresponding to the $RMSE$ developments in c) are shown in Fig. 4.4 d), at the iterations marked with the dots. Figure 4.4 c) illustrates in addition that the $RMSE$ may increase from one iteration to the other. The spectrum corresponding to the increasing $RMSE$ is plotted in Fig. 4.4 e) for iteration no. 29 in blue, and iteration no. 32 in red. It is evident that the rise of the $RMSE$ is due to artifacts in the baseline, which arise mainly close to the amide I absorption band. The $RMSE$ is not observed to increase when weighting is not applied. The reason for this is not completely clear and will be subject for future investigations. Further, it is evident that a too high number of iterations may introduce

more distortions in the baseline, as seen when comparing the red spectrum in Fig. 4.4 d) and e). It is important to note that these distortions in the corrected spectrum are most likely due to an artifact in the raw spectrum. In Fig. 4.4 f) it is shown that the distortions in the region between 2,100-2,300 cm^{-1} arise due to an artifact which is present in the raw spectrum.

When applying the improved stop criterion, termination occurs when the model has separated and parameterized the chemical and physical features. The final $RMSE$ may vary within the data set as show in Fig. 4.4 b), but strong deviations from the mean $RMSE$ can be taken as an indication of an unsuccessful correction. Thus, a simple quality test is implemented by letting the user set a limit for the maximum $RMSE$. As default, this limit is set to infinity. An example of how this quality test can be used in the initialization parameter estimation is given in section 4.6.

Weighting the reference spectrum

When infrared spectra are recorded, often the same background intensity I_0 is used for several measurements. When the experimental environment changes, such as changes in humidity or carbon dioxide concentration in the air, the background spectrum does no longer represents the background in the measurements, and consequently I_0 does not any longer correct these contributions in the spectrum. Therefore, absorption bands corresponding to disturbances from the system or surroundings will emerge in the measured spectrum. An example is absorption by carbon dioxide molecules in the air around 2,300 cm^{-1}, which lies in the silent region of the absorbance spectra. As these absorption modes are not caused by absorption by the sample, they should not be modelled as a characteristic of the scatterer. Disturbances in the absorbance spectra should therefore not be included when calculating the imaginary part of the refractive index n', as these modes are not deformed by the Mie scattering.

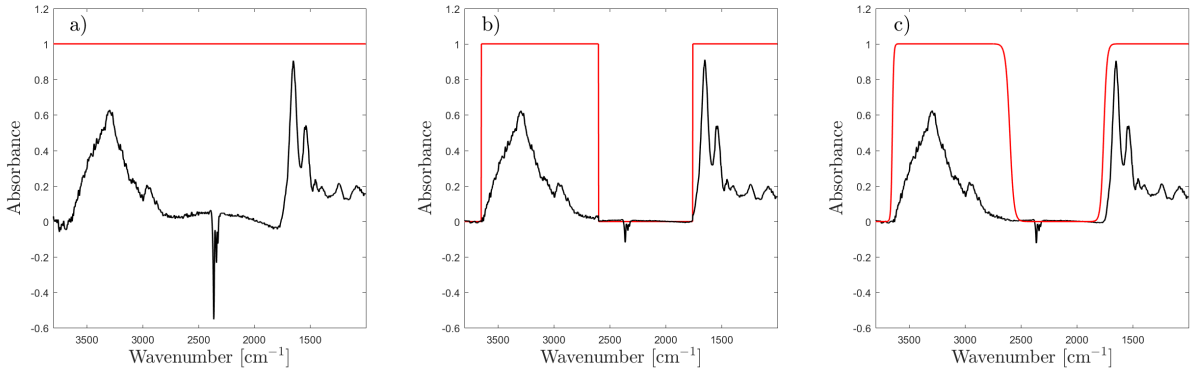


FIGURE 4.3: The effect of weighting the reference spectrum. a) No weighting. b) Weighting with a boxcar function. c) Weighting with hyperbolic tangent funtions between chemically active and inactive regions.

We handle disturbances in the silent region by weighting the reference spectrum, which affects both the calculation of the complex refractive index by Eq. 3.1 and the parameter estimation in the EMSC in Eq. 2.41. In regions where the scatterer is chemically active, the reference spectrum is weighted by 1, while the chemically inactive regions are weighted by 0. The effect of weighting the reference spectrum is illustrated in Fig. 4.3, where a) shows the correction without weighting. When weighting the background spectrum with a boxcar function, as shown in Fig. 4.3 b), the corrected spectrum shows non-smooth transitions between the chemically active and the chemically inactive regions. This can be seen by the transitions in the regions around 1,760 cm^{-1} and 2,600 cm^{-1}.

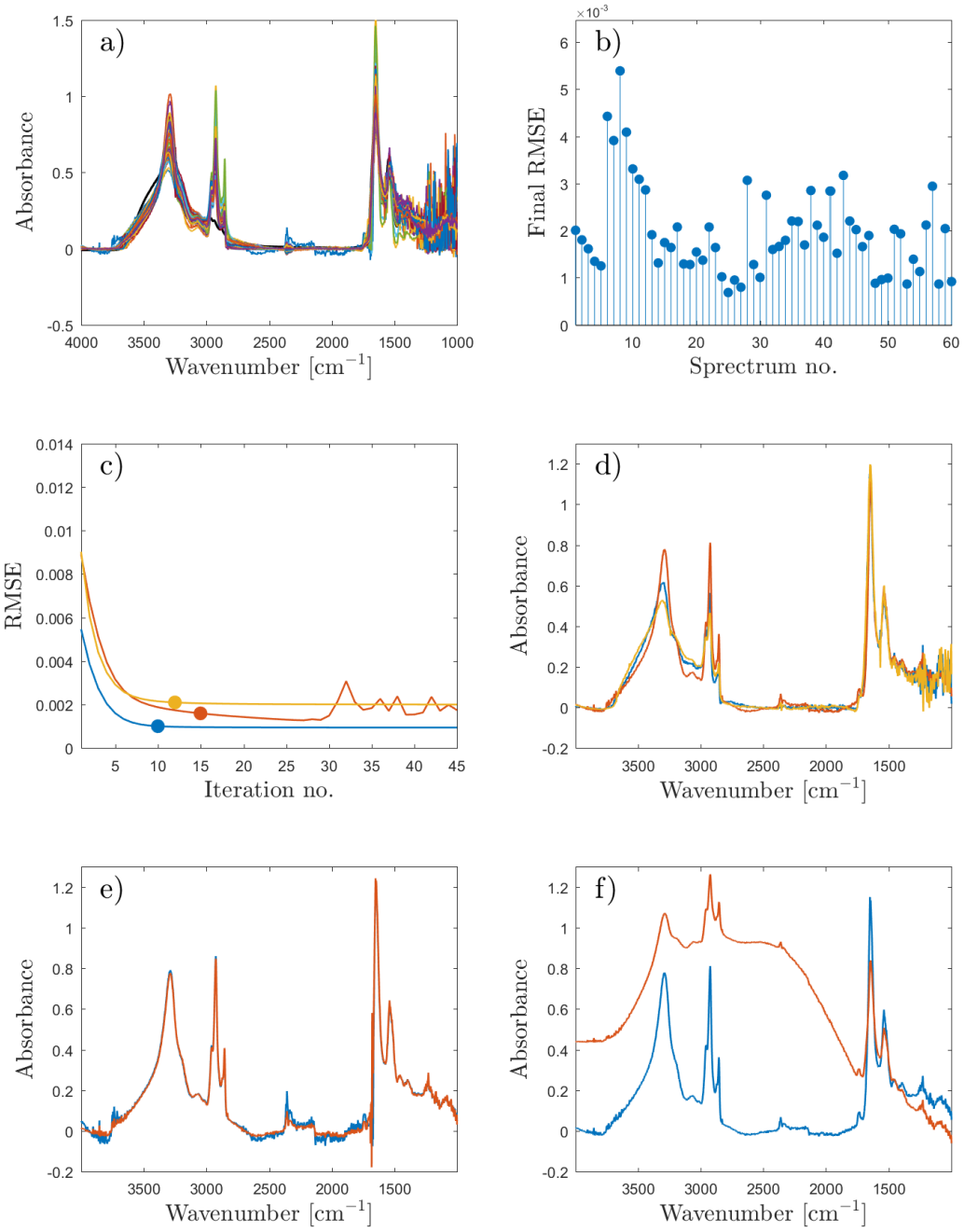


FIGURE 4.4: a) In total 60 apparent absorbance spectra obtained from lung cancer cells [24] are corrected with 45 iterations. b) The final $RMSE$ for each spectrum after 45 iterations. c) The development of the $RMSE$ for three spectra. It is evident that the $RMSE$ may increase from one iteration to the other. The markings indicates termination by the new stop criterion. d) The corrected spectra corresponding to the markings in c). e) The spectrum corresponding to the red $RMSE$ in c), plotted for iteration no. 29 and 32. f) The raw spectrum and the corrected spectrum corresponding to the red graph in c). The new stop criterion is applied.

In order to achieve a smooth transition between the chemically active and inactive regions, we employed a hyperbolic tangent functions $W(\tilde{\nu})$ according to

$$W(\tilde{\nu}) = \frac{1 \pm \tanh\left(\kappa(\tilde{\nu} - \tilde{\nu}_0)\right)}{2} \tag{4.1}$$

for the transitions between chemically active and chemically inactive regions. The parameter κ determines the slope of the function and $\tilde{\nu}_0$ is the inflection point. The sign of κ determines if the transition is from a chemically inactive region to an inactive region or vice versa. Figure 4.3 c) shows the corrected spectrum in black, plotted with the weight function built up by hyperbolic tangent functions $W(\tilde{\nu})$ in red. As a default, the inflection points of the weight function is set to 3,700 cm^{-1}, 2,550 cm^{-1} and 1,900 cm^{-1}, with $\kappa = 1$ for all points. The weight function should be customized to the different data sets by specifying the region for chemical activity. By regulating κ, a smooth transition is obtained. It is obvious that the hyperbolic tangent functions achieve smoother transitions points than a boxcar function.

Applying weighting of the reference spectrum is observed to stabilize the baseline correction, and in addition prevent the absorption bands in the chemically inactive regions to increase. This effect has been observed in the correction of spectra with strong features in this region. Further examples illustrating the effect of weighting is given in section 4.5.

Guaranteed positive reference spectrum

In ref [27], Konevskikh et al. proposed that the negative parts of the reference spectrum should be set to zero when calculating the imaginary part of the refractive index according to Eq. 3.1. This modification was done due to physical considerations stating that the absorbance should not assume negative values. For the same reason, we propose to set negative parts of the reference spectrum to zero in the EMSC parameter estimation as well. No significant changes in the correction has been observed when implementing this adjustment, as illustrated in Fig. 4.5 a). This figure shows a spectrum corrected with and without this functionality. In Fig. 4.5 b) the small chances are visualized by plotting the differences between the two spectra in figure a). The vertical red lines marks the position of the inflection points of the weight function, and it is evident that the small differences appear in this region. Although the differences are negligible, we suggest to implement this due to physical considerations.

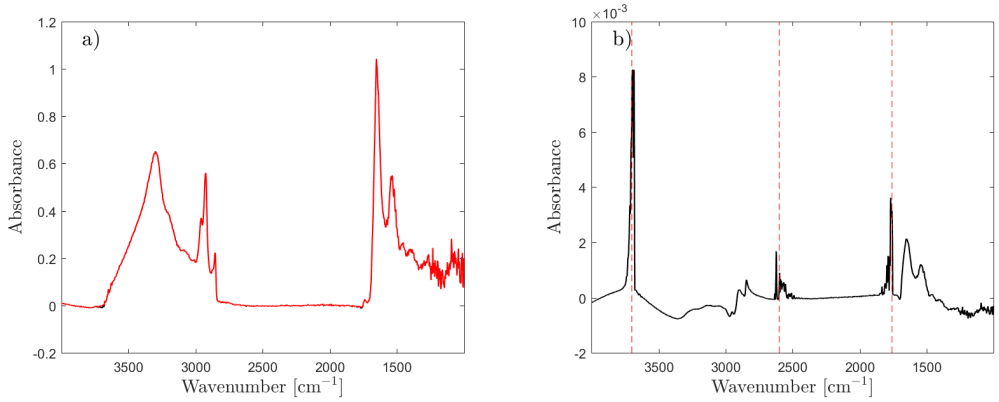


FIGURE 4.5: The effect of setting negative parts of the reference spectrum to zero. a) The red spectrum is corrected with this functionality, the black spectrum is corrected without. The differences are not visible. b) The black spectrum is subtracted from the red spectrum, and the difference is plotted. The red dotted lines indicates the inflection points of the applied weight function.

Standardization of the initialization parameters

The parameters α_0 and γ relate to physical properties of the sample through Eq. 2.48, and should therefore be adjusted to these properties. Because of this, these parameters are set by specifying a, n_0 and h. Some *a priori* knowledge about the sample is therefore desired. We propose the following default settings for a, n_0 and h:

$a \in [2\,\mu\text{m}, 7.1\,\mu\text{m}]$

$n_0 \in [1.1, 1.4]$

$h = 0.25$

With the data sets at hand for this study, it was observed that a stable correction is obtained, when the grid spacing of α_0 and γ is non-equidistant. The grid spacing showed in Fig. 3.2 is chosen as a default setting. This grid spacing is obtained by letting the ranges of a and n_0 consist of equidistant points, and subsequently calculating α_0 and γ by vector multiplication.

As mentioned previously, the reference spectrum should be both baseline corrected and normalized in order to standardize the initialization parameters α_0, γ and h. The scaling of the reference spectrum should thus be preserved in each iteration. The Mie EMSC is expected to control the scaling in each iteration step. Nevertheless, since the reference spectrum is updated in each iteration step, and may therefore also have slight scaling changes, an additional scaling is introduced by employing a basic EMSC in each iteration step. The reference spectrum is scaled to match the scaling of the initial reference spectrum, according to

$$Z^k_{ref}(\tilde{\nu}) = c + bZ^0_{ref}(\tilde{\nu}) + d\tilde{\nu} + e\tilde{\nu}^2 \tag{4.2}$$

where Z^k_{ref} is the reference spectrum in iteration no. k, and Z^0_{ref} is the initial reference spectrum. By implementing this step in the iterative algorithm, small but gradual changes with respect to the scaling is avoided. It has been observed that for some data sets, the reference spectrum tends to grow by 1 % in each iteration. The adjustment in each iteration is observed to be small, which is taken as an indication of stability of the algorithm.

4.2 Validation of the algorithm

The improved iterative Mie correction algorithm proposed in the previous section was validated with a set of simulated apparent absorbance spectra. When previous versions of the algorithm has been validated with a set of simulated apparent absorbance spectra, it was concluded that the simulated apparent absorbance spectra contained features that were not observed in measured spectra. Therefore, in this thesis one of the aims was to establish a simulated data set of apparent absorbance spectra that mimics measured data, and which at the same time is built up according to Mie theory. This simulated data set enables us to validate if the algorithm retrieves the chemical features of the underlying pure absorbance spectra, and if the algorithm is stable. The performance of the algorithm is evaluated by comparison of the corrected spectra with underlying pure absorbance spectra and cluster analysis with PCA.

4.2.1 Simulation of pure absorbance spectra

As a basis for comparison in the cluster analysis, a set of pure absorbance spectra were simulated according to the method described in section 3.2. The data set represents two chemically different groups, with some random chemical variation within each group. In Fig. 4.6 a), the two groups of pure absorbance spectra are plotted together with the Matrigel spectrum, which

was used as a template for the simulations. Spectra from group A are plotted in red, and spectra from group B are plotted in blue. The Matrigel spectrum is plotted in black. By PCA on the set of pure absorbance spectra, the main sample variation patterns can be expressed by only a few principal components. Here and in the following, the PCA is performed on the spectral range from 1,000 cm^{-1} to 1,400 cm^{-1}, as the main chemical differences are introduced in this region. In a score plot of the second and first principal component, we obtain two distinct clusters which separates the two groups, as seen in 4.6 b). The Matrigel spectrum is located in the middle of the two clusters. The first and second loadings are shown in Fig. 4.6 c) and 4.6 d), respectively. It is evident that the two loadings resolve chemical variations in the data set. The score plot in Fig. 4.6 b) will in the follwoing be used as a reference to determine whether the correction can be considered successful or not by means of classification. Apparent absorbance spectra are simulated by using one spectrum from each group as input for Eq. 3.6, as described in the following.

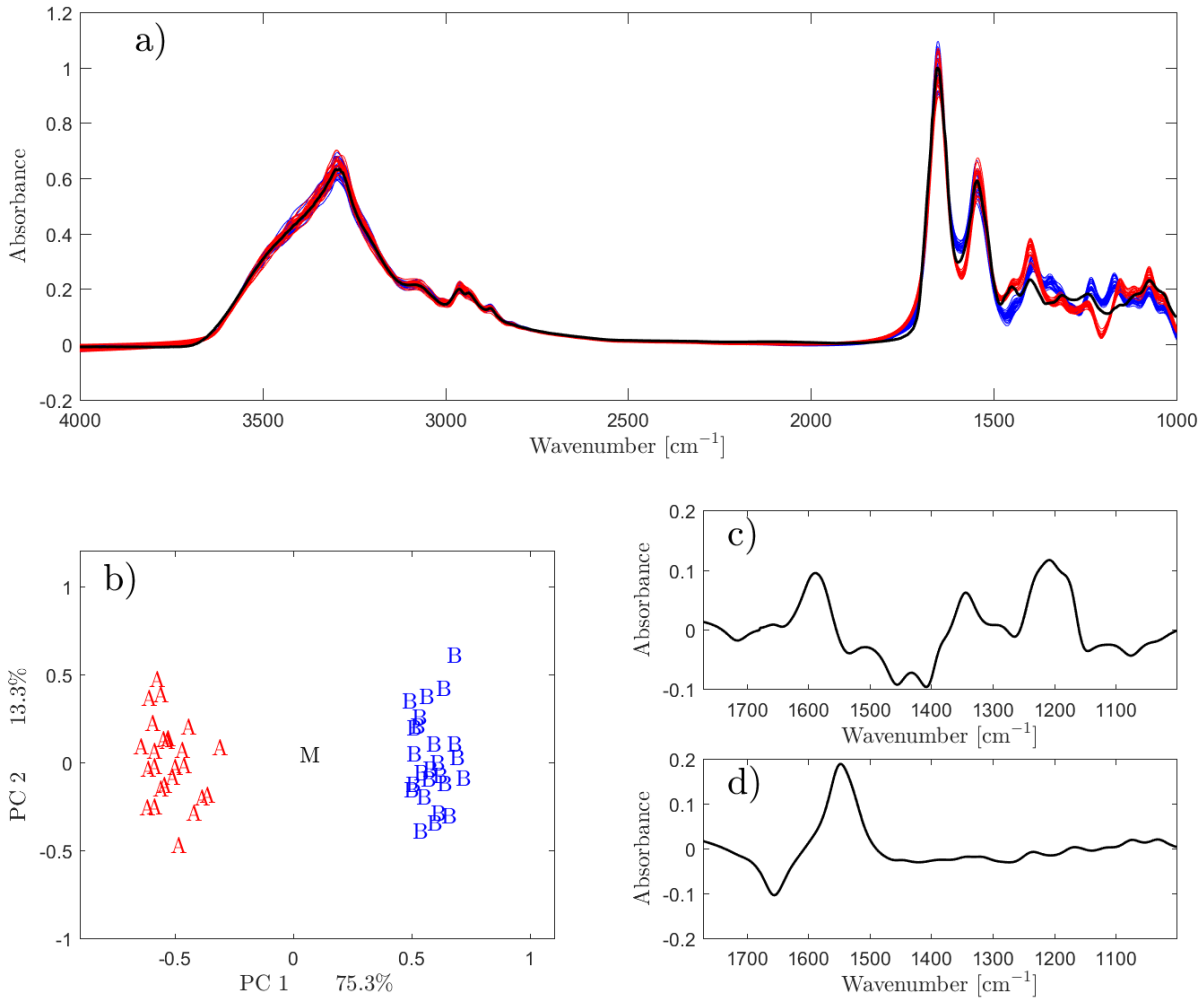


FIGURE 4.6: a) The simulated set of pure absorbance spectra which are based on the Matrigel spectrum. Group A is shown in red, group B is shown in blue. b) The score plot of the two first principal components of a PCA on the set of pure absorbance spectra. Two clusters are obtained. c) The first loading from the PCA, and b) the second.

4.2.2 Simulation of apparent absorbance spectra

The simulated apparent absorbance spectra were obtained by using Mie theory, by employing a real and imaginary part of the refractive index as input for the simulation of the extinction efficiency. In addition, they were simulated to mimic measured absorbance spectra, by estimating scattering parameteres directly from a set of measured apparent absorbance spectra. This is an improvement with respect to previous simulations of apparent absorbance spectra. In Bassan et al. [5], apparent absorbance spectra were simulated by adding simulated pure absorbance spectra to Mie extinction curves that were obtained by employing a constant real refractive index. The obtained simulated apparent absorbance spectra were thus based on the assumption that scattering and absorption are additive effects. The same assumption was subsequently used in the Mie EMSC model that was used for correction. In ref. [27], Konevskikh et al. simulated apparent absorbance spectra by full Mie theory, resulting in spectra with ripples that are not observed in measured spectra. It was observed that the algorithm could not correct the obtained ripples [27].

In this thesis, simulations of apparent absorbance spectra were done according to the method described in section 3.2. One pure absorbance spectrum from each group was used as a basis for the simulations, and a variety of scatter contributions were obtained from 59 experimentally obtained spectra. The different scattering signatures are illustrated in Fig. 4.7 a), where the apparent absorbance spectra are shown. Spectra with an underlying pure absorbance spectrum from group A is plotted in orange, and spectra corresponding to group B is plotted in light blue..

The scattering features observed in Fig. 4.7 a) dominates the main variance in the set of apparent absorbance spectra, such that a score plot of the first two loadings from a PCA on this data set does not allow for chemical classification. The score plot is shown in Fig. 4.7 b), and it is evident that the apparent absorbance spectra are spread out and the sample grouping that was observed in the pure absorbance spectra cannot be seen anymore. The first and second loading is presented in Fig. 4.7 c) and d) respectively. It is evident that the first two loadings expresses scattering featured rather than chemical features.

A similar effect of the Mie scatter disturbances in simulated apparent absorbance spectra was observed in previous simulations [5, 27].

4.2.3 Retrieval of pure absorbance spectra

The simulated apparent absorbance spectra shown in Fig. 4.7 a) were corrected with the improved Mie correction algorithm, with the Matrigel spectrum as the initial reference spectrum. The initialization parameters α_0, γ and h were set to the default values. The number of loadings included in the Mie EMSC model was set to 7, as a result of a level of explained variance at 99.96 %. Figure 4.8 a) shows the retrieved pure absorbance spectra. The red and blue spectrum are the underlying pure absorbance spectra from group A and B, respectively, and the Matrigel spectrum is shown in black. The corrected spectra are shown in orange and light blue, corresponding to the pure absorbance spectra from group A and B, respectively. By visual inspection, it is clear that the correction algorithm retrieves the chemical information from the apparent absorbance spectra. In order to evaluate the correction with respect to the grouping of the pure absorbance spectra, the corrected spectra were projected into the score plot from the PCA on the pure absorbance spectra shown in Fig 4.6 a). The projection is shown in Fig 4.8 b). From this figure we see that the corrected spectra cluster around the corresponding underlying pure absorbance spectrum used for the simulations. The spread of the corrected spectra is significantly less than the spread within the chemical group, and so by means of classification with PCA, the correction is considered highly successful.

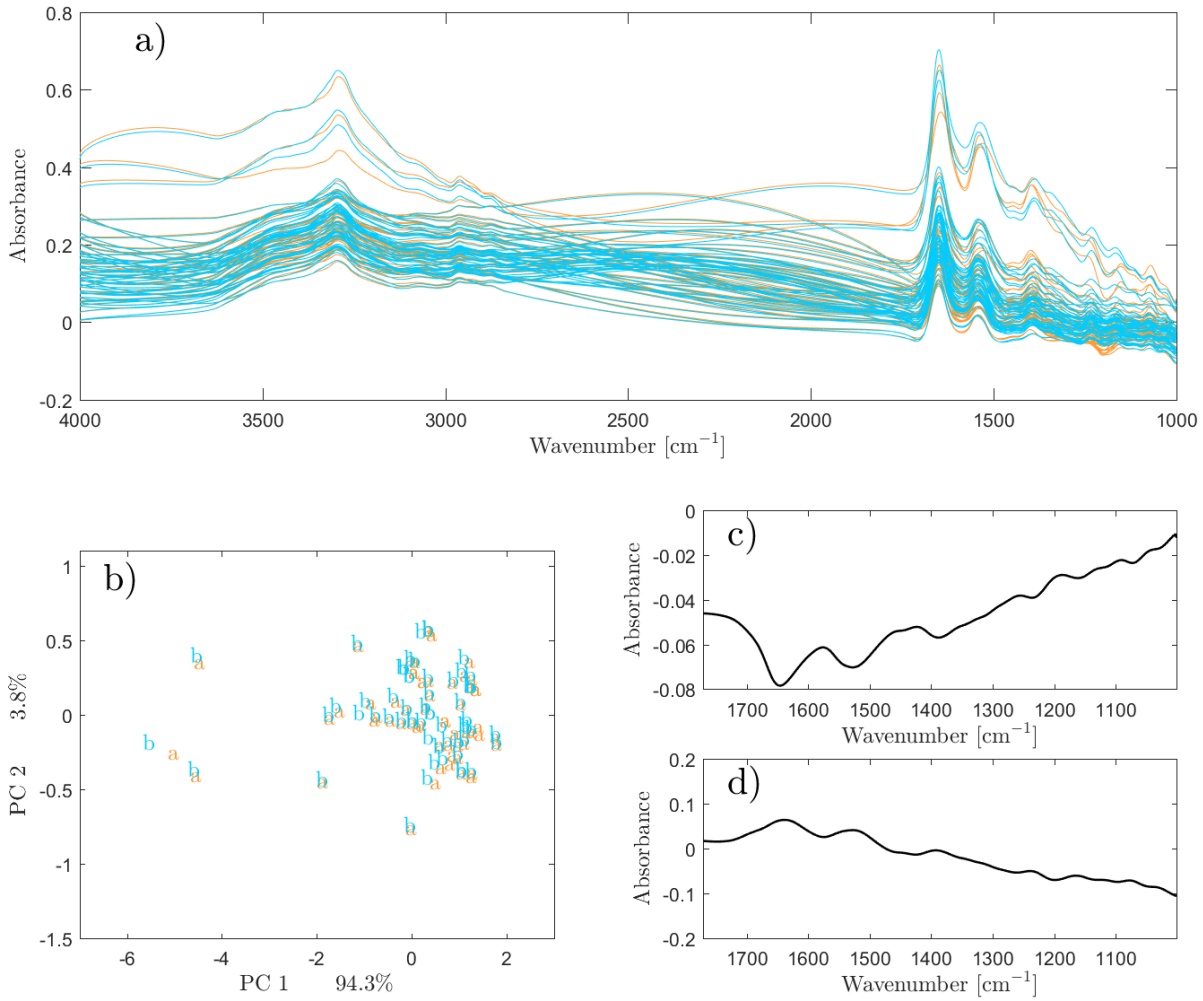


FIGURE 4.7: a) The set of simulated apparent absorbance spectra, where orange indicates an underlying absorbance spectrum from group A, and light blue indicates a pure absorbance from group B. b) In the score plot of the first two principal components from a PCA on the spectra in a). c) and d) shows the first two loading vectors.

In Bassan et al. [5], the spread of the corrected spectra obtained from a simulated data set was also evaluated by projecting the corrected spectra into the PCA model obtained from the simulated pure absorbance spectra. It was concluded that the corrected spectra have a strong tendency to assume features from the reference spectrum employed, as illustrated in Fig. 4.9. The boxes are added to mark the regions of which the corrected spectra were located in the projection. The corrected spectra adapted chemical features from the initial reference spectrum, which can be clearly seen in Fig. 4.10 c) [5]. When validating our code, this tendency was not observed anymore. The algorithm is shown to be reliable by means of classification with PCA, and the pure absorbance spectra are retrieved.

4.3 Dependency on the reference spectrum

The reference spectrum plays an essential role in the Mie EMSC model. In general, it is an advantage if the reference spectrum that is used for the initialization of the algorithm contains

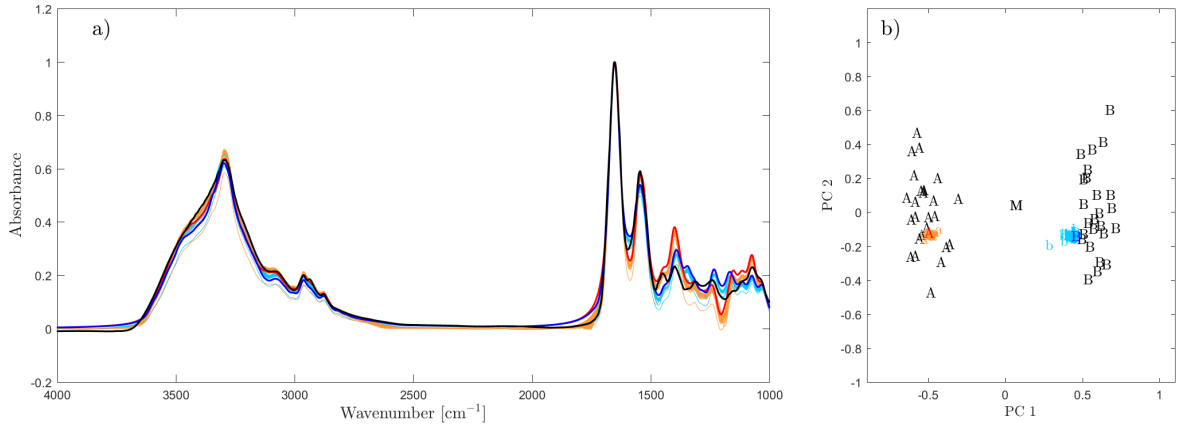


FIGURE 4.8: a) The correction of the simulated apparent absorbance spectra in Fig. 4.7 a). b) The corrected spectra are projected into the score plot shown in Fig. 4.6 b)

similar major chemical features as the true pure absorbance spectrum. Its overall shape relating to the total amount of carbohydrates, protein, fat and water should show similar features as the pure absorbance spectra underlying the apparent absorbance spectra. The reference spectrum serves as our initial best guess for the pure absorbance spectrum, and the more chemical features the reference spectrum and the apparent absorbance spectra share, the faster the algorithm will converge towards the true underlying pure absorbance spectrum. In addition, modelling around a reference spectrum stabilizes the correction. As mentioned above, earlier versions of the Mie correction algorithm have shown to be biased by the reference spectrum used to initiate the algorithm. If chemical features of the reference spectrum are adapted by the corrected spectrum, the chemical features in the corrected spectrum are erroneous and interpretation of chemical features and subsequent data analysis is biased.

In the resonant Mie correction algorithm the reference spectrum gradually approaches the true pure absorbance spectrum. It is important to note that there is a fundamental difference between the EMSC models obtained in every iteration step and the EMSC models obtained in for example a basic EMSC model. In a basic EMSC model, scattering and absorption is considered independent, and chemical differences between the reference spectrum and the pure absorbance spectrum are therefore collected in the residuals. The residuals can thus be used to express chemical variations in a complete dataset. In the resonant Mie EMSC model this is different: the chemical differences between the reference spectrum and the pure absorbance are expected to decrease in each iteration step. Each iteration step gradually changes the reference spectrum and make it more similar to the pure absorbance spectrum underlying the measured spectrum. In the following, we demonstrate that in the improved algorithm this convergence towards the true underlying absorbance spectrum actually takes place and that the iterative correction process is not biased by the reference spectrum.

4.3.1 Reference spectrum with altered O-H stretching region

A modified version of the Matrigel spectrum was generated by lowering the absorption in the O-H stretching region, to the left of the peak at 3,300 cm^{-1}. This was done, since the characteristics of the O-H stretching region of the Matrigel spectrum was adapted by the corrected spectrum in earlier versions of the resonant Mie correction algorithm, as seen in Fig. 4.10 c). Therefore, we wanted to modify this region, in order to check if the corrected spectrum adapted the shape of the reference spectrum, or if the shape of the true absorbance spectrum

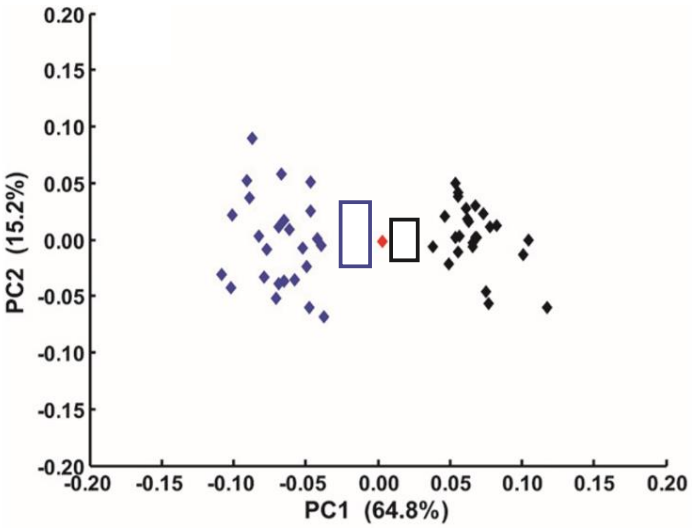


FIGURE 4.9: A score plot of the first and second loading from a PCA on a simulated set of pure absorbance spectra, used for validating the Mie correction algorithm of 2010 [5]. The blue and black diamonds corresponds to two chemically different groups of pure absorbance spectra. The boxes are added to indicate where the corrected spectra are located when they are projected into the score plot. Reproduced with permission from ref. [5]

was adapted. The modified version of the Matrigel spectrum is shown in black in Fig. 4.11. This spectrum was used as the initial reference spectrum for the correction of the simulated apparent absorbance spectra based on the pure absorbance spectrum from group A. The corrected spectra are shown in orange in Fig. 4.11. The underlying pure absorbance spectrum is shown in red, and was chosen from group A. As is evident from this figure, the corrected spectra do not adapt to the chemical features of the reference spectrum in the O-H stretching region. For two of the spectra, spectrum no. 21 and 34, the correction was not successful, and these spectra are not shown. The final $RMSE$ value, which was much higher than for the rest of the spectra, showed clearly that the forward model did not predict the apparent absorbance spectra very well. In order to successfully correct these spectra, a higher number of loadings should be included in the Mie EMSC model. We will later discuss how the final $RMSE$ can be used for assessing if a correction was successful or not.

4.3.2 Reference spectrum from another group

The simulated apparent absorbance spectra with an underlying pure absorbance spectrum from group A were in addition corrected with an initial reference spectrum selected from group B. In the fingerprint region, the chemical differences between group A and B are more prominent than between group A and the Matrigel spectrum, as seen in Fig. 4.6 a). The corrected spectra are shown in orange in Fig. 4.12, while the underlying pure absorbance spectrum and the initial reference spectrum are plotted in red and black, respectively. By visual inspection, it is evident that the correction retrieves the true pure absorbance spectra. The correction failed for one spectrum, i.e. spectrum no. 34, which is not shown in the figure. This was evident from the relatively high $RMSE$ value, and also relates to the number of loadings in the Mie EMSC model.

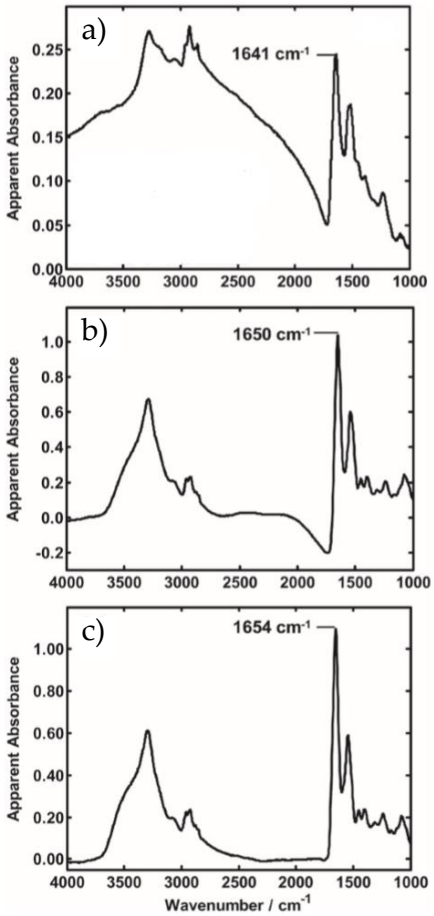


FIGURE 4.10: Correction of an IR spectrum obtained from a humane prostate cancer cell. a) Measured absorbance spectrum. b) Corrected spectrum by using the non-resonant Mie EMSC model of 2008. c) Corrected spectrum using the resonant Mie correction algorithm of 2010 . From [5]

The results presented in this section have shown that corrected spectra do not assume chemical features of the reference spectrum.

4.4 Ability to retrieve the true amide I peak position

As mentioned in section 2.1.5, Mie scattering can shift peak positions towards lower wavenumbers. In infrared spectra, this effect has been observed to affect the amide I peak position in particular. The amide I peak position and band shape carry important information about the secondary structure of proteins, and can be essential for the classification of e.g. healthy and deceased tissue [39]. The retrieval of a reliable peak position is therefore required. In ref. [5], Bassan et al. demonstrate that a reliable peak position for the amide I band can be retrieved. However, this is not shown for the algorithm developed by Konevskikh et al. [28]. In addition, it is essential to investigate whether the retrieval of the amide I peak position is a feature of the Mie EMSC model, or simply a consequence of adaption to the corresponding peak position in the reference spectrum.

For the experimentally obtained lung cancer cell spectra used in this study, the raw spectra do not show a strong shift in the amide I peak position. The peak positions are shifted from on average 1,649.8 $\pm$ 2.1 cm^{-1} in the raw spectra, to 1,651.8 $\pm$ 1.7 cm^{-1}, when the spectra are

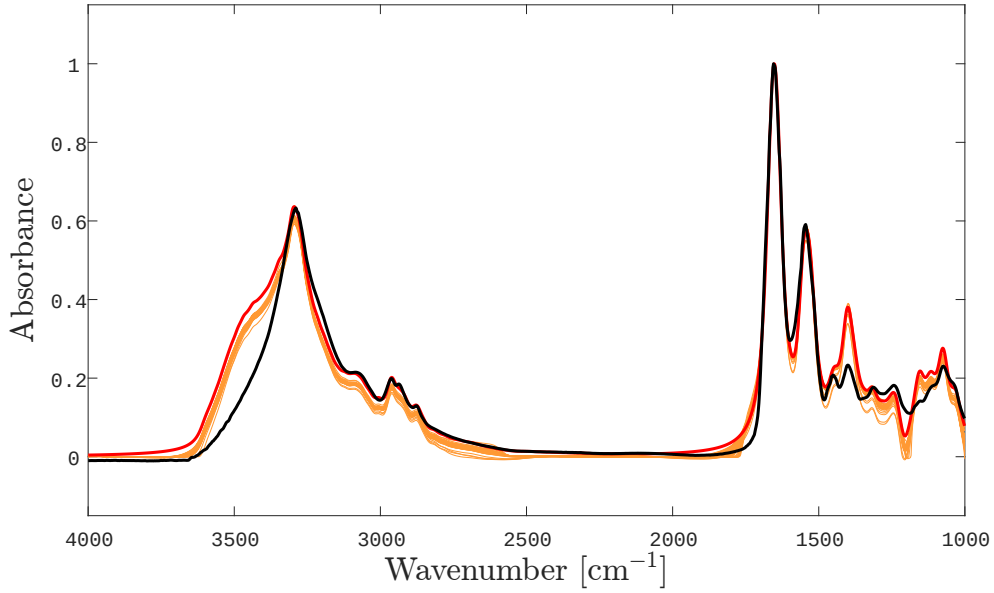


FIGURE 4.11: The simulated apparent absorbance spectra with an underlying pure absorbance spectrum from group A are corrected with a modified version of the Matrigel spectrum as the initial reference spectrum, where the absorbance has been lowered in the O-H stretching region.

corrected with the resonant Mie scatter algorithm. Accordingly, in the simulations of the apparent absorbance spectra, which are based on these measured spectra, the shifts in the amide I peak positions are rather small. In the pure absorbance spectra in group A, the amide I peak is located at 1,653.0 cm^{-1}. In the simulated apparent absorbance spectra based on the pure absorbance spectrum from group A, the peak was shifted to on average 1,649.5 $\pm$ 1.6 cm^{-1}. The correction brought the peak positions back to 1,652.7 $\pm$ 0.3 cm^{-1}. The algorithm is shown to retrieve a reliable peak position, based on the correction of the simulated apparent absorbance spectra. However, corrections should also be evaluated for spectra where the shift in the amide I peak position is strong.

Correcting simulated spectra where the shift in the amide I peak position is significant

In order to study the retrieval of the amide I peak position in spectra where the shift is significant, 8 experimentally obtained spectra from breast cancer cells were used as templates for scatter contributions in simulations of apparent absorbance spectra, provided by Nick Stone and his group at the University of Exeter. The breast cancer cells had been incubated in an osteogenic mix containing ascorbic acid, B-glycerophosphate and dexamethasone, for three days. Spectra were recorded with a FTIR Agilent microscope, with the spectral range of 3,900-900 cm^{-1} and a spectral resolution of 4 cm^{-1}. Simulations were done according to the method described in section 3.2, and a simulated pure absorbance spectrum from group A was used as the underlying Z_{pure}.

In the measured breast cancer cell spectra, the amide I peak positions are located at on average 1,635.3 $\pm$ 4.8 cm^{-1}. In the correction, the peaks were moved to 1,651.0 $\pm$ 1.2 cm^{-1}. Figure 4.13 shows an example of a simulated apparent absorbance spectrum with scattering features based on this data set. The underlying pure absorbance spectrum is shown in Fig. 4.13 a) together with the Matrigel spectrum, which was used as a template for the simulations. The amide I peak in the pure absorbance spectrum is located at 1,653.0 cm^{-1}. In Fig. 4.13

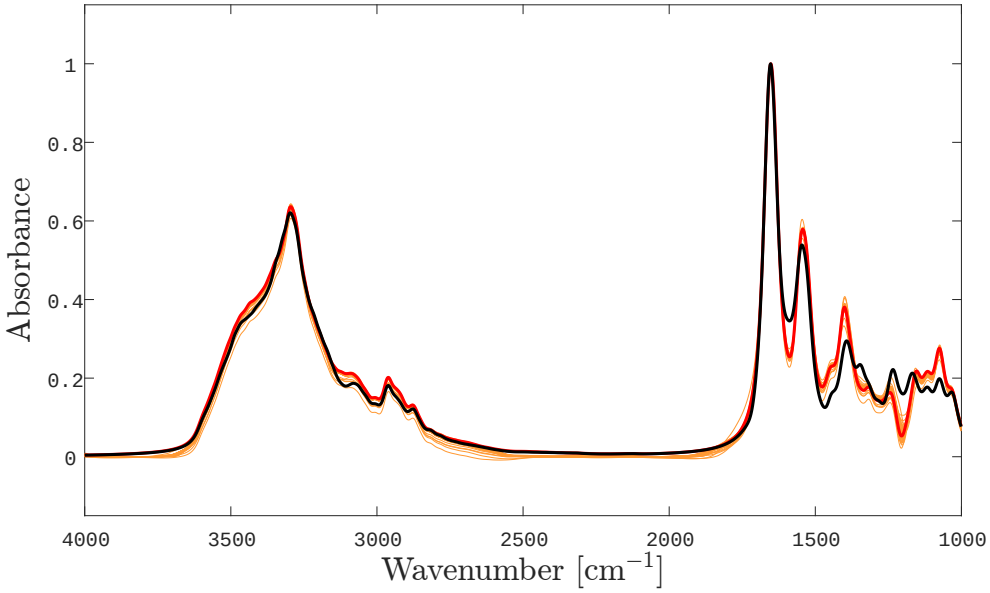


FIGURE 4.12: The simulated apparent absorbance spectra with an underlying pure absorbance spectrum from group A are corrected with a pure absorbance spectrum from group B as the initial reference spectrum.

b), the simulated apparent absorbance spectrum is shown with the experimentally obtained spectrum, which is used for obtaining the scattering features. In the simulated apparent absorbance spectra, the amide I peak positions were shifted to on average 1,643.0 $\pm$ 1.9 cm^{-1}. The correction moved the peak position back to 1,650.3 cm^{-1}, which is shown in the example in Fig. 4.13 c).

In this section, it was demonstrated that the algorithm retrieves a more reliable amide I peak position, even when the shift in the peak position is significant. As the correction is shown to be less dependent on the initial reference spectrum, this is taken as a strong indication that the retrieval of the true peak position is a feature of the model.

Reference spectrum with a dislocated amide I peak position

In order to verify that the retrieval of the amide I peak position is not a consequence of the corrected spectra adapting to the corresponding peak position in the reference spectrum, the Matrigel spectrum was modified by dislocating the amide I peak position.

The Matrigel spectrum is shown together with the the modified Matrigel spectrum in Fig. 4.14 a). The amide I peak was shifted towards higher wavenumbers, from 1,655 cm^{-1} to 1,674 cm^{-1}, which can be seen in the figure. The modified Matrigel spectrum was used as the initial reference spectrum in the correction of the simulated dataset, which is based on the pure absorbance spectrum from group A, and with scatter contributions from the measured lung cancer cell spectra [24]. The corrected spectra are shown in Fig. 4.14 b) in orange. The underlying pure absorbance spectrum is shown in red, and the modified Matrigel spectrum in black. It is evident that the true amide I peak position is retrieved, and the corrected spectra do not adapt to the peak position in the initial reference spectrum. This proves further, that the algorithm is stable towards variations in the reference spectrum used for initializing the algorithm. The correction failed for one spectrum, i.e. spectrum no. 34, which is not shown in the figure. This was evident from the relatively high $RMSE$ value, and is related to the number of loadings which should be higher for this spectrum.

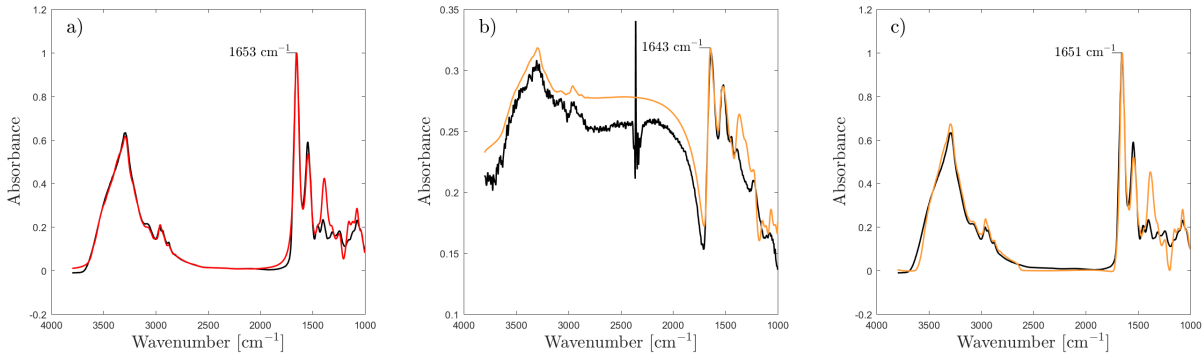


FIGURE 4.13: a) The amide I peak position is located at 1,655 cm^{-1} in the pure absorbance spectrum, showed in red. The Matrigel spectrum is plotted in black. b) A simulated apparent absorbance spectrum in orange, based on the pure absorbance spectrum in a). Scattering features are obtained from a measured spectrum from breast cancer cells, shown in black. The amide I peak position is shifted to 1,643 cm^{-1}. c) The corrected spectrum is shown in orange, and the Matrigel is shown in black. The amide I is shifted to 1,651 cm^{-1} in the correction.

4.5 Sensitivity towards initialization parameters

The optimal parameters for initializing the algorithm depends on the data set. In the following, a description of how a manual parameter optimization may be performed is given. Therefore, this section may serve as a user manual for the Mie correction code. In general, as the algorithm is currently set up, the parameter optimization needs to be done manually for each new data set. If the data set under investigation is large, it is expedient to use a smaller subset for the parameter adjustments. Later we will make a suggestion as to how the parameter ranges in future may be set automatically, when for example working with large data sets such as imaging data. The following sections will be related to a concrete example, namely the correction of the lung cancer cell spectra. For the parameter estimation, 7 spectra are chosen from the set of 60 spectra. When adjusting the initial parameters, neither weighting nor the guaranteed positivity of the reference spectrum should be applied. By applying these functionalities, the effect of adjusting the parameters may be diminished, as will be illustrated.

Setting a, n_0 and h

For most of the data sets in this study, the default parameter ranges for a, n_0 and h resulted in a stable correction. However, this is not the general case since morphology of the samples under consideration may differ considerably. We observed that by choosing unsuitable parameter ranges for the Mie meta-model, the model is not able to estimate the Mie scattering signatures in the apparent absorbance spectra. In general, this causes the correction to fail, and the corrected spectra do not resemble typical pure absorbance spectra. Unfit parameter ranges rather results in artifacts in the spectra. When the iterative algorithm results in a corrected spectrum with a relatively high $RMSE$ and the corrected spectrum contains artifacts, the corrected spectrum does not predict the apparent absorbance spectrum well when it is used as input for the Mie model. Therefore, the $RMSE$ is a good indicator for a successful or unsuccessful correction. A simple quality test based on the $RMSE$ for sorting out for which spectra the correction has gone wrong is presented later. Within the parameter ranges that do fit the data set, the correction is observed to be stable. It is not observed that the correction retrieves incorrect chemical features from the apparent absorbance spectra, and the correction rather fails than giving deceptive results.

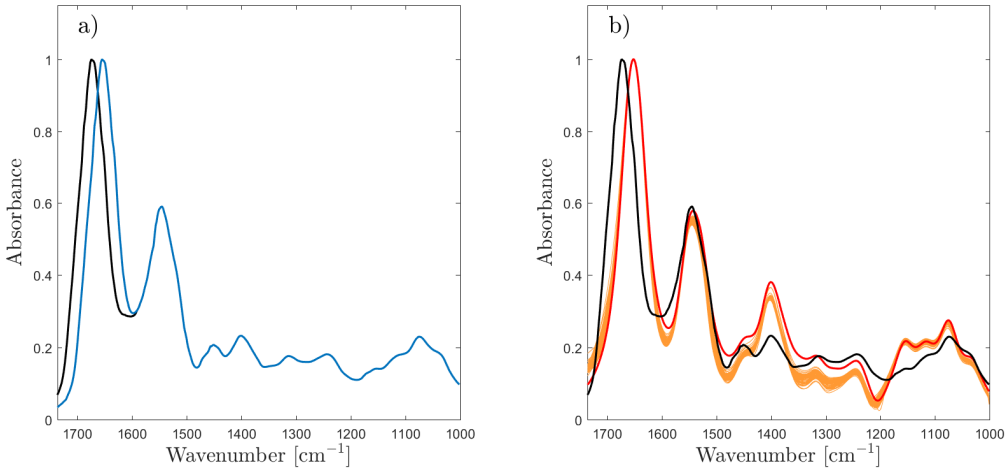


FIGURE 4.14: a) The Matrigel spectrum is plotted in blue with a modified version, where the amide I peak position is moved from 1,655 cm^{-1} to 1,674 cm^{-1}, is plotted in black. b) The simulated apparent absorbance spectra with an underlying pure absorbance spectrum from group A are corrected with a modified Matrigel spectrum. The corrected spectra still retrieves a reliable amide I peak position.

In order to test the stability of the algorithm with respect to the initialization parameters a, n_0 and h, corrections were performed with different parameter settings. The corrections were performed on the simulated apparent absorbance spectra which were based on the lung cancer cell spectra [24]. In Fig. 4.15 the corrected spectra are shown. As previously, the red and dark blue spectrum is the pure absorbance spectrum from group A and B, respectively, and orange and light blue is the corresponding corrected spectra. The Matrigel spectrum is shown in black. The parameters for the correction of spectra from group A is set to the upper limit of what resulted in a stable correction; $a \in [3, 8.1]$, $n_0 \in [1.16, 1.46]$ and $h = 0.40$. For group B the parameters were set to the lower limit, with $a \in [1.6, 6.7]$, $n_0 \in [1.08, 1.38]$ and $h = 0.20$. It is evident from Fig. 4.15 a) that artifacts has been introduced in the baseline of two of the corrected spectra from group B. When extending the parameter range beyond a certain limit, such artifacts may be introduced. However, this does in general not pose a problem as it is clear from the final $RMSE$ value and by visual inspection whether the parameter range is suitable or not. As mentioned, the correction does not retrieve wrong chemical features. As is shown in Fig. 4.15 b), the corrected spectra can still easily be classified by PCA. Figure 4.15 b) shows a projection of the corrected spectra into the score plot of the pure absorbance spectra in Fig. 4.6 b), and it is evident that the spectra still show the correct grouping.

In section 4.6, a description on how to handle data sets, which consists of samples that require different parameter ranges, is given. When the sample set is very homogeneous, i.e. for example spectra of the same type of lung cancer cells from a cell line that were treated with the same condition, it is expected that the same parameter ranges may be sufficient. If spectra are obtained by infrared spectroscopy of images, the situation may be different, since the tissue may display different morphological features, when large tissue regions are considered.

Setting the number of loadings A_{comp}

The number of loadings included in the Mie EMSC model affects how precisely the Mie oscillations are represented. In the Mie correction code, the number of loadings is set through the desired level of explained variance in the Mie extiction curves. With the data sets at hand

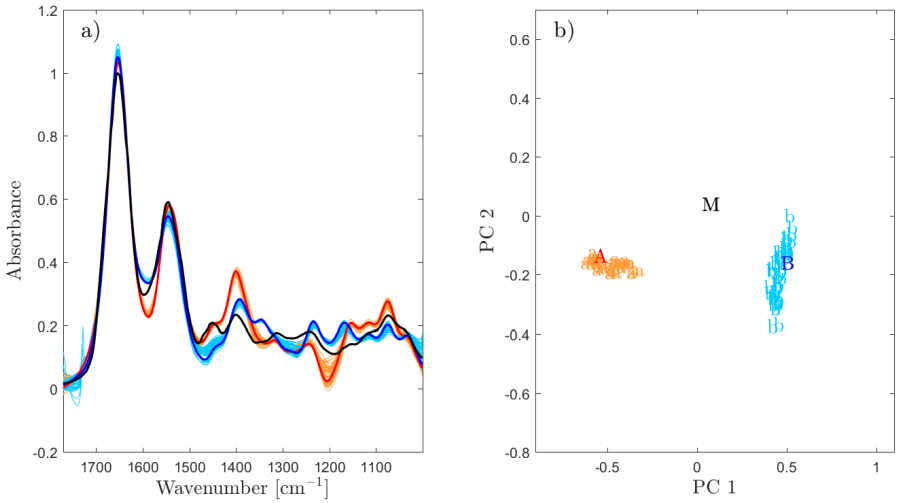


FIGURE 4.15: Corrected apparent absorbance spectra corresponding to the red pure absorbance spectrum in group A (in orange), and to the blue pure absorbance spectrum from group B (in light blue). In the correction, the parameters were changed to $a \in [3, 8.1]$, $n_0 \in [1.16, 1.46]$ for group A, and $a \in [1.6, 6.7]$, $n_0 \in [1.08, 1.38]$ and $h = 0.20$ for group B. b) The corrected spectra are projected into the score plot of the pure absprbance spectra in Fig. 4.6 b).

for this study, a level of explained variance at 99.96 - 99.99 % has shown to yield a stable and precise correction. The optimal level of explained variance is determined by increasing the limit until there is no significant change in the result of the correction.

In Fig. 4.16 a), the example data set is corrected with a level of explained variance at 99.96 %. With the default settings for α_0 and γ, this results in 7 loadings. Here the reference spectrum is not weighted and negative parts of the reference spectrum are not set to zero. As the correction shows, the Mie oscillations are not completely modelled in Fig. 4.16 a). In order to correct the remaining Mie oscillations, the limit of explained variance is increased to 99.99 %, resulting in 9 loadings. The correction is shown in Fig. 4.16 b), and it is evident that the Mie oscillations are more precisely modelled.

By setting the negative parts of the reference spectrum to zero in the EMSC parameter estimation, or by weighting the reference spectrum, the effect of increasing the level of explained variance would not be as evident. This is shown in Fig. 4.16 c), where 7 loadings were used. The applied weight function is plotted in red. In this case the weighting reduces the Mie oscillations, which could not be corrected above where weighting of the reference spectrum was not applied.

Setting the weight function

Weighting of the reference spectrum leads to a stable baseline correction with less disturbance in the chemically inactive regions. Firstly, the chemically active regions should be specified. A smooth transition between the chemically active and inactive regions can be achieved by adjusting the slopes of the tangent function. For the example data set, the chemically active regions are set to 1,000-1,750 cm^{-1} and 2,550-3,700 cm^{-1} in Eq. 4.1. The slope is determined by κ and is set to 1 for each inflection point. Figure 4.16 d) shows the result of weighting the reference spectrum when correcting example spectra. After suitable parameters for the weight function are found, negative parts of the reference spectrum should be set to zero.

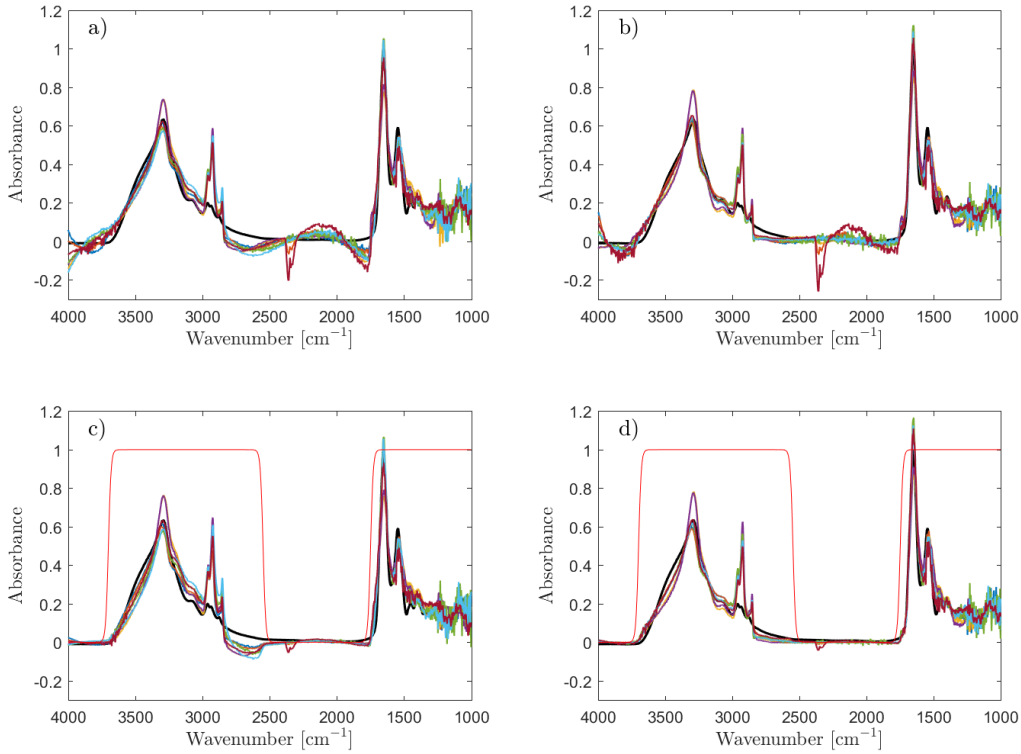


FIGURE 4.16: Illustration of the effect of changing the number of loadings included in the Mie EMSC model. a) Five spectra from the lung cancer cell data set [24] are corrected with $A_{comp} = 7$. The Mie oscillations are not accurately predicted. b) By changing the the number of loadings to $A_{comp} = 9$, the Mie oscillations are more precisely reproduced. c) When applying a weight function (shown in red) the effect of choosing parameters that are not optimal are reduced. d) Weighting should be applied when the optimal parameters are found.

We emphasize the importance of using a baseline corrected reference spectrum when weighting of the reference spectrum is applied. If the reference spectrum has a non-zero baseline, the weighting would result in a strong deformation of the corrected spectra.

Correcting the whole data set

When the initialization parameters are set, the whole data set can be corrected. All the 60 corrected spectra are shown in Fig. 4.17. If the correction fails for some of the spectra in the data set, it is usually evident from the final $RMSE$ value. In general, the final $RMSE$ values for spectra which could not be corrected, are significantly higher than for the successfully corrected spectra. A simple quality test based on the final $RMSE$ is implemented in the Mie correction code. By visual inspection of the final $RMSE$, an upper limit for the $RMSE$ is set. Corrections with a higher $RMSE$ than the upper limit is discarded. How this quality test works in practice is illustrated in the following section.

4.6 Correcting spectra from imaging data

A Mie correction code which can be used on IR imaging data, is highly desired by the infrared biomedical community. FTIR imaging has been developed to be used as a diagnostic tool for

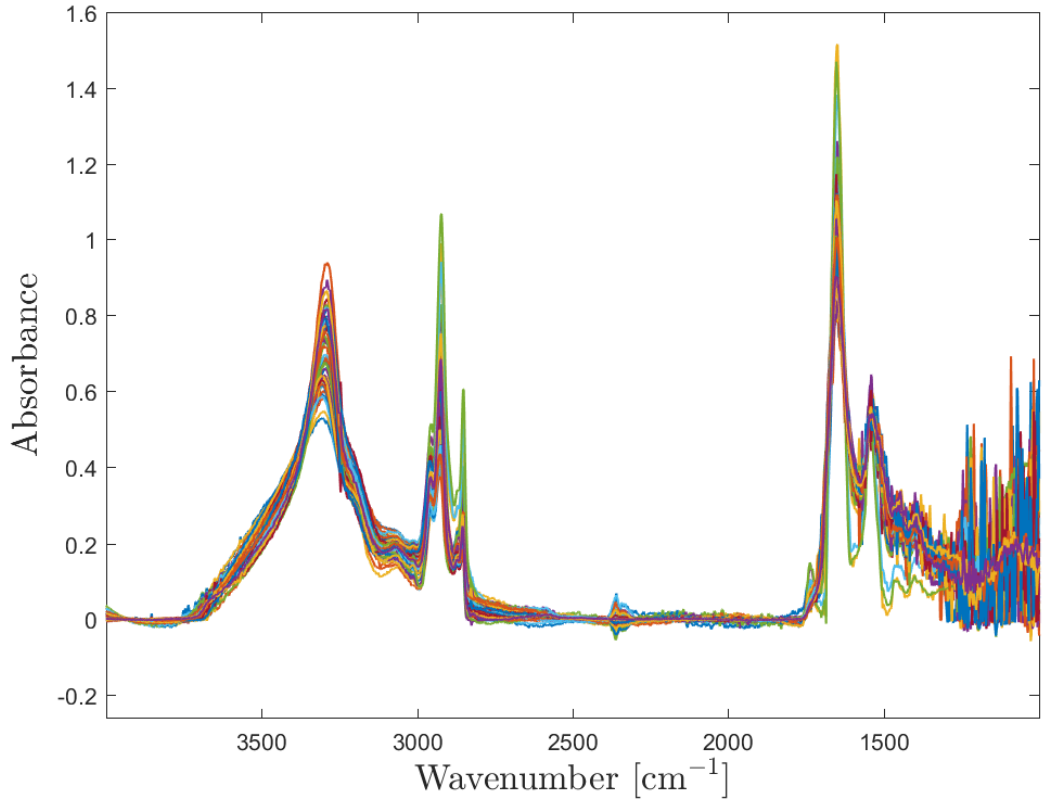


FIGURE 4.17: Correcting of the whole data set consisting of 60 spectra obtained from lung cancer cells [24], after the initialization parameters are determined.

cancer diagnosis, and is currently on its way to clinical practice. Earlier versions of the Mie correction code [5] are currently used for correcting imaging data. The use of the Mie correction code on infrared imaging data, requires adjustments to the code with respect to quality testing, in order to make an effective distinction between low quality spectra and sample spectra, prior to correction. In addition, the choice of parameter should be optimized and automatized in order to handle data sets consisting of different types of samples. As will be demonstrated below, an IR data set may contain spectra obtained from samples which differ in morphology in addition to chemical content. Parameter settings should therefore be adjusted in order to obtain optimal correction of the different groups. Adjustments for handling infrared imaging data are outside the scope of this thesis. However we provide an example where the Mie correction code is used on spectra extracted from an FTIR imaging data set. This example illustrates how parameters can be set manually for a data set consisting of samples which differ in morphology, and how the final $RMSE$ can be used as a tool for quality testing and parameter estimation.

In total, 147 spectra were extracted from an FTIR imaging data set obtained from sample of a thin slice of colon tissue. The data was obtained from Dr. Dennis Petersen in the group of Prof. Gerwert at the Ruhr University in Bochum. The data set consists of measurements from three different groups; submucosa (connective tissue), muscle tissue and lumen of crypts. In the following, the groups are denoted A, B and C, respectively. The measured absorbance spectra are shown in Fig. 4.18 a), where group A is plotted in red, group B in blue and group C in green. These color codes are kept throughout the following sections. From the raw spectra it is evident that the scattering signatures differs significantly between the groups.

By assuming no *a priori* knowledge about the group affiliation of the spectra prior to correcting the data, the whole data set is first corrected with the default parameter settings. The corrected spectra are shown in Fig. 4.18 b), and it is evident that for some spectra, the parameter settings were not suitable. By inspecting the final $RMSE$ in Fig. 4.18 c), it is clear that the correction fails for a group, i.e. group C which includes spectra no. 98-147. We established a quality control by setting the upper limit of the final $RMSE$ to 1.0×10^{-2}. Most of the spectra from group C did not pass this quality test. By visual inspection of Fig. 4.18 d) we see that for the spectra in group C which passes the quality test, the baseline correction in the region 1,750-2800,0 cm^{-1} is not optimal. The fact that the baseline could not be corrected is not evident from the final $RMSE$. While a separation could be made based on chemical differences between the groups, which are prominent in the fingerprint region, the $RMSE$ does not show that the correction of the baseline in this group was less successful than in other groups. The problem might be addressed by establishing an $RMSE$ for different spectral regions. Currently the $RMSE$ is calculated for the whole spectrum and contains model errors with respect to both chemical and scatter estimation. When an $RMSE$ is defined for chemically active and chemically inactive regions separately, we might be able to detect whether the scatter correction is successful or not. In order to automatize the process of parameter estimation, this problem has to be investigated further.

Correcting group C
The relatively high final $RMSE$ in the correction of group C indicates that the parameter ranges should be adjusted. We observed that by changing from the default parameter range for the radius a to $a \in [3.2\,\mu\text{m}, 8.3\,\mu\text{m}]$, a significantly better correction could be obtained, as illustrated in Fig. 4.19 a) and b). However, for some of the spectra, the baseline still contains broad oscillations, which indicates that the level of explained variance in the Mie extinction curves should be increased. By increasing the level of explained variance to 99.997 %, resulting in 10 loadings, most spectra in group C could be corrected. The correction is illustrated in Fig. 4.19 a) and b). The correction for some of the spectra was still not optimal. When inspecting the raw spectra visually, no reason for why the correction was not optimal could be found. This issue has to be investigated further and might be addressed by implementing an automatic optimization of the model parameters, such as the total number of loadings used in the Mie EMSC model, or by optimizing the parameter ranges.

We can conclude that the imaging data set could be corrected with the Mie correction code. However, it was necessary to adjust the parameter ranges for the different groups in order to account for different scattering contributions in the data set. The corrected spectra are shown in Fig. 4.20 b), while the final $RMSE$ values for all spectra are shown in Fig. 4.20 a). In order to adapt the Mie correction code such that it can handle infrared imaging data, or data sets consisting of different types of samples, the final $RMSE$ can be used to determine whether the parameter settings are suitable or not. Further investigation is needed to automate and increase the precision of this quality test. As mentioned above, the use of different $RMSE$s calculated for different spectral regions (chemically active and inactive regions) could be a possible solution.

4.7 Separating and investigating scatter and chemical information

After correcting the FTIR image spectra of the colon tissue from the previous section, biochemical differences can now be explored. In the following we demonstrate that when scattering features are removed from the spectra, separation of the different groups can be obtained by a PCA analysis.

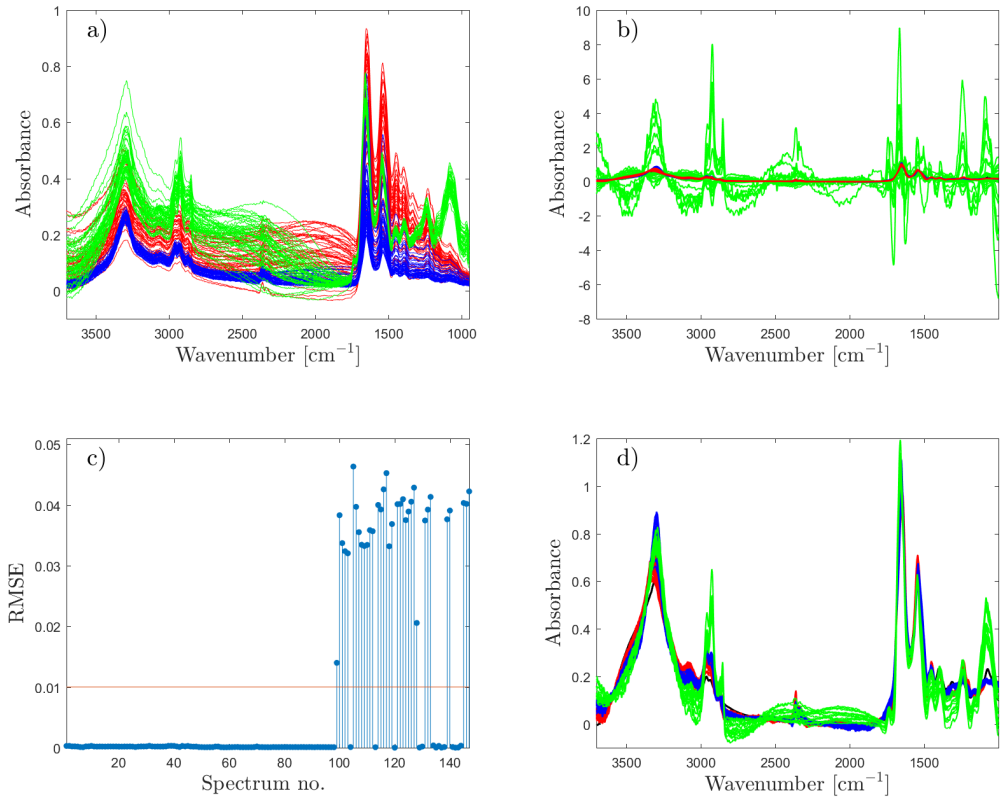


FIGURE 4.18: Correcting spectra from an infrared image data set obtained from a colon tissue sample. a) Raw spectra from three groups, submucosa (group A, red), muscle tissue (group B, blue) and lumen of crypts (group C, green). b) Corrected spectra by use of the default initialization parameters. c) The final $RMSE$ for each spectrum, and the upper limit is shown in red. Most of the spectra from group C do not pass this quality test. d) The spectra with a lower $RMSE$ than the upper limit.

From the raw spectra in Fig. 4.18 a), it is evident that scattering signatures vary between the different groups. This indicates differences in morphology between the samples. We have already seen that this has consequences for the parameter estimation. In the following section we demonstrate that the parameters obtained from scattering signatures are sufficient to separate the different groups. Since both scatter parameters and chemical signatures contain information that can be used for separating the different groups, the question may arise, if the removal of the scatter information may lead to data that has a lower discrimination ability or not. The machine learning community has the philosophy to work on raw data in order to prevent removal of important information from the data. This has in many cases been very successful, while it requires large amounts of calibration data. This is often not available, since it is very expensive. Furthermore, the pre-processing employed in this thesis separates the scatter information and the chemical information, which both are available after pre-processing. We discuss these issues in the following by considering the results of the separation of the groups based on PCA on the EMSC parameters, and when comparing this with a PCA on the chemical absorbance.

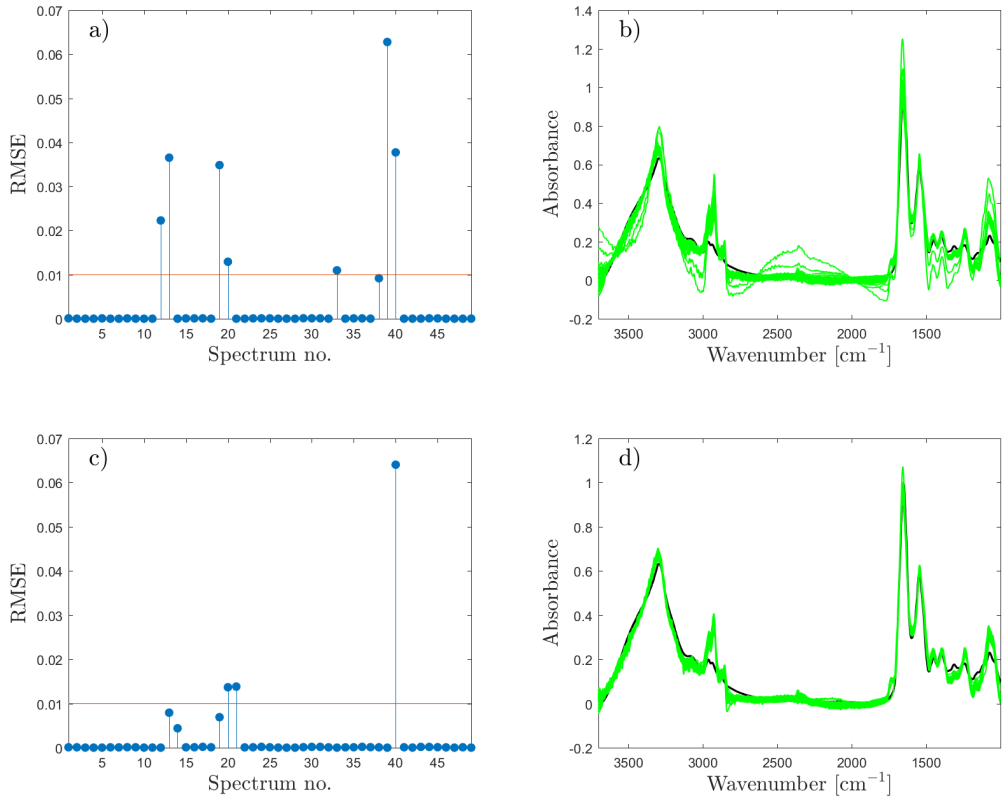


FIGURE 4.19: Correcting group C separately. a) The final $RMSE$ after adjusting the radius to $a \in [3.2\,\mu\text{m}, 8.3\,\mu\text{m}]$. b) Spectra which passes the quality test in a). The Mie oscillations are not removed completely. c) By setting the explained variance in the Mie extinction curves to 99.997 %, resulting in 10 loadings, the oscillations are more precisely represented. d) The spectra which passes the quality test in c).

4.7.1 PCA on raw spectra

The raw spectra of the FTIR image data of the colon tissue are plotted in Fig. 4.18 a). Both morphological and chemical differences between the groups are visible, and these differences are most prominent in the region between 1,000 - 3,000 cm^{-1}. By performing a PCA on this spectral region, the first and second principal components allow separating group C from group A and B, as shown in the score plot in Fig. 4.21 a). It is however not possible to separate group A and B from each other, not even when investigating the other principal components as illustrated in the score plot of the second and third principal component in Fig. 4.21 b). This is expected, as both scatter and chemical signatures are more distinctive for group C than for group A and B. From Fig. 4.21 c) and d) it is evident that the first loading describes a multiplicative effect, while the second loading describes mainly Mie scattering features. By performing a PCA on group A and B only, the two groups do still not separate. This is illustrated by the score plot in Fig. 4.21 e).

4.7.2 PCA on corrected spectra

After correcting the raw spectra of the FTIR image data of the colon tissue with the Mie correction algorithm as described in section 4.6, scattering features are removed from the absorbance

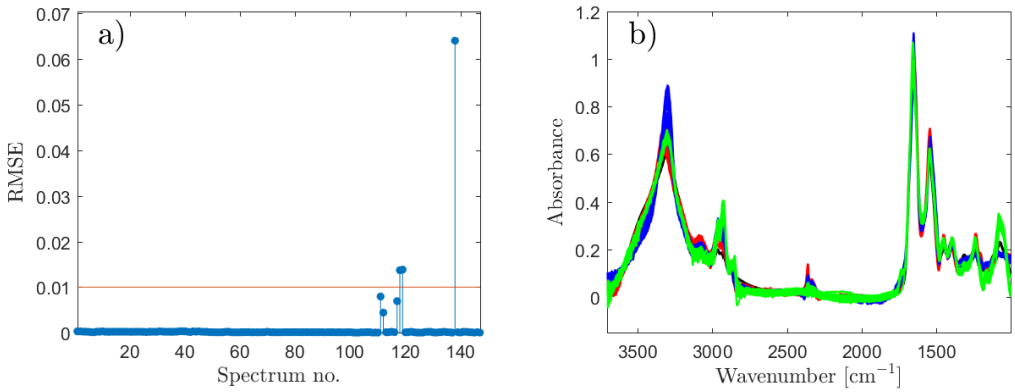


FIGURE 4.20: Correcting all spectra in the colon tissue data set, with optimized parameter settings. a) The final $RMSE$ for all spectra. b) A successful correction could be obtained by adjusting the initialization parameters.

spectra. When investigating the score plot of a PCA of the set of corrected spectra, we observe that group C can be clearly distinguished from group A and B, as shown in Fig. 4.22 a). By investigating the third principal component, group A and B are easily separated. This is illustrated in the score plot of the second and third principal component in 4.22 b). The first and second loading expresses almost exclusively chemical features, as seen in Fig. 4.22 c) and d). In addition, a PCA performed on group A and B only, allows separation of the two groups. This is shown in Fig. 4.22 e). It is evident that by removing scattering features from the spectra, a simple cluster analysis is sufficient to separate the three groups.

4.7.3 PCA on the Mie EMSC parameters

As mentioned previously, scattering features that are removed from the raw spectra are not lost after the correction. This offers a great opportunity, since the Mie EMSC parameters from Eq. 2.41 which relates to scattering features can be investigated separately: Through the EMSC modelling, scattering features are parameterized, and are therefore accessible for interpretation. This has been demonstrated for FTIR imaging data in ref. [25], where basic EMSC parameters have been plotted as parameter images and EMSC parameters have been used to separate between different experimental groups. In the following, the parameters from the Mie EMSC are used to separate the groups by PCA.

When PCA has been previously performed on EMSC parameters, the EMSC parameters from the spectra investigated in the same PCA analysis all referred to the same EMSC model functions. This is not the case for the resonant Mie EMSC model, as the model parameters p_i from Eq. 2.41 are different for each spectrum. The reason for this is that the imaginary part of the refractive index, which is used as input for the Mie extinction efficiencies in Eq. 2.33, contains chemical information that is different from spectrum to spectrum, and that is changing in each iteration step. Therefore, the parameters g_i from Eq. 2.41 do not refer to exactly the same feature in each spectrum, since the p_i in Eq. 2.41 are slightly different from spectrum to spectrum. However, the main features of the scattering components p_i are determined through the parameter ranges of α_0 and γ, and not the chemical information. In Figure 4.23 a)-c) the first three loadings of the Mie EMSC model from Eq. 2.41 are shown for one spectrum of each group (A, B, C) of the FTIR image data of the colon tissue. We observe that the loadings are different for each spectrum. However, it is obvious that the loadings which refers to the same component share common scattering features. The differences between the loadings are due to chemical chemical features that were introduced through the imaginary part of the refractive index n'. For the spectra in group C, the parameter range for a was changed as well. It

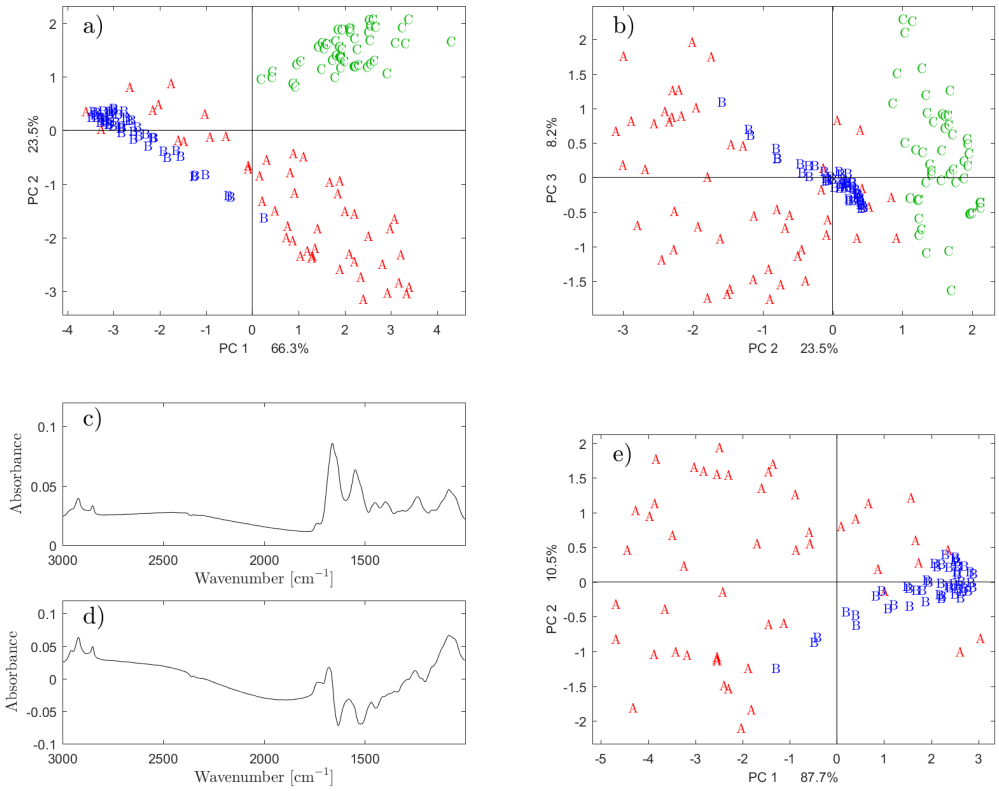


FIGURE 4.21: Performing a PCA on the raw colon tissue spectra shown in Fig. 4.18 a). a) In a score plot of the two first principal components, group C can be separated from group A and B. b) Group A and B is not possible to separate, illustrated by a score plot of the second and third principal component. c) and d) shows the first and second loading, respectively. It is evident that the two first loadings discriminates on scattering features. c) Separation is not achieved in a separate PCA on only group A and B.

is evident that this relatively small change in the parameter range has an insignificant impact on the first three loadings obtained on the set of extinction curves. Since the loadings that were used as EMSC model functions for each spectrum share the same scattering features, it is meaningful to perform a PCA on the Mie EMSC parameters estimated in the EMSC model for these loadings. These model functions differ as mentioned in their chemical features from spectrum to spectrum, but they share scattering features, which can be parameterized by the Mie EMSC model and be used for separating samples into groups by PCA.

In order to put all variables on the same footing, the EMSC parameters were mean-centered column-wise and normalized by dividing each column by its standard deviation. The normalized Mie EMSC parameters were then analysed by PCA. Fig. 4.24 a) shows that group C can be separated from group A and B by the first two principal components.

Further separation of group A and B could not be obtained, as illustrated by the score plot of the second and third principal component in Fig. 4.24 b), and a separate PCA on group A and B only, shown in Fig. 4.24 c). Since the scattering features are causing the main spectral variation in the raw spectra shown in Fig. 4.18 a), we assume that a PCA performed on the raw spectra discriminates on the same features as a PCA shown in Fig. 4.24 a) performed on the EMSC parameters. This is also supported by the scattering features visible in the loadings

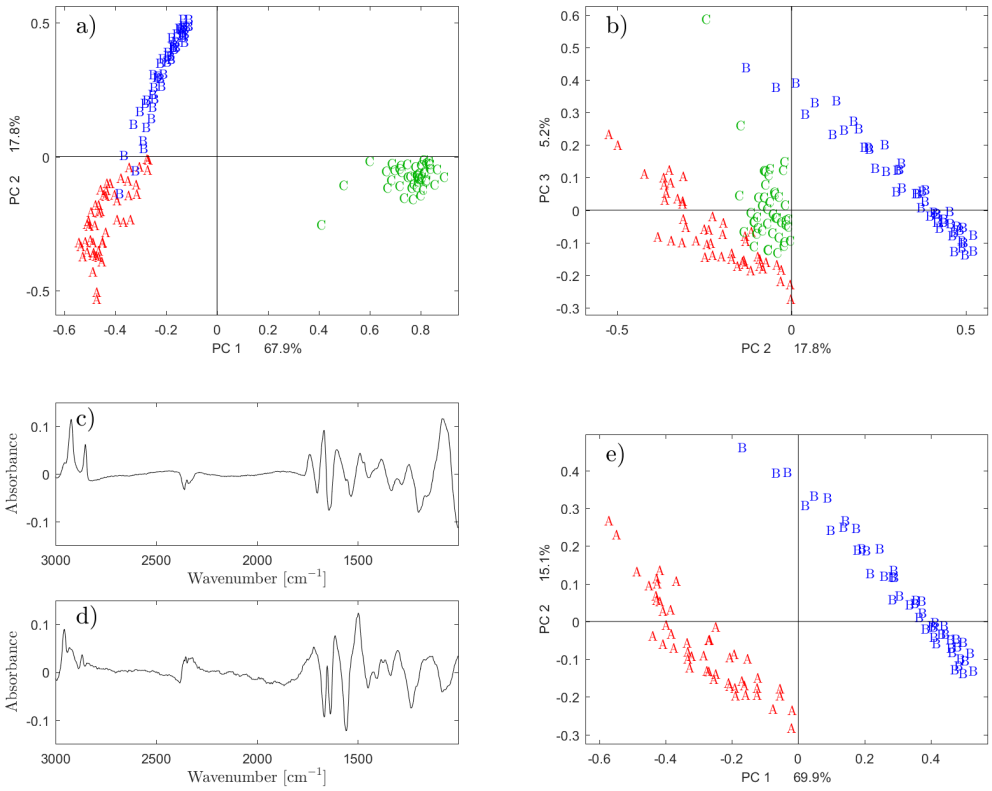


FIGURE 4.22: A PCA on the corrected spectra from the colon tissue sample allows separation of all groups. a) Group C is separated from group A and B by the first two principal components. c) Group A and B is separated by the second and third principal component. Both the first and second loading, shown in c) and d), respectively, express chemical features. e) In a separate PCA on group A and B only, the two groups are easily separated.

in Fig. 4.21 c) and d).

In addition to parameterizing the scattering features in the absorbance spectra, the Mie EMSC model offers a great advantage with respect to reducing complexity in a classification model. As was shown in section 2.1.5, the Mie scattering formalism describes a highly non-linear relation between the apparent absorbance spectrum and the pure absorbance spectrum. In short, the apparent absorbance spectrum is assumed to be proportional to the extinction efficiency Q_{ext}, while the extinction efficiency is a highly non-linear function of the pure absorbance spectrum through the imaginary part of the refractive index n' according to Eq. 2.33. It is important to remember that the pure absorbance spectrum is first used to calculate the imaginary part of the refractive index n' and the fluctuating real part of the refractive index n_{kk} is obtained from the imaginary part of the refractive index by Kramers-Kronig transform according to 2.27.

It is well know that machine learning algorithms are very well suited to solve highly non-linear problems. Thus, it may be possible that well-working classification models could be established on raw data without further pre-processing. Machine learning algorithms have been employed for identification of microorganisms in FTIR spectroscopy since the 1990s [23, 54].

Based on random forest and artificial neural network algorithms, flexible classifiers can

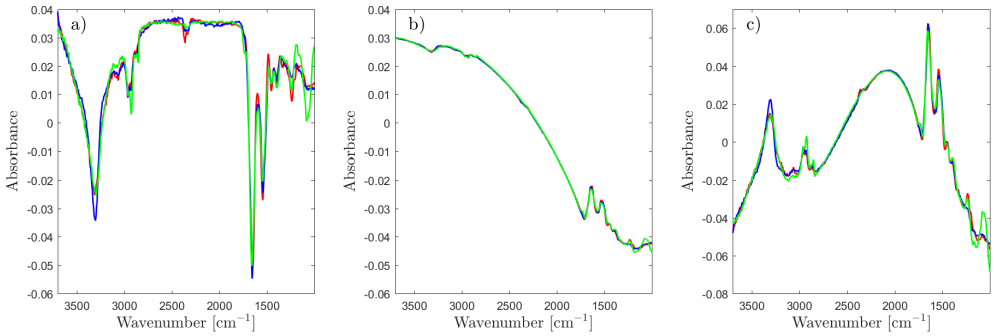


FIGURE 4.23: The first three loadings p_i included in the Mie EMSC model in Eq. 2.41, shown in a)-c), for one spectrum from each group (A: red, B: blue, C: green). Despite of chemical differences between the spectra, and different parameter ranges for α_0 and γ for the Mie extinction curves based on the spectra from group C, the loadings share common scattering features.

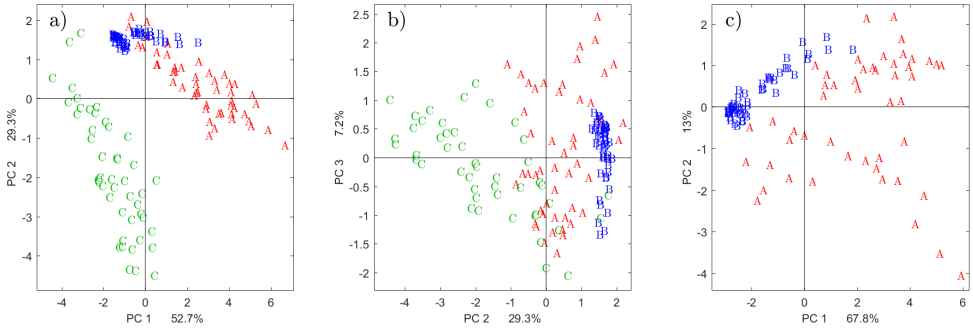


FIGURE 4.24: Score plots of the principal components from a PCA on the Mie EMSC parameters from the correction of the spectra obtained from the colon tissue samples. a) Group C can be separated from group A and B by the first two principal components. b) Further separation of group A and B was not obtained. c) In a separate PCA of only group A and B, the two groups could not be separated.

be trained that may be able to predict classes from measured raw spectra. However, when complicated non-linear relationships are present in the spectra, huge data sets are needed for training in order to resolve the non-linear relations between the apparent absorbance spectrum and the underlying chemical features. With the Mie EMSC model, a combination of electromagnetic theory and some *a priori* knowledge is used to separate the chemical and physical features of the absorbance spectra, and reduce the complexity considerably. Biochemical analysis can subsequently be performed with far less complex models. As shown previously in this section, separation of the chemical groups was obtained with a relatively small data set, by use of both the pure absorbance spectra and the Mie EMSC parameters in a simple PCA model.

Chapter 5

Conclusions and Outlook

5.1 Conclusions and outlook

In this thesis, an algorithm for separating Mie scattering and absorption in infrared absorbance spectra of single cells and tissues [28, 26] has been further developed and as a result, an open source code has been provided to the biomedical infrared spectroscopy community. Infrared microspectroscopy is employed in the biomedical infrared spectroscopy community for establishing new diagnostic tools for cancer diagnosis via infrared spectrocopy and imaging of cells and tissue samples. In the infrared spectroscopy of single cells and tissues, Mie scattering features are often dominating the absorbance spectra and thus create serious problems for subsequent data analysis of spectra and the interpretation of spectral bands.

Currently, the algorithm of Bassan et al. [5] is used by the biomedical community for correcting Mie-scatter distorted spectra of cells and tissues. The presence of scattering features in the measured infrared spectra is considered a serious problem for subsequent classification. It has been reported that the algorithm of Bassan et al. [5] has several problems. A major issue is the fact that spectra corrected by the Bassan algorithm have the tendency to adapt features of the reference spectrum employed. Konevskikh et al. [28, 26] developed the Mie scatter correction algorithm further during recent years. While several important improvements were done, the code was not yet in an user-friendly form that could be provided to the biomedical community. The overall aim of this thesis was to improve and stabilize the exiting Mie scatter algorithm by Konevskikh et al. [26], such that it could be published as an open source code. This has been accomplished and the code is published at https://bitbucket.org/biospecnorway/mie-emsc-code.

An important issue before publishing the code was the validation of the algorithm by a simulated data set where the underlying pure absorbance spectra are known. A data set that consists of apparent absorbance spectra which are distorted by Mie scattering, while at the same time the underlying pure absorbance spectra are known, is in general hard to obtain experimentally. To provide such a data set experimentally, samples have to be prepared from a material resembling a biological cell, and shaped like imperfect spheres or tissue. In practice, this is an impossible task. The fact that biological samples are not perfectly spherical and homogeneous suggests that exact Mie theory, which describes perfect and homogeneous spheres, cannot not be used for simulating apparent absorbance spectra of cells and tissues. Therefore we followed a different route: The simulated data set was obtained by using the Matrigel spectrum as a template for simulating pure absorbance spectra, and scattering features were introduced by mimicking features from experimentally obtained spectra. The scattering features are estimated from real measured spectra, which resulted in apparent absorbance spectra with a variety of scattering signatures, which clearly resembled the scattering features observed in measured data. The simulated data set may serve as a bench mark set for validation of future versions of the Mie EMSC correction and is published together with the source code provided by this thesis. The algorithm developed by Konevskikh et al. [27] and all improvements of

the algorithm made in this thesis were thereafter validated by the set of simulated apparent absorbance spectra.

In order to stabilize the code, and thus make it more user-friendly, the choice of different model parameters needed to be investigated further. These parameters include the number of principal components used in the Mie EMSC model, the stop criterion used for terminating the algorithm and the initialization parameters used to establish the Mie meta-model. In the published code, the number of principal components used in the Mie EMSC model is set automatically, based on a desired level of explained variance in the Mie extinction curves. Furthermore, the criterion for terminating the algorithm has been reviewed, and a flexible stop criterion is implemented based on the convergence of the forward model. A flexible stop criterion enables a stable termination of the iterative algorithm by the use of the root mean squared error ($RMSE$) of the forward model. When evaluating the stop criterion for different data sets it became obvious that the same $RMSE$ could be used as a spectral quality test. Therefore this thesis suggested a simple quality test based on the error of the forward model and implemented this quality test in the published code. While this quality test is not yet elaborated for being applied on big spectral data from imaging, it works well for the analysis of smaller data sets. By visual inspection of the final $RMSE$ of all spectra, an upper limit can be set by the user. Strong deviations from the mean error in a data set indicate an unsuccessful correction, and the user may decide to revise initialization parameters. The development of an automated quality test that can be used on images in order to decide if a spectrum can be scatter corrected or if it has to be discarded will be objected to a future study.

Further stabilization of the algorithm was achieved by weighting the reference spectrum. By down weighting the chemically inactive regions in the reference spectrum, a stable baseline correction is obtained, and baseline variations in the spectra are reduced. Additional stability is obtained by setting negative parts of the reference spectrum to zero, which is done due to physical considerations.

In order to standardize the initialization parameters, the initial scaling of the reference spectrum should be preserved throughout the correction. This is handled by performing a basic EMSC in each iteration. In addition, default settings of the initialization parameter ranges and distributions are provided. The default parameter ranges are observed to be suitable for most of the data sets at hand in this study.

It was demonstrated that the Mie correction algorithm in general retrieves the true chemical features of the pure absorbance spectrum with high precision. Further, we demonstrated that the correction is not sensitive to the chemical characteristics of the reference spectrum, which is a critical feature of the algorithm. We show that a more reliable amide I peak position is retrieved by the Mie correction algorithm, which is shown to be a feature of the Mie meta-model. Both the shape and peak position of the amide I absorption band plays a crucial role in classification of cells and tissues.

The sensitivity towards the initialization parameters is assessed by use of both a set of simulated apparent absorbance spectra and measured apparent absorbance spectra. Clear directions on how to initialize the Mie correction algorithm are provided. A demonstration of how the correction algorithm can be applied to imaging data is given. The example illustrates how parameter estimation can be performed manually for a reasonably sized data set, as a preliminary approach on using the algorithm on imaging data. In this context we demonstrated that EMSC parameters obtained by the Mie correction algorithm contain meaningful information that can be used for discimination. The use of Mie EMSC parameters may be further explored for their use in feature extraction for machine learning.

In order to have a fully functional Mie correction code for its application to imaging data, measures should be done to decrease the computational time and automatize the parameter choices. By means of replacing the Fourier transform by the Hilbert transform and reducing

the number of model parameters, Konevskikh et al. [28] considerably improved the speed of the algorithm. However in order to handle large data sets the code should be adapted to GPU programming or parallel programming.

From the correction of the spectra obtained from the FTIR image on the colon tissue sample, it was evident that an effective and accurate quality test is required in order to handle data sets, which contain different types of tissues. As differences in morphology in general imply the need for different parameter settings, the parameter choices need be automatized in order to handle large and inhomogeneous data sets. For imaging data, the possibility of using information from neighbouring pixels in the correction should be investigated. This could possibly lead to a faster and more efficient choice of parameters. Further, optimization algorithms should be explored in order to achieve the most suited parameter ranges, and minimize the error of the forward model.

Bibliography

[1] Nils Kristian Afseth and Achim Kohler. "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial". In: *Chemometrics and Intelligent Laboratory Systems* 117.Supplement C (2012). Special Issue Section: Selected Papers from the 1st African-European Conference on Chemometrics, Rabat, Morocco, September 2010 Special Issue Section: Preprocessing methods Special Issue Section: Spectroscopic imaging, pp. 92–99. ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2012.03.004. URL: http://www.sciencedirect.com/science/article/pii/S0169743912000494.

[2] Murat Bagcioglu et al. "Monitoring of plant-environment interactions by high throughput FTIR spectroscopy of pollen". In: 8 (July 2017), p. 870.

[3] R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. "Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra". In: *Appl. Spectrosc.* 43.5 (May 1989), pp. 772–777.

[4] Andreas Barth and Christian Zscherp. "What vibrations tell about proteins". In: *Quarterly reviews of biophysics* 35.4 (2002), pp. 369–430.

[5] Paul Bassan et al. "Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples". In: 135 (Feb. 2010), pp. 268–77.

[6] Giuseppe Bellisola and Claudio Sorio. "Infrared spectroscopy and microscopy in cancer research and diagnosis". In: *American journal of cancer research* 2.1 (2012), p. 1.

[7] B Bird and J Rowlette. "High definition infrared chemical imaging of colorectal tissue using a Spero QCL microscope". In: *Analyst* 142.8 (2017), pp. 1381–1386.

[8] Maren Anna Brandsrud. "Understanding Resonant Structures of Coupled Disks for Light Management in Photovoltaics". MA thesis. Norwegian University of Life Sciences, 2015.

[9] Susanne W Bruun et al. "Correcting Attenuated Total Reflection–Fourier Transform Infrared Spectra for Water Vapor and Carbon Dioxide". In: *Applied spectroscopy* 60.9 (2006), pp. 1029–1039.

[10] Lin-P'ing Choo et al. "In situ characterization of beta-amyloid in Alzheimer's diseased tissue by synchrotron Fourier transform infrared microspectroscopy". In: *Biophysical Journal* 71.4 (1996), pp. 1672–1679.

[11] FRS Clark and DJ Moffatt. "The elimination of interference fringes from infrared spectra". In: *Applied Spectroscopy* 32.6 (1978), pp. 547–549.

[12] Laura Corte et al. "Development of a novel, FTIR (Fourier transform infrared spectroscopy) based, yeast bioassay for toxicity testing and stress response study". In: *Analytica chimica acta* 659.1-2 (Feb. 2010), pp. 258–265. ISSN: 0003-2670. DOI: 10.1016/j.aca.2009.11.035. URL: https://doi.org/10.1016/j.aca.2009.11.035.

[13] Alexandre Dazzi et al. "AFM–IR: combining atomic force microscopy and infrared spectroscopy for nanoscale chemical characterization". In: *Applied Spectroscopy* 66.12 (2012), pp. 1365–1384.

[14] Ariane Deniset-Besseau et al. “Monitoring triacylglycerols accumulation by atomic force microscopy based infrared spectroscopy in streptomyces species for biodiesel applications”. In: *The journal of physical chemistry letters* 5.4 (2014), pp. 654–658.

[15] M Diem et al. “A decade of vibrational micro-spectroscopy of human cells and tissue (1994–2004)”. In: *Analyst* 129.10 (2004), pp. 880–885.

[16] Thomas van Dijk et al. “Recovery of Absorption Spectra from Fourier Transform Infrared (FT-IR) Microspectroscopic Measurements of Intact Spheres”. In: *Applied Spectroscopy* 67.5 (2013). PMID: 23643044, pp. 546–552. DOI: 10.1366/12-06847.

[17] Marc F Faggin and Melissa A Hines. “Improved algorithm for the suppression of interference fringe in absorption spectroscopy”. In: *Review of scientific instruments* 75.11 (2004), pp. 4547–4553.

[18] P. Geladi, D. MacDougall, and H. Martens. “Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat”. In: *Appl. Spectrosc.* 39.3 (May 1985), pp. 491–500. URL: http://as.osa.org/abstract.cfm?URI=as-39-3-491.

[19] Carol Hirschmugl and Ghazal Azarfar. “NSF project meeting”. Advancing 3D Chemical Imaging: FTIR Spectro-microtomography, FTIR Spectro-microlaminography and Hyperspectral Data Analysis National Science Foundation (NSF) (CHE-1508240), Milwaukee, Wisconsin.

[20] Hendrik C. van de Hulst. *Light scattering by small particles*. Dover Publications, 2014. ISBN: 978-0-486-64228-4.

[21] J. L. Ilari, H. Martens, and T. Isaksson. “Determination of Particle Size in Powders by Scatter Correction in Diffuse Near-Infrared Reflectance”. In: *Applied Spectroscopy* 42.5 (1988), pp. 722–728. DOI: 10.1366/0003702884429058. eprint: https://doi.org/10.1366/0003702884429058. URL: https://doi.org/10.1366/0003702884429058.

[22] Adele CM Julier et al. “Chemotaxonomy as a tool for interpreting the cryptic diversity of Poaceae pollen”. In: *Review of Palaeobotany and Palynology* 235 (2016), pp. 140–147.

[23] Angela Kallenbach-Thieltges et al. “Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections”. In: *Journal of biophotonics* 6.1 (2013), pp. 88–100.

[24] Achim Kohler et al. “Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction”. In: 62 (Apr. 2008), pp. 259–66.

[25] Achim Kohler et al. “Extended Multiplicative Signal Correction as a Tool for Separation and Characterization of Physical and Chemical Information in Fourier Transform Infrared Microscopy Images of Cryo-sections of Beef Loin”. In: 59 (July 2005), pp. 707–16.

[26] Tatiana Konevskikh, Rozalia Lukacs, and Achim Kohler. “An improved algorithm for fast resonant Mie scatter correction of infrared spectra of cells and tissues”. In: (Aug. 2017), e201600307.

[27] Tatiana Konevskikh et al. “Fringes in FTIR spectroscopy revisited: understanding and modelling fringes in infrared spectroscopy of thin films”. In: *Analyst* 140 (12 2015), pp. 3969–3980. DOI: 10.1039/C4AN02343A. URL: http://dx.doi.org/10.1039/C4AN02343A.

[28] Tatiana Konevskikh et al. “Mie scatter corrections in single cell infrared microspectroscopy”. In: *Faraday Discuss.* 187 (0 2016), pp. 235–257. DOI: 10.1039/C5FD00171D. URL: http://dx.doi.org/10.1039/C5FD00171D.

[29] Peter Lasch, Anthony Pacifico, and Max Diem. "Spatially resolved IR microspectroscopy of single cells". In: *Biopolymers* 67.4-5 (2002), pp. 335–338. ISSN: 1097-0282. DOI: 10.1002/bip.10095. URL: http://dx.doi.org/10.1002/bip.10095.

[30] Kristian Hovde Liland, Achim Kohler, and Nils Kristian Afseth. "Model-based pre-processing in Raman spectroscopy of biological samples". In: *Journal of Raman Spectroscopy* 47.6 (2016), pp. 643–650.

[31] Rozalia Lukacs et al. "Recovery of absorbance spectra of micrometer-sized biological and inanimate particles". In: 140 (Mar. 2015).

[32] Harald Martens, S. A. Jensen, and P. Geladi, eds. *Multivariate linearity transformation for near-infrared reflectance spectrometry*. Proc. Nordic Symp. on Applied Statistics. Stokkand Forlag Publishers, p. 205.

[33] Harald Martens and Tormod Næs. *Multivariate Calibration*. Wiley, 1991. ISBN: 978-0-471-93047-1.

[34] Harald Martens, Jesper Pram Nilsen, and Søren Engelsen Balling. "Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures". In: *Anal. Chem.* 74.3 (2003), pp. 394–404. DOI: 10.1021/ac020194w.

[35] Harald Martens and Edward Stark. "Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy* 1". In: 9 (Feb. 1991), pp. 625–35.

[36] Harald Martens et al. "Pre-processing in biochemometrics: correction for path-length and temperature effects of water in FTIR bio-spectroscopy by EMSC". In: *Journal of chemometrics* 20.8-10 (2006), pp. 402–417.

[37] Lisa M Miller et al. "Synchrotron-based infrared and X-ray imaging shows focalized accumulation of Cu and Zn co-localized with β-amyloid deposits in Alzheimer's disease". In: *Journal of structural biology* 155.1 (2006), pp. 30–37.

[38] Brian Mohlenhoff et al. "Mie-type scattering and non-Beer–Lambert absorption behaviour of human cells in infrared micro-spectroscopy." In: 88 (2005), pp. 3635–40.

[39] Andreas Nabers et al. "Amyloid-β-secondary structure distribution in cerebrospinal fluid and blood measured by an immuno-infrared-sensor: A biomarker candidate for Alzheimer's disease". In: *Analytical chemistry* 88.5 (2016), pp. 2755–2762.

[40] Andreas Nabers et al. "An infrared sensor analysing label-free the secondary structure of the Abeta peptide in presence of complex fluids". In: *Journal of Biophotonics* 9.3 (2016), pp. 224–234. ISSN: 1864-0648. DOI: 10.1002/jbio.201400145. URL: http://dx.doi.org/10.1002/jbio.201400145.

[41] Dieter Naumann, Dieter Helm, and Harald Labischinski. "Microbiological characterizations by FT-IR spectroscopy". In: *Nature* 351 (6321 1991), pp. 81–82.

[42] Astrid Oust et al. "Fourier transform infrared and Raman spectroscopy for characterization of Listeria monocytogenes strains". In: *Applied and environmental microbiology* 72.1 (2006), pp. 228–232.

[43] Sushmita Paul and Mina Ray. "Periodically repeated gap-independent transmission in waveguide coupled microring Whispering Gallery Resonators (WGR)". In: *Journal of Optics* 42.3 (2013), pp. 203–207.

[44] Stanford Encyclopedia of Philosophy. *Models in Science*. 2012. URL: https://plato.stanford.edu/entries/models-science/ (visited on 12/13/2017).

[45] C. A. Rebuffo et al. “Reliable and rapid identification of Listeria monocytogenes and Listeria species by artificial neural network-based Fourier transform infrared spectroscopy”. In: 72.2 (2006), pp. 994–1000.

[46] Cecilia A Rebuffo-Scheer et al. “Identification of five Listeria species based on infrared spectra (FTIR) using macrosamples is superior to a microsample approach”. In: *Analytical and Bioanalytical Chemistry* 390.6 (Mar. 2008), pp. 1629–1635. ISSN: 1618-2642. DOI: 10.1007/s00216-008-1834-1. URL: https://doi.org/10.1007/s00216-008-1834-1.

[47] A. Savitzky and M.J.E. Golay. “Smoothing and differentiation of data by simplified least squares procedures”. In: *Anal. Chem.* 36 (1964), pp. 1627–1639. URL: http://dx.doi.org/10.1021/ac60214a047.

[48] Maarten Scholtes-Timmerman et al. “A novel approach to correct variations in Raman spectra due to photo-bleachable cellular components”. In: *Analyst* 134.2 (2009), pp. 387–393.

[49] Volha Shapaval et al. “A high-throughput microcultivation protocol for FTIR spectroscopic characterization and identification of fungi”. In: *Journal of Biophotonics* 3.8-9 (2010), pp. 512–521. ISSN: 1864-0648. DOI: 10.1002/jbio.201000014. URL: http://dx.doi.org/10.1002/jbio.201000014.

[50] Volha Shapaval et al. “A novel library-independent approach based on high-throughput cultivation in Bioscreen and fingerprinting by FTIR spectroscopy for microbial source tracking in food industry”. In: 64 (Oct. 2016).

[51] Patrick Suppes. “Models of data”. In: *Studies in the Methodology and Foundations of Science*. Springer, 1969, pp. 24–35.

[52] David B. Tanner. *Optical effects in solids*. URL: https://www.phys.ufl.edu/~tanner/notes.pdf (visited on 12/12/2017).

[53] Ali Tfayli et al. “Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1724.3 (2005), pp. 262–269.

[54] Thomas Udelhoven, Dieter Naumann, and Jürgen Schmitt. “Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria”. In: *Applied Spectroscopy* 54.10 (2000), pp. 1471–1479.

[55] Irvine University of California. *Iris Data Set*. URL: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data (visited on 12/13/2017).

[56] M. Wenning, H. Seiler, and S. Scherer. “Fourier-Transform Infrared microspectrometry, a novel and rapid tool for identification of yeasts”. In: *Appl. Environ. Microbiol.* 68 (2002), pp. 4717–4721.

[57] Rolf Wolthuis et al. “IR spectral imaging for histopathological characterization of xenografted human colon carcinomas”. In: *Analytical chemistry* 80.22 (2008), pp. 8461–8469.

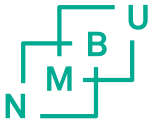
[58] Andrew Zangwill. *Modern Electrodynamics*. Cambridge University Press, 2016. ISBN: 978-0-521-89697-9.

[59] Boris Zimmermann. “Chemical characterization and identification of Pinaceae pollen by infrared microspectroscopy”. In: (Sept. 2017), pp. 1–10.

[60] Boris Zimmermann and Goran Baranović. “Determination of Phase Transition Temperatures by the Analysis of Baseline Variations in Transmittance Infrared Spectroscopy”. In: *Appl. Spectrosc.* 63.10 (Oct. 2009), pp. 1152–1161. URL: http://as.osa.org/abstract.cfm?URI=as-63-10-1152.

[61] Boris Zimmermann and Achim Kohler. "Optimizing Savitzky–Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy". In: *Applied Spectroscopy* 67.8 (2013). PMID: 23876728, pp. 892–902. DOI: 10.1366/12-06723.

[62] Boris Zimmermann et al. "Analysis of Allergenic Pollen by FTIR Microspectroscopy". In: 88 (Nov. 2015), pp. 803–811.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway