Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS
Faculty of Science and Technology

Preprocessing strategies for infrared spectral data with limited numbers of spectral channels

Tiril Aurora Lintvedt
Environmental physics and renewable energy

Preface

This master thesis is a result of years of curiosity and studying (to gratify this curiosity). The goal to understand the world around me has always pushed me forward and motivated me to look closer into any problem at hand. One subject that has intrigued me from a young age is quantum physics, and at the time of introduction to this field, I perceived the ideas within it as abstract and mysterious. It was in many ways the fascination for this subject that led me to the topic of this thesis. In 2017, I was introduced to my supervisor Achim Kohler through a quantum physics course at the Norwegian University of Life Sciences (NMBU). This course was to me one of the most rewarding subjects I had studied at the time. Something which I had earlier perceived as abstract and mysterious, now became a little bit more tangible and I wanted to grab it. By the end of this course, I was determined to apply some of this acquired knowledge in my master thesis. Through the concept of vibrational spectroscopy, which is based on quantized absorption of radiation in matter, I have been able to apply quantum physics in a very real way. At the same time, I have been lucky to be involved in an EU project, called Miracle. Here, the concept of vibrational spectroscopy meets challenges in the health care domain, making the work in this thesis truly rewarding for me.

I would like to thank the BioSpec research group at NMBU for receiving me with open arms. Special thanks to the Miracle project group, including Valeria Tafintseva, Johanne Solheim, Boris Zimmermann and Achim Kohler, for helpful discussions and follow-up conversations through a whole year of studying for this thesis. I am particularly grateful I was given the opportunity to travel to Minsk and present my work during the BioSpecMLC conference in August 2019. Additional thanks to the rest of the Miracle consortium, for experimental work and data acquisition done prior to my thesis.

Ås, December 16th 2019

Tiril Aurora Lintvedt

Abstract

Joint health is vital for mobility and well being of all people. In an EU project called Miracle, a mid-infrared arthroscopic probe for diagnosis of joint diseases during surgery is being developed. The focus is mid infrared measurements of cartilage tissue. To deal with instrumental spatial issues, the probe will emit only a few single wavenumbers, thus presenting a challenge for data preprocessing. The state of the art preprocessing technique Extended Multiplicative Signal Correction (EMSC) is a well established technique which corrects for physical effects such as scaling and different types of baseline variations in the data. The method is built on broad-band spectra, and for data with few wavenumber channels the stability of the EMSC can not be guaranteed. In thesis, this issue was investigated further. As the Miracle project is an ongoing project, and the final probe was not ready for operation during this master thesis, measurements employing the probe were not available. Therefore, we simulated a data set in order to develop a preprocessing strategy. The simulation was based on applying PCA on a data set of existing broad-band spectra measurements that were obtained from healthy and diseased samples by various project groups and on different conventional lab instruments. We identified several interference and measurement variations from the experimental broad-band data, including variations in water vapor, carbon dioxide, noise and cartilage signal strength. Spectra completely without cartilage signal was also found, which were linked with high degradation cartilage samples. However, it was shown that the high-degradation class membership for such spectra is not guaranteed in practice, and therefore it is concluded that such spectra will not give any meaningful value to classification tasks of healthy and diseased cartilage. This demonstrates the importance of the development of an automatic detection algorithm for measurements which deviates highly from the expected cartilage signal. Further, it was suggested that physical effect can carry discriminative information about healthy and diseased cartilage for broad-band spectra. It was however shown, that for the seven wavenumber channel data, corresponding EMSC type correction methods was not as accurate for seven wavenumber channels data as for broad-band spectra, and that the difference increased with the complexity of the EMSC model. Therefore, it is concluded that the estimated physical effects baseline parameters from the EMSC correction of seven wavenumber channels data most likely does not correctly describe physical phenomena in the sample, but may in stead express a trend in the relationship between absorbance levels for the seven wavenumbers. In total 11 EMSC type preprocessing strategies for seven wavenumber channels data were suggested, and validated using the simulated data set. The best performance across four tested classifiers was obtained by using the conventional MSC. The inclusion of the estimated MSC parameters as extra input variables to the classifier led to further increase in accuracy, though marginal. In combination with the Random Forests classifier, the maximum accuracy of 81,2 % was achieved, which represented an increase of 6,2 % with respect to classification based on raw data. Lastly, we demonstrated that water vapor is disturbing for classification based on seven wavenumber channels data. By including water vapor in simulation, we found that the classification accuracy (Random Forest) decreased by 5 %. Based on this, it is recommended that instrumental precautions are made to try and minimize presence of water vapor.

Sammendrag

Leddhelse er viktig for alle menneskers mobilitet og velvære. I et EU-prosjekt kalt Miracle, utvikles et midt-infrarødt kikkertinstrument for diagnostisering av leddsykdommer under operasjoner. Fokuset er IR målinger på bruskvev. For å håndtere instrumentale utfordringer vil kun noen få utvalgte bølgelengder utstråles, og dermed føre til en utfordring for preprosessering av data. Den moderne forbehandlingsteknikken Utvidet Multiplikativ Signalkorreksjon (EMSC) er en veletablert teknikk som korrigerer for fysiske effekter som skalering og forskjellige typer baseline-variasjoner i spektre. Metoden er bygget på kontinuerlige spektre med bredt spektralområde, og for data med få bølgetallkanaler kan ikke stabiliteten til EMSC garanteres. Denne problemstillingen ble undersøkt nærmere i masteroppgaven. Ettersom Miracle-prosjektet er et pågående prosjekt, og det endelige kikkertinstrumentet er ferdigstilt, var ikke målinger fra instrumentet tilgjengelige. Derfor simulerte vi et datasett for å utvikle en forbehandlingsstrategi. Simuleringen var basert på anvendelse av PCA på et datasett med eksisterende bred-område spektre som ble oppnådd fra friske og skadede bruskprøver av forskjellige prosjektgrupper og med forskjellige konvensjonelle laboratorieinstrumenter. Vi identifiserte flere interferens- og målevariasjoner fra eksperimentelle bred-område data, inkludert variasjoner i vanndamp, karbondioksid, støy og signalstyrke fra brusk. Spektra helt uten brusksignal ble også funnet, og ble koblet til bruskprøver med skade. Det ble imidlertid vist at ikke alle spektre uten brusksignal stammet fra prøver med stort skadeomfang, og derfor konkluderes det med at slike spektra ikke vil gi noen meningsfull verdi i klassifiseringsoppgaver av friskt og skadet brusk. Dette demonstrerer også viktigheten av utviklingen av en automatisk deteksjonsalgoritme for målinger som avviker sterkt fra forventede brusk-signal. Videre ble det foreslått at fysiske effekter kan gi diskriminerende informasjon om friskt og skadet brusk. Det ble imidlertid demonstrert at for data med kun syv bølgetallkanaler, var korresponderende EMSC-type korreksjonsmetode ikke like nøyaktig som bred-område spektre, og at forskjellen økte med inkludering av bølgetall-avhengige beselines til EMSC-modellen. Derfor konkluderes det med at estimerte fysiske effekter fra EMSC-korreksjon av syv bølgetallkanaldata mest sannsynlig ikke beskriver fysiske fenomener i prøven korrekt, men likevel uttrykker en trend i forholdet mellom absorbansnivåer for de syv bølgetallene. Totalt 11 EMSC-type forbehandlingsstrategier for syv bølgetallkanaldata ble foreslått og validert ved bruk av det simulerte datasettet. Den beste ytelsen over fire testede klassifiseringsalgoritmer ble oppnådd ved bruk av MSC. Inkluderingen av de estimerte MSC-parameterne som ekstra variabler for klassifikatoren førte til ytterligere økning i suksessrate, men marginal. I kombinasjon med klassifiseringsalgoritmen Random Forest oppnådde vi en maksimal nøyaktighet på 81,2 %, noe som representerte en økning på 6,2 % med hensyn til klassifisering basert på rådata. Til slutt demonstrerte vi at vanndamp er forstyrrende for klassifisering basert på syv bølgetallkanaldata. Ved å inkludere vanndamp i simuleringen gikk suksessraten til klassifikatoren (Random Forest) ned med 5 % i sammenlikning med simulering som ikke inkluderte vanndamp. Basert på dette anbefales det at det tas instrumentelle forholdsregler for å prøve å minimere tilstedeværelsen av vanndamp.

Contents

Terminology and Abbreviations

ANN Artificial Neural Networks. 16

EMSC Extended Multiplicative Signal Correction. 14

MSC Multiplicative Signal Correction. 15

OA Osteoarthritis. 1

PG Proteoglycan. 9

QCL Quantum Cascade Laser. 1, 2

RF Random Forests. 15

SVM Support Vector Machines. 16

Chapter 1

Introduction

1.1 Motivation

Joint health is vital for mobility and well being of all people. Each and everyday our joints are carrying the load of our body, and are going through high strains such as heavy lifting and sports activities. The articular cartilage tissue in the joints may be subject to small or major trauma, and as a result suffer small or major damages. If cartilage damages are not treated or incorrectly treated, the damage can develop through time and ultimately lead to chronic diseases such as Osteoarthritis (OA). OA is a chronic joint disease characterized by degenerative changes to the bones, cartilage, menisci, ligaments, and synovial tissue [1]. It was estimated in 2010 that 4.7 % of the global population suffer from osteoarthritis (hip and knee), of which 3.8 % represent knee osteoarthritis [1]. The current evaluation methods for articular cartilage during surgery have been reported to be subjective and invasive. In addition, the current evaluation methods do not allow discovery of degeneration on an early stage. It is thus desirable to develop new tools for aiding the evaluation of articular cartilage, which will be objective and noninvasive. This is the aim of an ongoing ICT Horizon 2020 EU project (Miracle, ICT-30-2017: Photonics KET 2017: Mid-infrared arthroscopy innovative imaging system for real-time clinical in depth examination and diagnosis of degenerative joint diseases).

1.2 The Miracle project

In the Miracle project, a mid-infrared arthroscopic probe for diagnosis of joint diseases during surgery is developed. The main focus is examination of articular cartilage in knee joints. To this purpose, Quantum Cascade

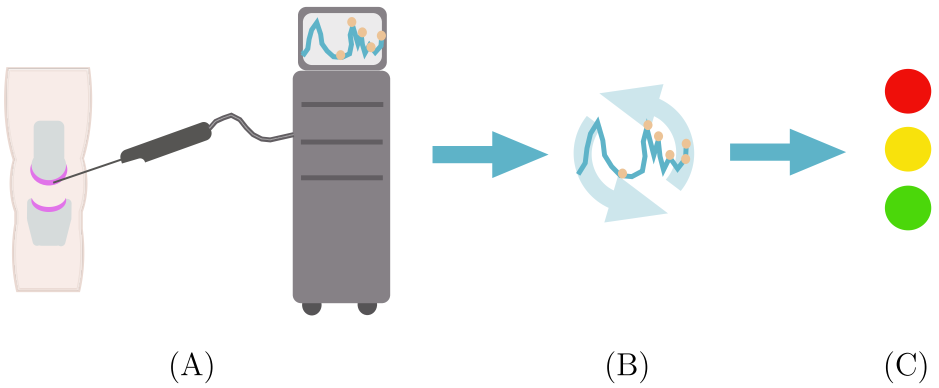


Figure 1.1: This figure illustrates the probe system which is under development in the Miracle project. The system consists of a mid infrared probe providing an in-situ measurement (A), real-time preprocessing (B) and classification (C) of cartilage damage.

Laser (QCL) elements are employed together with lasers which emit selected wavelengths for producing a frequency combination structure. For coupling these selected wavelengths into the input waveguide of the actual probe, an on-chip beam combiner based on thin-film semiconductor waveguide technology is used. The final aim is a system illustrated in Fig. 1.1, where the Miracle probe can provide in-situ measurements (A) which subsequently are preprocessed (B) and used for classification (C) of cartilage damage degree, providing the surgeon with objective insights. The Miracle probe will provide a seven wavelength absorbance signal based on the ATR sampling technique. Feature selection for determination of these seven wavelengths was done by Partial Least Squares Discriminant Analysis (PLS-DA) and Sparse Partial Least Squares Discriminant Analysis (SPLS-DA) prior to this thesis.

1.3 Data preprocessing challenge

It is well established that preprocessing is an important part of IR-spectroscopy. For applications of mid-infrared spectroscopy, employing radiation sources in the region 2,5 μm - 25 μm, typical effects that interfere with informative signals are signals from water, baseline shifts and scaling effects due to variations in the contact between the probe and the sample. In conventional IR spectroscopy of biological materials, these effects are commonly removed from absorbance spectra by the model-based preprocessing technique Extended Multiplicative Signal Correction (EMSC) [2, 3, 4, 5, 6], which is a state of the art preprocessing technique. This technique is built on spectra for which whole spectral ranges are available, in essence for cases where a large number of wavenumbers are probed (>1000). In such data, there is a high co-linearity between variables. This co-linearity is not present in QCL-based waveguide data, and thus the stability of the EMSC used as preprocessing for such seven wavenumber channels data can not be guaranteed. Consequently, for preprocessing the infrared spectra produced by the Miracle-probe, preprocessing approaches must be investigated in greater detail.

1.4 Scope of thesis

The Miracle project is an ongoing project, and, according to the plan, the final probe was not produced and ready for operation during the master thesis. Thus, measurements employing QCL lasers were not available at the beginning of the Miracle project. The Miracle QCL lasers were expected to be finished after half of the project time. However, it was vital for the project to start investigating the consequences for the availability of just a few number of wavelengths for the preprocessing strategies. Therefore, it was decided to simulate a data set in order to develop a preprocessing strategy. The idea was to base the simulation on a data set that could be used for developing a preprocessing strategy, on existing spectra measurements that were obtained from healthy and diseased samples by various project groups and on different conventional lab instruments, that allow to acquire spectra over the full spectral range (broad-band spectra). Therefore the aim of this thesis was to (i) explore interferent and measurement variability in broad-band spectra, (ii) establish routines for detection of low quality broad-band spectra, (iii) use only selected wavelengths from the broad-band spectra (the wavelengths that were selected for the QCL lasers) and investigate preprocessing strategies based on only few wavelengths, (iv) to suggest preprocessing strategies for data with few wavelength channels, and finally (v) to simulate a data set based on the knowledge about interference effects from broadband spectra and use the simulated data set for validation of the suggested preprocessing strategies.

Chapter 2

Theory

2.1 Vibrational spectroscopy

In this section, the reader is provided with background material for the field of Vibrational Spectroscopy. This section is based on [7]. Vibrational spectroscopy, or infrared spectroscopy is a widely used tool in chemistry research [8, 9, 10]. It is a non-destructive tool, which can help scientists identify functional groups in both organic and inorganic samples by taking advantage of quantized absorption by the compounds in the sample. When a sample is measured in vibrational spectroscopy, radiation in the infrared region of the electromagnetic spectrum is sent through the sample, for which present functional groups absorb characteristic wavelengths and give rise to changes in molecule vibrations, in essence transitions in vibrational energy states. The absorbed wavelengths are recorded by spectroscopic instrumentation, creating a so called *fingerprint* for the given sample. There are several ways to measure such characteristic absorption in molecules. Some possible instrumental setups are FTIR, Raman or AFMIR. In the following sections, the main focus is FTIR spectroscopy.

2.1.1 Molecule absorption of IR radiation

Absorption of IR radiation mainly cause changes to molecule vibrations. In this section, different types of vibrations molecules can have are shortly introduced. The vibrational modes are defined by stretching and bending modes. Stretching is when the atoms moves along the axis of the chemical bond between them. There are two types of stretching modes; asymmetric and symmetric. Bending is when the angle between two chemical bonds is continuously changing. There are four types of bending modes; wagging, twisting, rocking and scissoring. For diatomic and triatomic molecules, these are easily understood, but for more complex compounds, interactions between different modes of vibrations becomes very complicated and unique. The vibration of a molecule exist in quantized energy levels or so called vibrational energy states. The vibrational energy states $\mathrm{V}_{\mathrm{i}\nu}$ of chemical bonds can be described as anharmonic oscillations [7], by the following equation

$$V_{iv} = h\nu_i\left(\upsilon_i + \frac{1}{2}\right) + h\nu_i x_i\left(\upsilon_i + \frac{1}{2}\right)^2 \tag{2.1}$$

, where h is Planck's constant, ν_{i} is the characteristic frequency of vibrational mode i, υ_{i} is the vibrational quantum number of mode i (υ_{i} = 0,1,2,...) and x_{i} is the dimensionless anharmonicity constant for mode i. The first term is the energy states of harmonic oscillations, and the second term is the anharmonicity contribution. The energy difference between the fundamental state (i = 0) and the first excited state (i = 1) often correspond with frequencies in the mid infrared region. Thus, when bonds are illuminated by IR radiation, the bonds will absorb it. Notably, transitions between modes with larger energy gaps does not in general produce high signals in the mid infrared spectrum. However, absorption in the sample is not strictly limited to to vibrational energy transitions. Liquid and solid samples can have vibrational energy transitions, but small gaseous molecules such as water vapor and carbon dioxide can in addition have rotational transitions when illuminated by IR radiation. Spectra of such small molecules in the vapor phase show considerable fine structure because transitions between quantized rotational energy levels occur at the same time as vibrational transitions.

2.1.2 Lambert-Beer's law

Lambert-Beer's law is one of the most fundamental relations/equations in vibrational spectroscopy. While transmittance $T(\tilde{\nu})$ of a sample at a given wavenumber $\tilde{\nu}$ can experimentally found given by the ratio of the radiant power emerging from the rear face of the sample at that wavenumber $I(\tilde{\nu})$ to the power of the radiation at the front face of the sample$I_0(\tilde{\nu})$, the Lambert-Beer's law provide a useful approximation of how absorbance in the sample can be described. For a pure component sample, the Lambert-Beer's law takes the form in equation 2.2 [7], which is the simplest form of the equation.

$$T(\tilde{\nu}) = \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} = e^{-\alpha(\tilde{\nu})b} \tag{2.2}$$

, where b is the sample thickness and $\alpha(\tilde{\nu})$ is the linear absorption coefficient at $\tilde{\nu}$. From this relation the absorbance of the sample can be calculated. Taking into account that most samples are mixtures of several components which absorbs at $\tilde{\nu}$, the absorbance can be expressed as

$$A(\tilde{\nu}) \approx \sum_{j=1}^{J} [k_j(\tilde{\nu}) b c_j] \tag{2.3}$$

, where J is the number of absorbing constituents at $\tilde{\nu}$, $k_j(\tilde{\nu})$ is the absorptivity at $\tilde{\nu}$ of component j, and c_j is the concentration of component j.

2.1.3 Fourier transform infrared spectroscopy

In the FTIR spectroscope, a transmission spectrum for a sample is obtained by utilising the Michelson's interferometer [11] to produce an interferogram, and subsequently turned this into a transmission spectrum by utilising the Fourier transform. There are several ways to obtain a measurement of a sample with an FTIR spectrometer, and the sampling technique of choice depends on the application. The main possibilities include Transmission, Attenuated Total Reflection (ATR)[12] , Diffuse Reflectance [13, 14] and Specular Reflectance [15]. The most classical sampling technique in FTIR spectroscopy is the transmission sampling technique. However, the transmission sampling technique does not allow in-situ applications since samples must be very thin($\sim$10 μm [7]) and require careful sample preparation, techniques utilising reflection instead of transmission have an advantage in this area. In the next paragraph, the ATR sampling technique, which is one of the techniques which utilise reflections on the sample surface, is shortly explained.

ATR sampling technique

The ATR sampling technique is based on the phenomenon of total internal reflection, and the sampling setup of a single-bounce system is illustrated in Fig. 2.1. In this setup, the changes which occur in an internally reflected infrared beam (1 reflection for single bounce system) which comes in contact with the sample through a crystal or diamond (high refractive index) is measured. Upon contact with the sample, an evanescent wave, which extends into the top surface of the sample ($\sim$1-2 μm), is generated. Thus, the evanescent wave will be attenuated by absorption of chemical bonds in the sample surface [16]. The exact penetration depth depends on the particular wavelength in the beam and several other factors, such as the difference in refractive indices of the sample and the crystal, the angle of incidence of the beam, the number of reflections [17]. Since the penetration depth is only around ∼1-2 μm, for measurements of solids it is important that the ATR diamond tip is applied with pressure on the sample. Comparison between spectra of different sampling techniques should be made with caution, since different techniques will involve different types of physical phenomena. For example, it should be noted that ATR spectra have a shift to lower frequencies compared to transmission spectra [18].

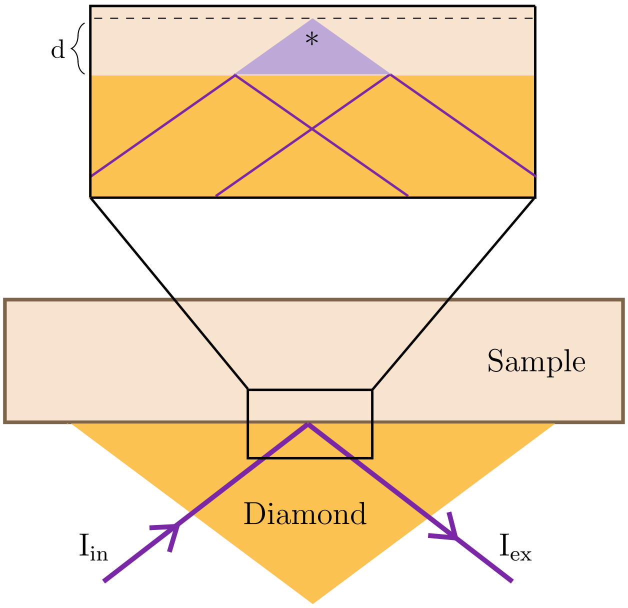


Figure 2.1: This figure shows a simple schematic of a single bounce ATR system. The incident beam I_{in} is reflected once on the sample. An evanescent wave* penetrates the sample with depth d, resulting in an attenuated exit beam I_{ex}.

2.1.4 The infrared absorbance spectrum

The mid infrared region is often divided into a so called functional region above 1500 cm^{-1} and the fingerprint region below 1500 cm^{-1}. The functional region is the region including absorbance of separate functional groups within the molecule, while the fingerprint region contain absorption due to complex deformations of the molecule. However, this assignment is not strict, since the two region in practice will overlap. For the purpose of this thesis, we consider the cartilage IR spectrum which will be encountered in this thesis. As a comparative note, the IR spectra of bone is also considered in this section. For a full overview of absorption peaks in the fingerprint region associated with cartilage tissue, see table 2.1. Some main peaks expected are collagen-associated peaks, protein-associated peaks (Amide I- III) and peaks associated with proteoglycans. As representatives for cartilage information, the 7 preselected wavenumbers for the laser sources in the Miracle project are 1800 cm^{-1} (Background), 1745 cm^{-1} (Lipids), 1620 cm^{-1} (Amide I), 1560 cm^{-1}(Amide II), 1210 cm^{-1} (Amide II), 1080 cm^{-1} (Collagen) and 850 cm^{-1} (Water/COS). Lipids and water bands are not included as cartilage components in table 2.1 but they are still present in synovial fluid and cells (chondrocytes) in the cartilage [19], and are thus in practice expected to be measured as well. In figure 2.2 [20], the qualitative differences between bone and cartilage is highlighted by showing typical IR transmission spectra for the two. As can be seen, for the protein associated peaks Amide I, II and III are present for both bone and cartilage, although for cartilage, the peaks are in general stronger. In the region 1000-1100 cm^{-1}, the most apparent difference occurs. While bone tissue is characterised by steep phosphate associated peaks, the corresponding cartilage signal is expected to be considerably lower, and is dominated by Proteoglycan (PG) absorption. As noted in section 2.1.3, for ATR spectra, exact match of peak positions should not be expected. The main phosphate peak for bone in FTIR-ATR instrumentation is seen at 1010 cm^{-1} [21]. As a last remark, the Miracle laser with radiation of wavenumber 850 cm^{-1}, may be a measure of water or carbonyl sulfide, but if cartilage is so worn out that we measure in stead bone like tissue, it may include information about carbonate content, as seen from figure 2.2 (left).

2.2 Disturbances in IR spectra

For optical instruments and other types of sensors, there will always be factors disturbing the desired signal. It may be that there are chemical signals we measure in our samples that we are not interested in, or there may be physical effects in either the instrumentation or in the sample itself augmenting the recorded absorbance signal. In chemometrics, the desired information is more often than not to obtain pure chemical information by using optical instruments, for instance the FTIR spectrometer. However, for such instrumentation and the sample of interest, there can be physical phenomena such as scattering, interference of IR waves [31] and variations in optical path which may cause the recorded spectrum to take on different characteristics,

Table 2.1: This table shows assignments of absorption peaks to bond vibrations for cartilage tissue in fingerprint region.

<table>
<tr><th>Frequency (cm^{-1})</th><th>Vibration</th></tr>
<tr><td>1700-1600</td><td>C=O stretch (Amide I) [22, 23, 24, 25, 26]</td></tr>
<tr><td></td><td><table>
<tr><th>Frequency (cm^{-1})</th><th>Secondary structure of collagen</th></tr>
<tr><td>1691</td><td>β-turns</td></tr>
<tr><td>1679</td><td>β-sheets</td></tr>
<tr><td>1669</td><td>β-turns</td></tr>
<tr><td>1658</td><td>α-helix</td></tr>
<tr><td>1647</td><td>unordered</td></tr>
<tr><td>1637</td><td>triple helix</td></tr>
<tr><td>1626</td><td>β-sheets</td></tr>
<tr><td>1608</td><td>side chains</td></tr>
</table></td></tr>
<tr><td>1600-1500</td><td>C-N stretch and N-H bend (Amide II) [26]</td></tr>
<tr><td>1480-1440</td><td>CH3 and CH2 deformations [27, 28]</td></tr>
<tr><td>1400</td><td>COO- stretch of amino side chains [27]</td></tr>
<tr><td>1375</td><td>CH3 symmetric deformation of glycosaminoglycans [29]</td></tr>
<tr><td>1335</td><td>CH2 deformations of collagen side chains [27]</td></tr>
<tr><td>1300-1200</td><td>O=C-N-H stretch and bending (Amide III) vibration with significant mixing with CH2 wagging vibration from the glycine backbone and proline side chain [27]</td></tr>
<tr><td>1250-1220</td><td>S=O stretch (SO3-) of sulphated glycosaminoglycans [26, 30]</td></tr>
<tr><td>1200-900</td><td>C-O-C, C-O, C-C ring, C-OH vibrations [27, 28, 30]</td></tr>
<tr><td></td><td><table>
<tr><th>Frequency (cm^{-1})</th><th>Vibrations</th></tr>
<tr><td>1160</td><td>C-O-C stretch</td></tr>
<tr><td>1120</td><td>C-O-C antisymmetric stretch</td></tr>
<tr><td>1080</td><td>C-O stretch of the carbohydrate residues in collagen and proteoglycans</td></tr>
<tr><td>1064</td><td>C-O stretch of the carbohydrate residues in proteoglycans</td></tr>
<tr><td>1032</td><td>C-O stretch of the carbohydrate residues in collagen and proteoglycans</td></tr>
</table></td></tr>
<tr><td>1065</td><td>SO3 symmetric stretch of sulphated glycosaminoglycans [30]</td></tr>
<tr><td>850</td><td>C-O-S stretch [26]</td></tr>
</table>

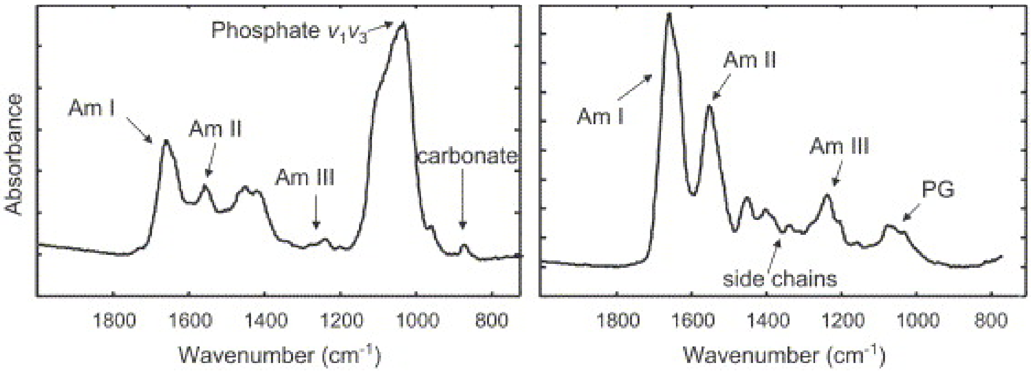


Figure 2.2: This figure shows the expected difference in the FTIR spectrum of cortical bone and articular cartilage, using instrumentation in transmittance mode. (Left) Healthy cortical bone and (Right) bovine articular cartilage. Reprinted from *FT-IR imaging of native and tissue-engineered bone and cartilage* by A. Boskey, and N. Camacho, 2007, Biomaterials, 28. Copyright 2006 by Elsevier Ltd.

such as multiplicative effects and baseline shifts. The disturbance of physical effects on the spectra may be more or less complicated, depending on the type of sample. For biological tissues, which are in general inhomogeneous, concentration differences of compounds may be one source of variability, and the presence of spherical structures such as cells may lead to specific scattering types, such as the Mie Scattering [3]. It can be noted that the Attenuated total reflection sampling technique is known to elliminate several of the spectral disturbances which are seen for other sampling techniqes [32], and the main issue is that radiation has increased penetration depth for lower wavenumbers. A spectrum which is not yet corrected for such physical effects is often called an *apparent absorbance spectrum.* After correction, the spectrum is referred to as *pure absorbance spectrum.* In to physical phenomena effects, random fluctuations in the spectrometer may disturb the recorded spectra in varying levels for different instruments.

In addition to physical effects, chemical information in itself can be seen as disturbances for a given application. In IR spectroscopy, one main concern are water signals. The water molecule is a polar molecule which has very high attenuation coefficient in the IR region. The exact absorption depends on the phase of the water. For liquid water, absorption due to vibrational transitions can often end up dominating the IR spectrum. The IR spectrum of water is shown in Fig. 2.3. High absorption bands are present at at 3500 cm^{-1} and 1635 cm^{-1} , which are caused by respectively O-H stretching and O-H-O scissor bending. Further a smaller band is located centred at 2120 cm^{-1}, which is the result of coupling of the scissors-bending and a broad liberation band in the near-infrared. The small absorption peak is for this reason called a combination band [17]. Often, challenges in sample preparation are due to high water content. For instance for measurements employing transmission mode of the FTIR instrument, it is required that such samples are very thin to not saturate the signal. For ATR measurements, the problem is not as pronounced because the penetration depth of the evanescent wave is typically very low, limiting the effective sample thickness. However, variability in water concentrations in the sample may still be a source of uncertainty. The Amide I is a known peak associated with protein absorption, and is expected for cartilage spectra as can readily be seen from figure 2.2. If the spectrum of liquid water is inspected in Fig. 2.3,it is also seen that one of the peaks are expected in the Amide I region. This is a good example of how water can disturb our spectra in perspective of further analysis. In the Amide I region it is hard to separate out signals we are interested in because of significant overlap of absorption bands, and whether a change in Amide I level originates from the sample constituents of interest or from less interesting constituents such as for example water, is difficult to determine. Other absorbing molecules of disturbance may be water vapor and carbon dioxide, which is often measured because air resides inside the instrument. Water vapor bands and carbon dioxide are shown in Fig. 2.4. As can be seen, water vapor (A) has to absorption regions, namely 3231 - 4000 cm^{-1} and 1205 - 2072 cm^{-1} originating from respectively stretching and bending. For carbon dioxide also two region exist, although only one is shown here. The one shown in Fig. 2.4 (B), namely in the region 2208 - 2442 cm^{-1} originates from asymmetric stretching and the second region 600 - 914 cm^{-1} originates from bending of the molecule [33]. Signals from carbon dioxide and water vapor are often measured due to air in the instrumentation, and in this case do not represent information about the sample itself. Such interference of the sample signal may disturb further analysis and preprocessing because of the sharp characteristic peaks associated with rotational transitions of small gas molecules. From this section it is understood that it is important to be aware of signals and phenomena that may disturb further analysis so that proper preprocessing of the spectra can be applied and limitations in analyses are known. In this section some common physical and chemical interferents were presented. In the next section, ways of preprocessing spectra to deal with such effects in spectra are introduced.

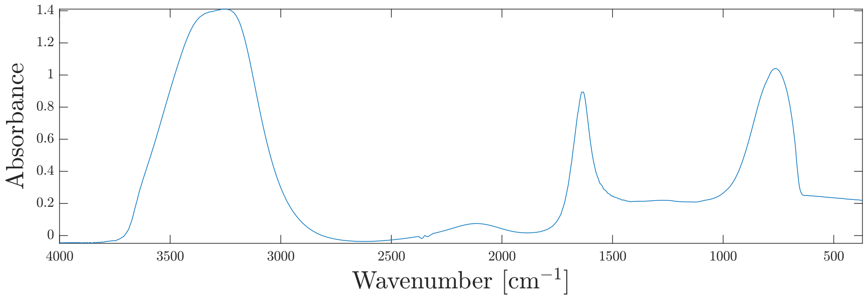


Figure 2.3: This figure shows a plot of the IR spectrum of liquid water, obtained by FTIR-ATR. Courtesy of Nebojsa Perisic and Achim Kohler.

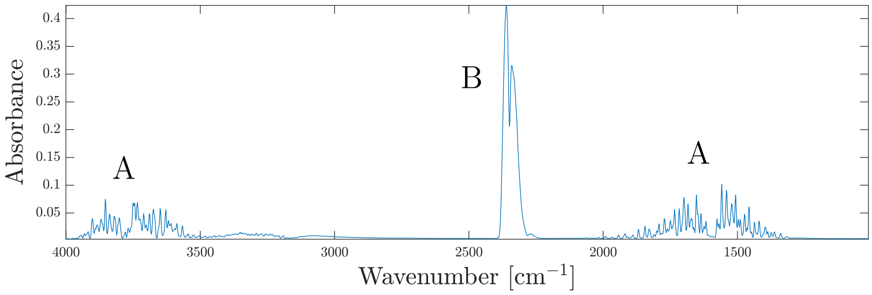


Figure 2.4: This figure shows a plot of the IR spectrum, obtained by FTIR-ATR, of water vapor (A) and carbon dioxide (B), which are interferents that can be associated with air inside the spectroscope. Courtesy of Nebojsa Perisic and Achim Kohler.

2.3 Preprocessing techniques for spectral data

To handle physical effects and intereferences as described in section 2.2, several common preprocessing methods can be mentioned, including Normalisation, derivative calculations (e.g. Savitsky-Golay), background subtraction, Standard Normal Variate (SNV), and Multiplicative Signal Correction (MSC) [34, 35]. In this section we consider an extended version of the MSC in more depth.

2.3.1 Extended Multiplicative Signal Correction

Extended Multiplicative Signal Correction (EMSC) is a well established algorithm for correction of physical effects in infrared spectra [3, 4, 2, 5]. It is a model based approach which can be used to correct both instrumental interference and interference in sample. It is a Least Squares method, where a predefined number of model component are fitted to the measured spectrum. These model components usually include both constant and wavenumber-dependent baselines and it is stabilised by using a reference spectrum, often chosen to be the mean spectrum in the data set. The general formulation of the EMSC model is summarized in equations 2.4 and 2.5.

$$A_{app}(\tilde{\nu}) = a + b \cdot \sum_{j=1}^{J} c_j \cdot k_j(\tilde{\nu}) + d \cdot \tilde{\nu} + e \cdot \tilde{\nu}^2 + \cdots + \epsilon \tag{2.4}$$

, where $A_{app}(\tilde{\nu})$ is the apparent absorbance spectrum, a is a constant baseline shift, b is a multiplicative effect representing effective optical thickness, c_j is the constituent concentrations in the sample, $k_j(\tilde{\nu})$ is the constituent's characteristic absorptivities, J is the number of absorbing species in the sample, d is linear baseline shift, and lastly e is a quadratic basline shift. This can be rewritten with respect to a reference spectrum $m(\tilde{\nu})$, by

$$A_{app}(\tilde{\nu}) = a + b \cdot m(\tilde{\nu}) + d \cdot \tilde{\nu} + e \cdot \tilde{\nu}^2 + \cdots + E \tag{2.5}$$

, where the information about the chemical differences between the reference spectrum and the measured apparent spectrum is assumed to be captured in the residual E. The larger b, the higher the probability that light is absorbed by a molecule. The parameters a, d and e is normally associated with scattering effects. This model is closely related with Lambert-Beer's law, as the description of absorption in the sample (sum term) in 2.4 is analogous to 2.3 as can be readily seen. The model components are fitted to the given spectrum by the Least Squares method, and the parameters are estimated and subsequently the physical effects can readily be separated from chemical information in the spectrum. If we only include the constant baseline (a term) and the multiplicative effect (b), the correction is referred to as Multiplicative Signal Correction (MSC). In theory, we can add any types of term to the model, such as polynomials or sinusoidal terms [36] [31]. In this thesis, only the addition of linear (d) and quadratic (e) terms are considered. It is distinguished between an EMSC model including only a linear term and a model including both linear and quadratic terms by referring to these as respectively MSC-L and EMSC.

2.4 Machine learning algorithms

Machine learning is the study of algorithms and statistical models which can be used for the purpose of learning dependencies and pattern in acquired data in order to make predictions for class membership of future observations. Data on which such learning algorithms are built is called a training set. Data, on which the learning algorithms are tested, is referred to as the test set. In this section, different Machine Learning algorithms used in this thesis are shortly presented to provide the reader with an overview. The methods are not explained in mathematical details. The section is based on [37, 38, 7, 39], and the main source used for each section is given at the end of the respective section. In addition to the below mentioned Machine Learning algorithms, PLS-DA is used for classification in this thesis. PLS-DA is a special case of Partial Least Squares Regression (PLSR), where the PLSR output is simply mapped into classes. It is assumed that the reader is familiar with PLSR.

2.4.1 Random Forests

The Random Forests (RF) classifier belongs to the Decision tree family. It sets up multiple decision trees, usually referred to as an ensemble of trees, where each tree is built from a random selection of samples from the original data (i.e by bootstrapping) and each node is optimised using a random subset of variables. After a tree is built on a random subset of samples, the remaining samples which the tree was not built on (called out of bag) is passed through the tree to obtain a classification. After this is done for all trees, the overall classification of the samples is determined from a majority voting among all trees [38, 37].

2.4.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are based on designing a system of artificial neurons. An artificial neuron consists of a weighted summation of inputs from the data variables or other neurons and an activation function. If the weighted sum exceeds the threshold given by the activation function of choice, an output is provided. Examples of activation functions are Sigmoid function, Hyperbolic tangent function, Rectified linear unit function and the Softmax function. An artificial neural network consists of an input layer, hidden layers and an output layer. The input layer consists of the actual variable values in the data set, and in the hidden layers, we can find the system of neurons. The first hidden layer takes input from one or more variables in the input layer and provides an output which can be passed on to one or more neurons in the next hidden layer, and so forth. In each hidden layer there can be several neurons, but these do not interact with each other. In the output layer, an overall prediction is produced from the output of the neurons in the last hidden layer [37, 7].

2.4.3 Support Vector Machines

The Support Vector Machines (SVM) learning algorithm is a soft margin classifier. It constructs a decision boundary to separate the classes by maximising the margin. The margin is defined as the distance between the decision boundary and the samples (in the training data set) that are closest to the decision boundary. These samples which are closest to the decision boundary are called the support vectors. However, in construction of the decision boundary, some misclassifications are allowed which is why the margin is called *soft* This prevents the method from being very sensitive to outliers [37].

2.4.4 Evaluation metrics

In classification tasks, one predicts the class of a sample with more or less success. If reference data is at hand, one can evaluate the given classifier quantitatively, for instance by cross validation. In this section, different metrics for describing success and failure of a classifier in a certain task is presented. To understand the classification metrics, the concept of *True positives*(TP) and *True negatives*(TN) as well as *False positives*(FP) and *False negatives*(FN) must be explained. In binary classification, there are two classes to be predicted. In terms of prediction of class 1, we predict that the samples either belongs to class 1 (positive prediction) or that it does

Table 2.2: This table shows an overview of classification metrics defined in terms of counts of True positives (TP), True negatives (TN), False positives (FP) and False negatives (FN). These counts can be defined for binary classification.

Name	**Definition**
Specificity (True negative rate)	$TNR = \frac{FP}{FP + TN}$
Sensitivity (Recall)	$REC = \frac{TP}{FN + TP}$
Precision	$PRE = \frac{TP}{FP + TP}$
F1 score	$F1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC}$
Accuracy (Success rate)	$ACC = \frac{TP + TN}{FP + FN + TP + TN}$
Prediction error	$ERR = \frac{FP + FN}{FP + FN + TP + TN}$

not belong to class 1 (negative prediction). A sample belonging to positive class which is predicted as positive class is called a True positive, while a sample belonging to positive class and is predicted as negative class, it is called a False negative. Correspondingly, a sample belonging to negative class which is predicted as negative class is called a True negative, while a sample belonging to negative class and is predicted as positive class is called a False positive. Metrics are often summarized in a confusion matrix, for easy comparison, where the counts of TP, TN, FP and FN are given. In this thesis, values printed in the diagonal entries of the confusion matrices are the recall values for the given class. The definition of recall and other classification metrics commonly used are summarised in table 2.2 [37].

Chapter 3

Method

3.1 Available data

The analyses performed in this thesis are based on FTIR-ATR measurements of hydrated articular cartilage sections from knee joints in human, bovine and equine cadavers. The data consists of broad-band spectra, unlike the data that will be available from the Miracle imaging probe. In this section, we describe each data set and point out differences between them. For the analyses in this thesis, the main focus is human and bovine data while equine data is used for secondary purposes.

We start by defining what is considered as a *sample* in our data sets. We consider one sample as one location on the cartilage for a given leg and a given cadaver. Firstly, it should be noted that this means we have more samples than cadavers. Furthermore, the number of samples in each data set is not the same as the number of spectra, since there are technical replicas for each sample. The number of cartilage locations and number of replicas taken vary across data sets and will be specified in the upcoming subsections in which each data set type is considered in depth. A comparison of the data sets are summarised in table 3.1. All instruments used are run in ATR mode for comparability with the Miracle probe which is based on ATR instrumentation.

For grading of cartilage damages in the samples, there are many possible systems available for articular cartilage tissue, and one of the most used assessment systems for cartilage damage is the OARSI grading. This cartilage pathology assessment system is based on histology of small extracted sections of cartilage tissue. It is a grading system based on six grades, which reflect depth of the lesion and the extent of osteoarthritis over the joint surface [40]. Other often used grading systems are for example Mankin and ICRS [41] grading systems. ICRS grading is based on 3 grades and thus shows a less nuanced cartilage assessment than OARSI grading system. For the data in this thesis, samples are graded by OARSI or ICRS. This will be specified in the upcoming sections.

3.1.1 Bovine data as a model system

One bovine data set was available for this thesis, acquired by research group at Oulu University, using a Thermo Fischer Nicolet i5 with AP pyramidal diamond probe run in ATR mode. The bovine data set can be regarded a model system, and stands out from the other available data sets in the sense that different bovine cartilage samples are subjected to different kinds of treatment. In this way, a variation of artificial damages are created. It is a high control data set, since the damage ”ground truth” is known. OARSI and Mankin grading of the samples are also available, but not focused on in this thesis. The bovine data set consists of measurements of 72 samples, distributed across 10 bovine cadaver knees. There are 3 technical replicates per sample. For each cadaver, only one knee joint is available. So we do not for instance have both right and left leg from same cadaver. In addition, as shown in figure 3.1, each knee is divided in two main sections: lateral and medial. For two of the cadavers, experiments/measurements are run on both section, while for the rest only one of the sections are used. This is done to ensure complete balance in our data set with regards to treatment groups.

There are in total 396 spectra in the data set, and two types of control measurements are available , including i) control at same location as treatment measurement (prior to treatment) and ii) control at different location. See figure 3.1 for sample locations on bovine cartilage. Locations 1 and 7 are used for ”control at different location” for respectively the medial and lateral cartilage section. In the analysis for this thesis, only control at different location is used, because it is treated as a complete separate group. After these control measurements have been discarded, the total number of spectra is 216. These spectra are divided equally between the treatment groups. There are in total 6 different treatment groups G1-G6, where G1 is the control group consisting of measurements of untreated cartilage samples. The 5 remaining treatment groups consists of two mechanical damage groups and three enzymatic damage groups. These are respectively damages induced by impact (G3), abrasion (G5), trypsin treatment (G6), collagenase 1.5h treatment (G4) and collagenase 24h treatment (G2). Each treatment happens at assigned locations on the cartilage for the lateral and medial sections, as can be seen from figure 3.1.

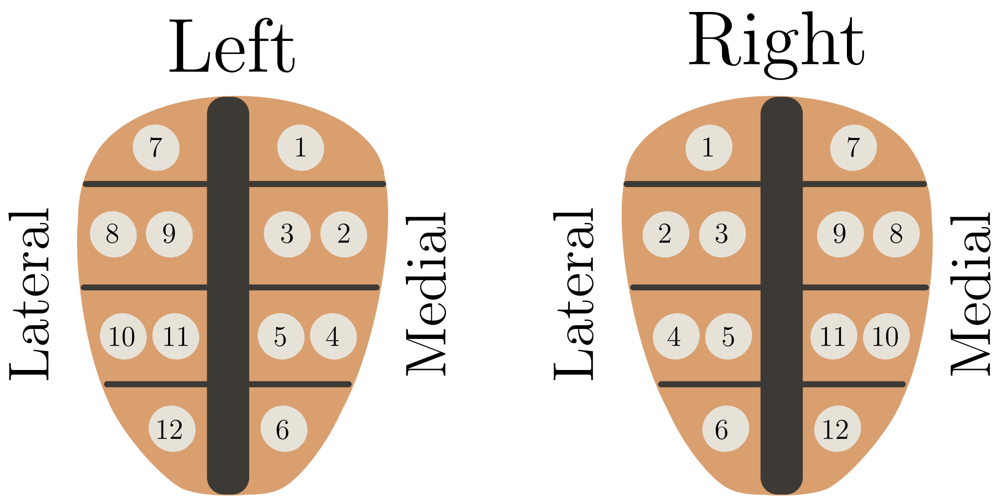


Figure 3.1: This figure illustrates the sample locations for bovine cartilage. Certain treatment groups are associated with certain sample locations. Treatment G2 (Collagenase 24 h) is applied to locations [2,8]. Treatment G4 (Collagenase 1,5 h) is applied to locations [4,10]. Treatment G6 (Trypsin) is applied to locations [6,12]. Treatment G3 (Impact) is applied to locations [3,9]. Treatment G5 (Abrasion) is applied to locations [5,11], and lastly no treatment is applied to locations [1,7] (control measurements).

3.1.2 Human data

At the time of this thesis, there were five available data sets of measurements on human articular cartilage. They are referred to as Human1-3 and Human11-12. The spectra are acquired by research groups at Ulm University (UULM), Art Photonics (AP) or Oulu University (UOULU) using different FTIR instruments. For specifications of instruments used for spectrum acquisition, the reader is referred to the summary table 3.1. For the human samples, no artificial damages are induced prior to measurements as for the bovine data set. This mean that these samples represent more realistic damages than the bovine samples. While the Human1-3 data sets contain measurements of the same two cadavers, data sets Human11-12 contain measurements of the same 9 cadavers including the two from data set 1-3. The human sample measurement locations are more detailed than for the bovine samples, as can readily be seen from figure 3.2. Data sets Human1-3 consists of measurement of 76 samples, distributed across two cadavers with ID tags KPO8 and KPO9. These data sets contains in total respectively 232, 226 and 228 spectra. Data sets Human11-12 consists of measurement of 282 samples distributed across 9 cadavers with ID tags KPO1-9. The total number of spectra are respectively 838 and 836.

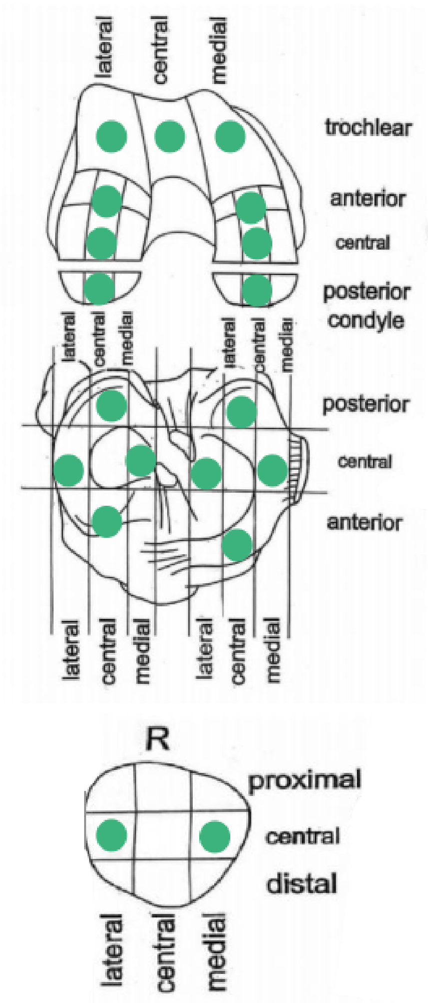


Figure 3.2: This figure shows the sample locations used on the human cartilage, for the right leg as an example. The corresponding locations are used for the left leg.

For reference data for damage degree of all human samples, an OARSI consensus grading is used. All samples are graded by three different experts independently, and later on an agreed OARSI grading is concluded. The OARSI grading is provided by reaserach group at University of Oulu (UOULU). For Human11 and Human12, some samples were ungraded. These were removed from further analysis, resulting in the final data set sizes of respectively 274 and 275 samples which yielded the total number of spectra respectively 802 and 797.

3.1.3 Equine data

There was one equine data set used in this thesis, acquired by research group at Ulm University using a Bruker Alpha 2 FTIR with Platinum ATR cell (single bounce diamond). No comprehensive analysis is executed on this data set, but it is used in tests of quality-check methods, to increase the variety of spectra for which the method performance is evaluated. There are measurements of 180 samples with three technical replicates distributed across 24 equine cadavers with ID tags H01-24. In total, there are 542 available spectra. It varies how many samples are measured from each cadaver(3-10 per horse). For equine articular cartilage damage degree, the reference data consists of the commonly used ICRS grading system.

Table 3.1: This table shows an overview of the available data sets for this thesis by summarising data set ID, how many cadavers were available in the data set, the total number of samples, the FTIR-ATR instrument the data set was acquired with and the type of available reference data.

Data set ID	**Cadavers IDs**	**Num. samples**	**Instrument**	**Reference data**
Bovine1	KBOV1-10	72	Thermo Fischer Nicolet i5 with AP pyramidal diamond probe	Treatment groups, OARSI
Human1	KPO8-9	76	Bruker Alpha 1 FTIR with Platinum ATR cell (single bounce diamond)	OARSI
Human2	KPO8-9	76	Bruker Matrix FTIR with AP Fiber probe head (diamond pyramid tip)	OARSI
Human3	KPO8-9	76	Bruker Alpha 1 FTIR with Platinum ATR cell (single bounce diamond)	OARSI
Human11	KPO1-9	282	Bruker Alpha 1 FTIR with Platinum ATR cell (single bounce diamond)	OARSI
Human12	KPO1-9	282	Bruker Alpha 2 FTIR with Platinum ATR cell (single bounce diamond)	OARSI
Equine4	H01-24	180	Bruker Alpha 2 FTIR with Platinum ATR cell (single bounce diamond)	ICRS

3.1.4 Balance of reference data

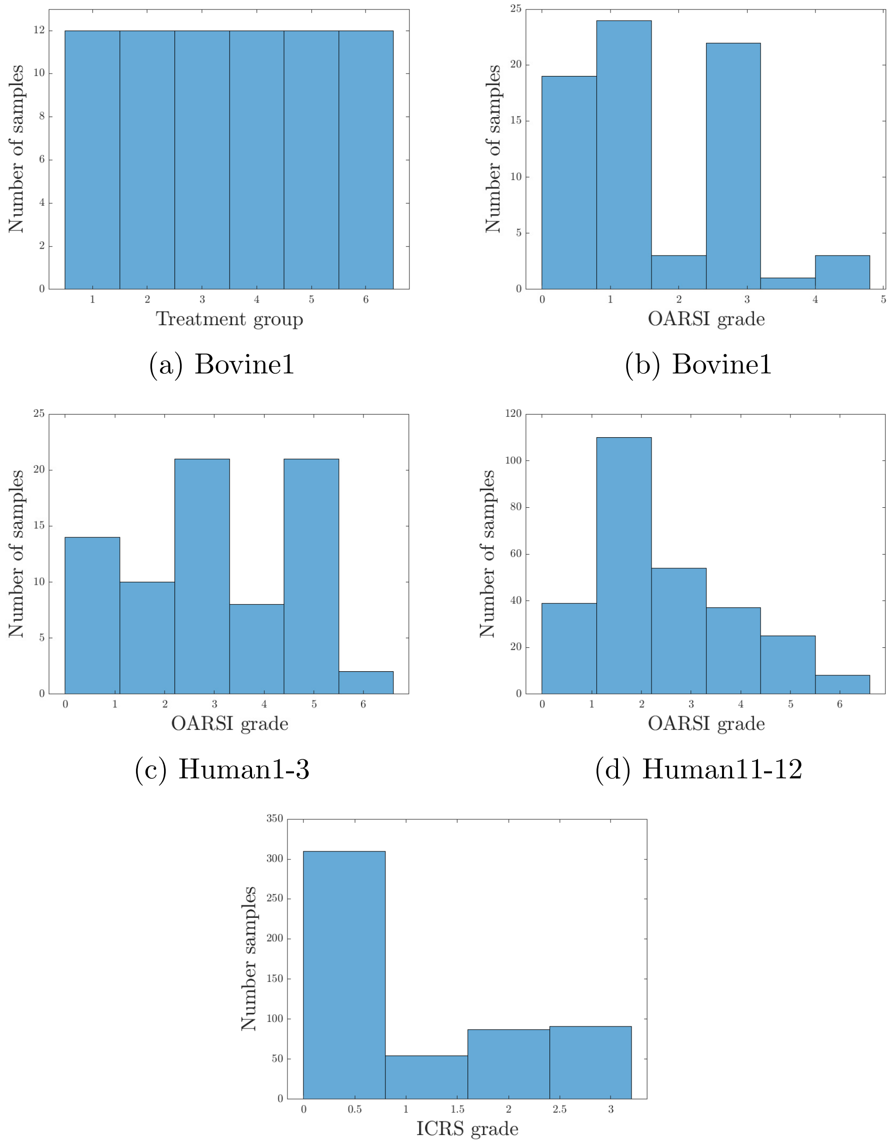
For classification tasks, balance in the data set with respect to reference data is important. In this thesis, the reference data are labels we can attach to the spectra which describes the damage degree of the cartilage tissue. In Fig. 3.3 we can see the distribution of samples with respect to the corresponding classes and treatment groups for the available data sets. The bovine data set is balanced by design with respect to treatment groups. In general, one note to make is that there are few OARSI grade 6 (high damage) samples.

3.2 Thesis pipeline

This section provides an overview of the data analysis steps and investigations making up this thesis. As described in the introduction, the sub goals was to (i) explore interferent and measurement variability in broad-band spectra, (ii) establish routines for detection of low quality broad-band spectra, (iii) use only selected wavelengths from the broad-band spectra (the wavelengths that were selected for the QCL lasers) and investigate preprocessing strategies based on only few wavelengths, (iv) to suggest preprocessing strategies for data with few wavelength channels, and finally (v) to simulate a data set based on the knowledge about interference effects from broadband spectra and use the simulated data set for validation of the suggested preprocessing strategies.

In this pipeline, all broad-band data sets first went through a quality check, where general quality measures such as signal to noise ratios and instrumental interferences were mapped. The quality check was concluded by running through developed routines for detection of spectra with low cartilage signal. These routines are explained further in the following subsection 3.3. Subsequently, an EMSC investigation was executed. This firstly included an investigation of different EMSC models for continuous broadband spectra, which resulted in the conclusion of an EMSC model which were applied for all broad-band spectra. In extension of this, EMSC for continuous spectra and EMSC correction for only the seven wavenumbers emitted by the QCL lasers developed in the Miracle project, were compared to demonstrate stability of EMSC correction for such data. Based on experience from this EMSC investigation, suggestions for preprocessing strategies for seven wavenumber channels data were made.

Subsequently, healthy and diseased cartilage spectra were simulated by a



(a) Bovine1

(b) Bovine1

(c) Human1-3

(d) Human11-12

(e) Equine4

Figure 3.3: These plots show the distribution of samples across reference data for the available data sets. We show (a) the distribution of samples across different treatment groups for Bovine1, (b) the distribution of samples across OARSI grades for Bovine1, (c) the distribution of samples across OARSI grades for Human1-3,(d) the distribution of samples across OARSI grades for Human11-12 and (e) the distribution of samples across ICRS grades for Equine4.

PCA based method. Data set variability of experimental broadband spectra Human12 with respect to damage degree was thus exploited. This simulation approach is described in more detail in subsection 3.4. Subsequently, the simulated data set was used to validate the suggested preprocessing strategies by classification using Random Forest (RF), Partial Least Squares Discriminant Analysis (PLS-DA), Artificial Neural Network (ANN) and Support Vector Machines(SVM). For Random Forest 150 trees was used. For SVM a polynomial kernel was applied, and ANN used 10 neurons with the Softmax activation function. Since the data set is simulated and do not distinguish between replicates and cadavers, cross validation with 20 random folds was used. It must be noted that for cross validation tasks, correction was done prior to, and completely separately, from the classification. As a remark, the conceptually correct method it is to perform correction of each test set separately in the cross validation. However, it is not expected that this will greatly effect the results. As a final step in this thesis, the simulated data set was used to investigate the influence of water vapor on classification performance of healthy and diseased samples.

3.3 Detection of Low absorbance spectra

Three methods for detection of spectra with low cartilage signals are tested for broad-band spectra. The three approaches are based on (i) absolute absorbance levels, (ii) derivative absorbance levels and (iii) exploitation of EMSC. The approaches are presented on a conceptual level in the following sections. All exact cutoff limits used for the respective methods are discussed in the Result and discussions section.

3.3.1 Absolute absorbance approach

This approach sets a criterion for absolute absorbance levels A in the spectra in the region 1100 - 1400 cm^{-1}. The applied definition of absolute absorbance is,

$$Abs = Max(A|_{1100-1400\text{cm}^{-1}}) - Min(A|_{1100-1400\text{cm}^{-1}}) \tag{3.1}$$

, which is in accordance with the methodology in the Opus quality check developed by Bruker [42]. Before calculation of the absorbance from equation 3.1, MSC-L correction is run with a normalised mean as reference. Thus a simple criterion is set for categorisation as spectrum with low absorbance signal, by $Abs < limit$, where the limit is tuned.

3.3.2 Derivative absorbance approach

This approach is analogous to the absolute absorbance approach. It sets a criterion for the 1st derivative of absorbance levels, $\dot{A}$, for the spectra in the region 1100-1400 cm^{-1}. The applied definition of criterion metric for this approach is thus,

$$Abs = Max(\dot{A}|_{1100-1400\text{cm}^{-1}}) - Min(\dot{A}|_{1100-1400\text{cm}^{-1}}) \tag{3.2}$$

, which is also in accordance with methodology in the Opus quality check developed by Bruker [42]. The derivative is found by applying Savitsky-Golay method. Before calculation of the criterion metric from equation 3.1, MSC-L correction is run with a normalised mean as reference. Thus a simple criterion is set for categorisation as spectrum with low absorbance signal, by $Abs < limit$, where the limit is tuned.

3.3.3 EMSC approach

This approach applies a different methodology than the two former low-cartilage-signal detection methods. Here MSC-L correction is run for spectra by i) using a normalised water spectrum as reference and ii) using a normalised mean as reference. From each of these correction methods, there are residuals from the model fitting, and the Root Mean Square Error (RMSE) is calculated from these (for each spectrum) in the region 980 - 1500 cm^{-1}. We call the two RMSEs respectively $RMSE_w$ and $RMSE_m$. The criterion metric of this approach is thus,

$$RMSE_{diff} = RMSE_w - RMSE_m \tag{3.3}$$

, which is expected to be a negative value if the spectrum is more similar to the water spectrum than the mean spectrum. The categorisation of a spectrum as low-cartilage-signal is thus $RMSE_{diff} < limit$, where the limit should be some negative value tuned for the specific use.

3.4 Simulation method

In this thesis, we simulate a data set based on Principal Component Analysis (PCA). In this section, the simulation is explained on a conseptual level, and it is assumed that the reader is familiar with the statistical method of PCA. Otherwise the reader is referred to litterature such as [39]. The simulation method exploits variations in experimental broad-band data sets and establishes a simulated data set of healthy and diseased spectra. The ground idea is to run PCA on an experimental data set of choice to obtain loadings (principal components) and scores which carry information about the spectral variations in healthy and diseased cartilage groups. The loadings are subsequently recombined with new scores, which are drawn from a normal distribution defined by the mean and standard deviation of scores from the experimental data set. Moreover, group specific perturbation by physical effects are added in the simulation by using estimated parameters from Extended Multiplicative Signal Correction on the experimental data set. Thus, the result is one unperturbed simulated data set $\tilde{X}_{pure}$ and one perturbed version of the same data set $\tilde{X}_{app}$. The approach is described in more detail in the following paragraphs.

Firstly, we note that when PCA is applied on a data set X of spectra, the spectral data block is decomposed by,

$$X = \bar{x} + TP^{'} + E \tag{3.4}$$

, where $\bar{x}$ is the mean spectrum, T is the score matrix, P is the loading matrix consisting of orthogonal components (principal components), and E is the residuals matrix. This is thus the basis for the data set simulation in this thesis. The simulation approach is summarized in Fig. 3.4. First, MSC-L correction was applied on the full data set X. After correction, a set β of estimated MSC-L parameters were obtained for each spectrum in the data set. The parameters belonging to respectively group healthy and group diseased were put in separate pools β_i, from which mean and standard deviation was calculated. This formed the basis of two separate normal distributions for the MSC-L parameters, representing group specific physical effects in spectra. These distributions were saved for later perturbation of simulated data set. Subsequently, the full MSC-L-corrected data set block X_{corr} were further altered by setting irrelevant absorbance bands in region 1780 - 2600 cm^{-1} to zero by applying a window function based on Tukey [43]. The data set was split into healthy and diseased categories. From each of these data set groupsof experimental data, denoted by i, new healthy and diseased groups were simulated by

1. Running PCA to find scores (T_i) and loadings (P_i) corresponding to equation $X_i = \bar{x} + T_i P_i^{'} + E_i$.

2. Calculating mean μ_{T_i} (i.e 0) and standard deviation σ_{T_i} of scores T_i for A number of loadings, where A is the number of loadings chosen to be included in simulation.

3. Drawing new scores $\tilde{T}_i$ for each loading included in simulation (1:A) randomly from respective normal distributions found in experimental data set, $\tilde{T}_i \sim N(\mu_{T_i}, \sigma_{T_i})$. The random drawing has a feedback loop which is activated if scores higher than maximum or lower than minimum scores obtained in experimental data set are drawn. This is to prevent very unrealistic score values being drawn.

4. Recombining the A first loadings of P_i from experimental data set with newly drawn scores ($\tilde{T}_i$), in accordance with equation 3.4.

After the recombination of scores and loadings for healthy and diseased groups respectively, the groups are merged into one data set again, and a new MSC-L correction is run on the data set. This is done to make sure no artificial physical effects were created by the random recombination of loadings in the simulation. If this is the case, it would not be a pure absorbance spectrum, and the high control environment the simulation shall provide with respect to physical interferents would be compromised. The resulting data set is the final simulated pure absorbance spectrum found by merging the group-wise simulated pure absorbance data sets,

$$\tilde{X}_{pure,i} = \bar{x}_i + \tilde{T}_i P_i' \tag{3.5}$$

Subsequently the the simulated pure absorbance data set is perturbed by group specific MSC-L parameters $\tilde{\beta}_i$ drawn from the distributions calculated from the experimental data set. White noise vectors w is also added by randomly drawing from a uniform distribution with level similar to experimental data set (not group specific). The resulting data set is a simulated apparent absorbance spectrum,

$$\begin{aligned} \tilde{X}_{app,i} &= f(\tilde{X}_{pure,i}, \tilde{\beta}_i, w) \\ \tilde{\beta}_i &\sim N(\mu_{\beta_i}, \sigma_{\beta_i}) \\ w &\sim U(-b, b) \end{aligned} \tag{3.6}$$

, where [-b, b] is the chosen range from which noise levels are drawn. The physical effects perturbation is achieved by first multiplying the obtained pure absorbance spectra with the newly drawn multiplicative parameters. Subsequently for the baseline effects, model vectors from the MSC-L which were applied on the experimental data set, is reused and multiplied by the new drawn baseline parameters. Lastly the white noise drawn is simply added to the spectra. Thus, the result is two corresponding versions of the simulated data set $\tilde{X}_{pure}$ and $\tilde{X}_{app}$.

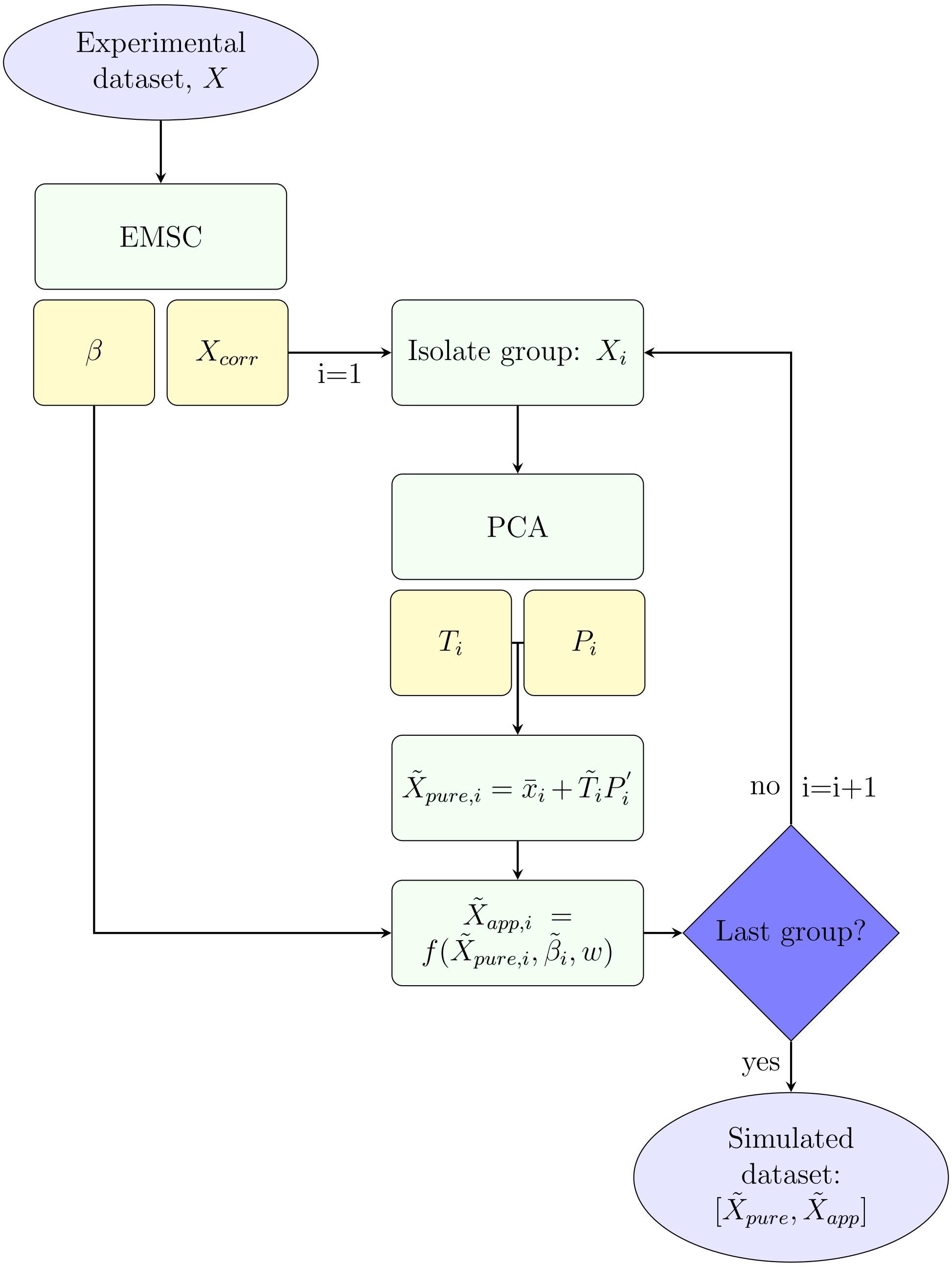

Experimental dataset, X
EMSC
β
X_{corr}
i=1
Isolate group: X_i
PCA
T_i
P_i
$\tilde{X}_{pure,i} = \bar{x}_i + \tilde{T}_i P_i'$
$\tilde{X}_{app,i} = f(\tilde{X}_{pure,i}, \tilde{\beta}_i, w)$
Last group?
no
i=i+1
yes
Simulated dataset: $[\tilde{X}_{pure}, \tilde{X}_{app}]$

Figure 3.4: This figure shows a flowchart for the PCA simulation. Blue blocks denote data sets, green blocks denote an action and yellow blocks denote results from the belonging green block.

Chapter 4

Results and discussion

4.1 Evaluation of quality and interference in broad-band spectra

In this section, we evaluate the quality of the available broad-band spectra and identify interference characteristics in the spectra. This is done by visual inspection of the raw spectra and calculations of mean noise and mean signal strength for each data set. We discuss some conditions which may marginalise the information in a spectrum. Before further processing a data set of infrared spectra, it is vital to know the quality of the spectra and to remove spectra with too low quality. Spectra may need to be removed if they do not contain relevant information or if the relevant information is marginal. In many cases it is possible to use preprocessing strategies to enhance the relevant information in the spectra and to remove non-relevant effects. Whether information in infrared spectra is relevant or not, depends on the purpose of the infrared analysis. Therefore, it depends on the final goal of the analysis if spectra will be considered as *high quality spectra* or not. Since one of the goals of the thesis is to create a realistic simulated data set for IR spectra of cartilage tissue, we aim at establishing a set of nearly pure absorbance spectra that can be used as a starting point for the simulation study. To obtain nearly pure absorbance spectra we wanted to select only high quality spectra from the measured spectra. All physical and scatter effects can be added later to the pure absorbance spectra in a controlled way for future studies of how they effect the classification for broad-band spectra and for spectra with only a limited number of wavenumbers e.g. when a number of QCL lasers are used such as in the Miracle project.

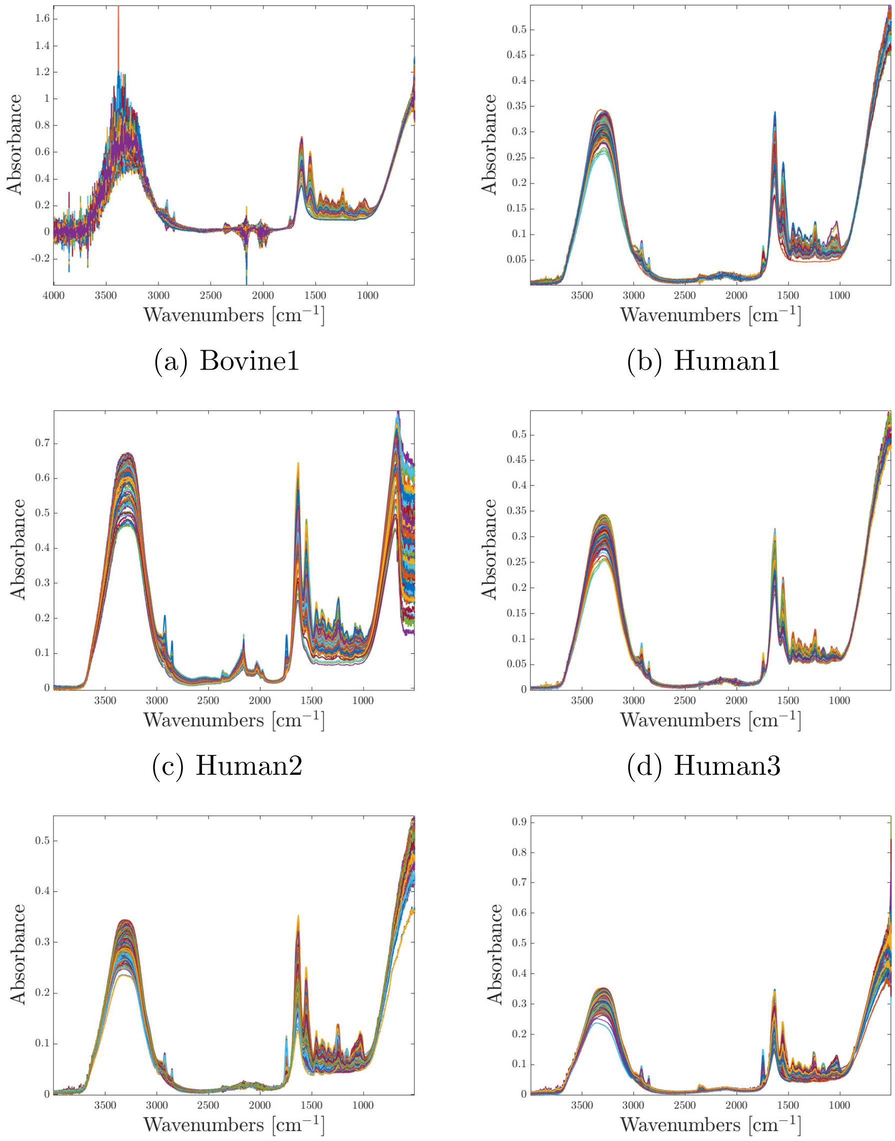
As a starting point, we want to inspect raw spectra from the different data sets visually. Different data sets of Bovine and Human samples obtained from different research groups in the Miracle project are presented in figure 4.1. In addition to these, one data set was available for equine samples. The raw equine data set is included in the appendix (Fig. 1). As there are many equine spectra with very high absorbance in the carbohydrate/phosphate region and there was a very high variation in this region which could not be explained, we have not focused on it in further analysis. For the remaining data sets shown in Fig. 4.1, we observe that the data sets obtained a large variety of cartilage and interferent features. We can for instance see that the Bovine1 data set (top left) contain high amplitude noise features in the region 3000 - 4000 cm^{-1} in comparison to the other data sets, and we observe across data sets a high absorbance variations in region 1850 - 2300 cm^{-1}. For instance the Bovine1 data set has again very disturbing characteristics in this region, while data set Human2 has more defined absorbance characteristics. The remaining data sets contain only minor absorbance in this region. In addition, we observe varying absorbance in region 2300 - 2400 cm^{-1} consistent with carbon dioxide for all data sets. For example the CO2 signal is stronger in Human12 data set than Human11 data set. On a related note, we see varying signal of water vapor. This can most clearly be seen by inspecting region 3700 - 4000 cm^{-1} for all data sets. For Human2, the water vapor signal is nearly non-existent, while for data sets such as Human11 and Human12, water vapor is clearly seen. There can also be seen variability in the cartilage signal itself, as wee can see by inspecting the fingerprint region. One main observation is that the levels and ratio of the levels of the absorption peaks at 1032 cm^{-1} and 1080 cm^{-1} varies internally in data sets and between data sets. For Human11 and Human12, the absorbance in the carbohydrate region for many of the spectra is dominated by the peak at 1032 cm^{-1}, while for data sets Human2 and Human3, the ratio between the two peaks are closer to one for all spectra. In addition, we see variability in the amount of signals from liquid water versus cartilage signals, which is evident from the variations in band height in region 3000 - 3500 cm^{-1}. These observed interferences and variations are commented in more depth in the following sub sections.

Furthermore, it is evident that several data sets contain some spectra that have very low absorbance in the fingerprint region, where we expect strong signals from cartilage, while the typical bands for water are strong. This can for instance clearly be seen for instance for Bovine1 (top left) and Human1 (top right) data set. We assume that the low cartilage signal in the spectrum and the high signal from water is due to the pressure used when the probe is put in contact during measurements. The pressure might have been too low. The spectra may contain too little information about the cartilage and consequently in the perspective of this thesis we consider these spectra as low-quality spectra. Since we apparently can obtain completely flat spectra in the region 1000 - 1500 cm^{-1} due to too low pressure, we expect any degree of peak weakening in the same region when different probe pressures are used. This is in agreement to what we observe in the figure 4.1, particularly for the data sets Bovine1 and Human2. Since such spectra are characterized by having much lower, but in general the same informative peaks as the rest of the spectra, it can further be hypothesised that spectral correction by EMSC can standardize these spectra. Therefore, provided that the difference from the reference spectrum, for which we use the average spectrum, is not too big, we can use the spectra. However, if spectra are nearly flat in the region 1000 - 1500 cm^{-1}, we discard the spectra because it is very different from the average spectrum. In the further paragraphs, we will consider the quality of spectra in grater detail by comparing noise levels in the experimental data sets and running developed detection algorithms for low cartilage signal spectra.

4.1.1 Comparison of noise levels

In this section we compare the noise levels in the broad-band experimental data sets, and discuss how this may effect our decision for which data set to base simulation on. As can be seen from table 4.1, the bovine data set has considerable higher noise levels than the Human data sets and therefore lower signal to noise ratios. Data sets Human1-3 have the highest signal to noise ratios, but while this is the case, Human11 and Human12 are much larger data sets. This is prioritized, and we choose therefore Human12 data set as a base for the simulation, since this data set have higher signal to noise ratios than Human11. The Bovine1 data set has particularly high noise levels, but it has approximately the same level of AmideII/Noise ratio, which indicate that the cartilage signal is still comparable to the other data sets. The Equine4 data set has also high quality with respect to noise, but as there are many equine spectra with very high absorbance in carbohydrate/phosphate region and there was a very high variation in this region which could not be explained, we have not focused on it further. However, equine and other data sets are used for illustrations and tests of robustness of low-cartilage-signal detection methods precisely because they have a large variability in quality.

Additionally, it is observed in the Fig. 4.1 that the Bovine data has in general higher absorbance levels than the other data sets. This means that



(a) Bovine1

(b) Human1

(c) Human2

(d) Human3

(e) Human11

(f) Human12

Figure 4.1: In this figure, we show raw cartilage spectra of human and bovine data sets.

Table 4.1: This table shows the mean noise levels and signal to noise ratios for all available data sets. Noise is defined as the difference between maximum and minimum derivative in region 1800-1850 cm^{-1}. *It should be noted that the region used to calculate this noise value can include some rotational transitions from water vapor. Thus the term noise is slightly misleading.

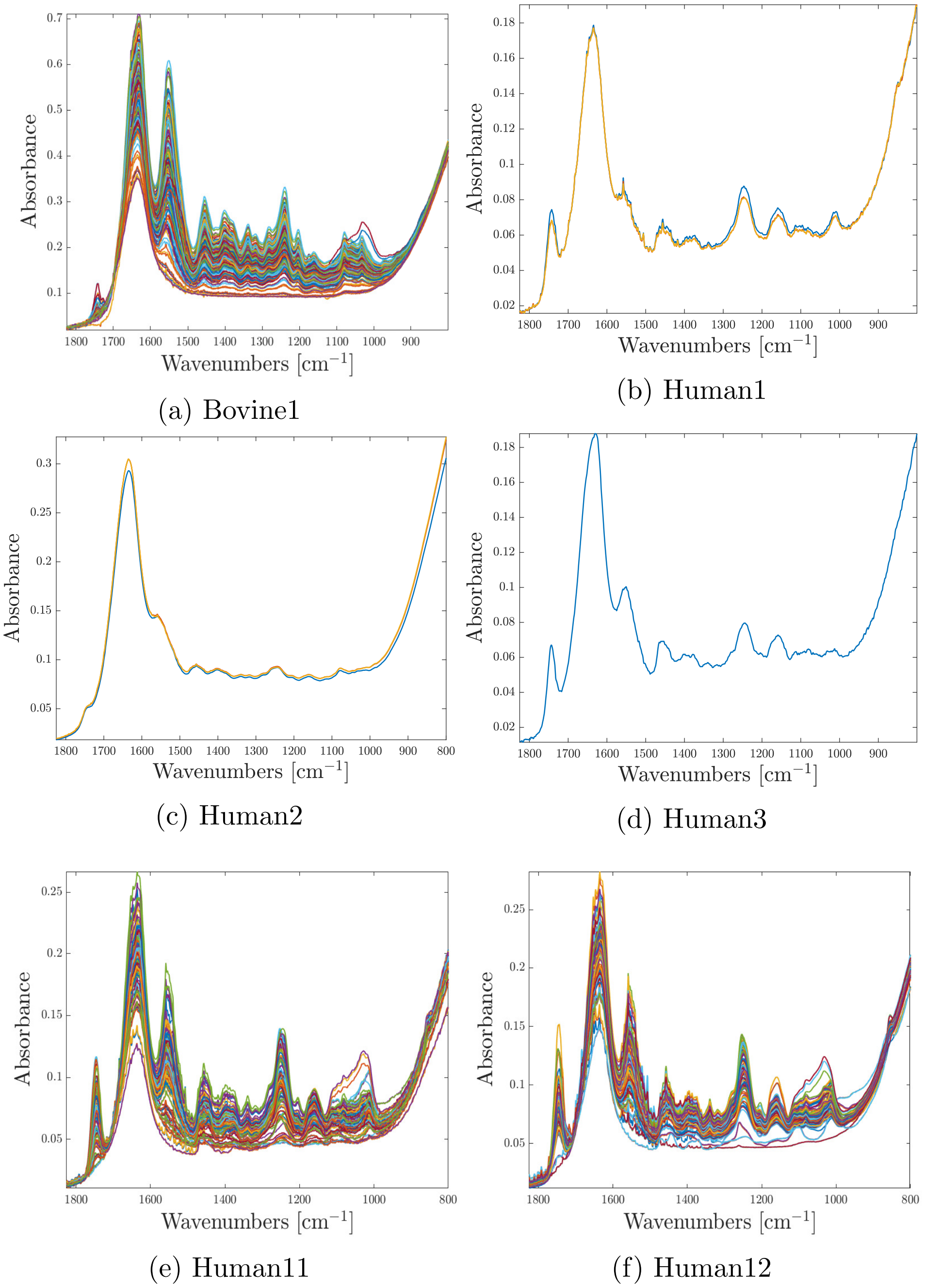
	noise*	**AmideI/noise***	**AmideII/noise***
Bovine1	0.000112	48.7	27.5
Human1	0.000068	95.9	32.0
Human2	0.000094	94.5	36.0
Human3	0.000068	95.4	31.5
Human11	0.000086	73.90	23.0
Human12	0.000089	78.7	26.7
Equine4	0.000074	95.7	37.0

when cutoff limits for distinction between high and low noise spectra are chosen, they should not be chosen based on absolute noise values. To strive for methods generalisation across different types of instruments, we identify high noise spectra, by using signal-to-noise ratios. Criteria for categorisation as *high noise spectra* is investigated, including i) AmideI/noise $<$ 50 and ii) AmideII/noise $<$ 10. *Noise* is defined as the difference between maximum and minimum derivative in region 1800 - 1850 cm^{-1}. This methodology is in correspondence with the Opus quality test designed by Bruker [42]. The region 1800 - 1850 cm^{-1} is chosen because it avoids the absorption bands that are clearly present in the spectra around 2000 - 2400 cm^{-1}, and is the region free of any broad absorbance bands which is closest to the fingerprint region. The quality of the fingerprint region is our main concern, and it is assumed that using the disturbance level here is the best possible measure for the disturbance in the fingerprint region, at least in the perspective of comparison between different spectra in a data set. It should be noted, however, that in this region bands associated with rotational transitions in water vapor are expected. Thus, in practice it may be a measure of water vapor disturbance of the signal. The spectra shown in Fig. 4.2 are identified as *high noise spectra* in accordance to criteria i), and spectra shown in Fig. 4.3 correspondingly for criteria ii). As we see, most bovine spectra are categorised as high noise, using the global criteria i) and ii) for all data sets. It is thus clear that we cannot exclude spectra based on criteria i) and ii) for the bovine data set. We can also observe that the no-cartilage-signal spectra (visually looks like water spectra) are categorised as high-noise spectra by criteria B (AmideII/noise). It seems like spectra with no cartilage signal have low amide signals while noise levels are in the same order as the other measurements. Since these are also assumed to be uninformative about cartilage tissue and we wish to detect them and remove them, this is a useful observation. In the Miracle imaging system, we might want to detect water spectra, and provide a feedback if the measurement lacks cartilage signal, encouraging the surgeon to remeasure. However, it cannot be trusted that low-absorbance spectra will always be noisy (i.e. for future data). Consequently, tools should be developed specifically for detecting such spectra. In section 4.2, some possible approaches are investigated for broadband spectra.

In the data sets, one can also see a variation of noise levels for different regions of the spectral range. For bovine data in particular, high noise levels are associated with absorption above 3100 cm^{-1}. For this reason, this region should be excluded in further analysis. For the remaining data sets, measurements below approximately 600 cm^{-1} are also associated with high noise levels. To prevent this disturbing further data preprocessing, the spectral region below approximately 600 cm^{-1} were excluded from further analysis, although the exact boundary for exclusion is individual for each data set.

4.1.2 Effects on spectra due to ATR crystal disturbance

In the region 2100 - 2400 cm^{-1} there are different types of absorption and noise levels for different data sets. It appears that data sets that were acquired with the same instrumentation have the same type of effects in this region. This suggests that this is an instrumental issue, perhaps caused by the specific ATR tip used. Comparing data sets Human1-3, which are data sets obtained from the same cadavers but measured by different instrumentation, we see clear differences in spectra in this region. It is concluded that this is indeed some sort of absorption caused by instrumentation, and in this perspective should be considered disturbance. This needs to be taken into account in further analysis, and not be mistaken for chemical variations in cartilage. In particular, it is not desirable to recreate these effects in simulated spectra, or let these effects influence the simulation in any way. In addition, high noise in this region which is observed particularly for the bovine data set, may interfere with preprocessing approaches such as EMSC when spectra are not down-weighted in this region. It is thus desirable to weight down this area in further analysis. This will be further discussed in section 4.3.2. It should also be noted that this region is also associated with a CO2 absorption band and water combination band. See section 2.2. As carbon



(a) Bovine1

(b) Human1

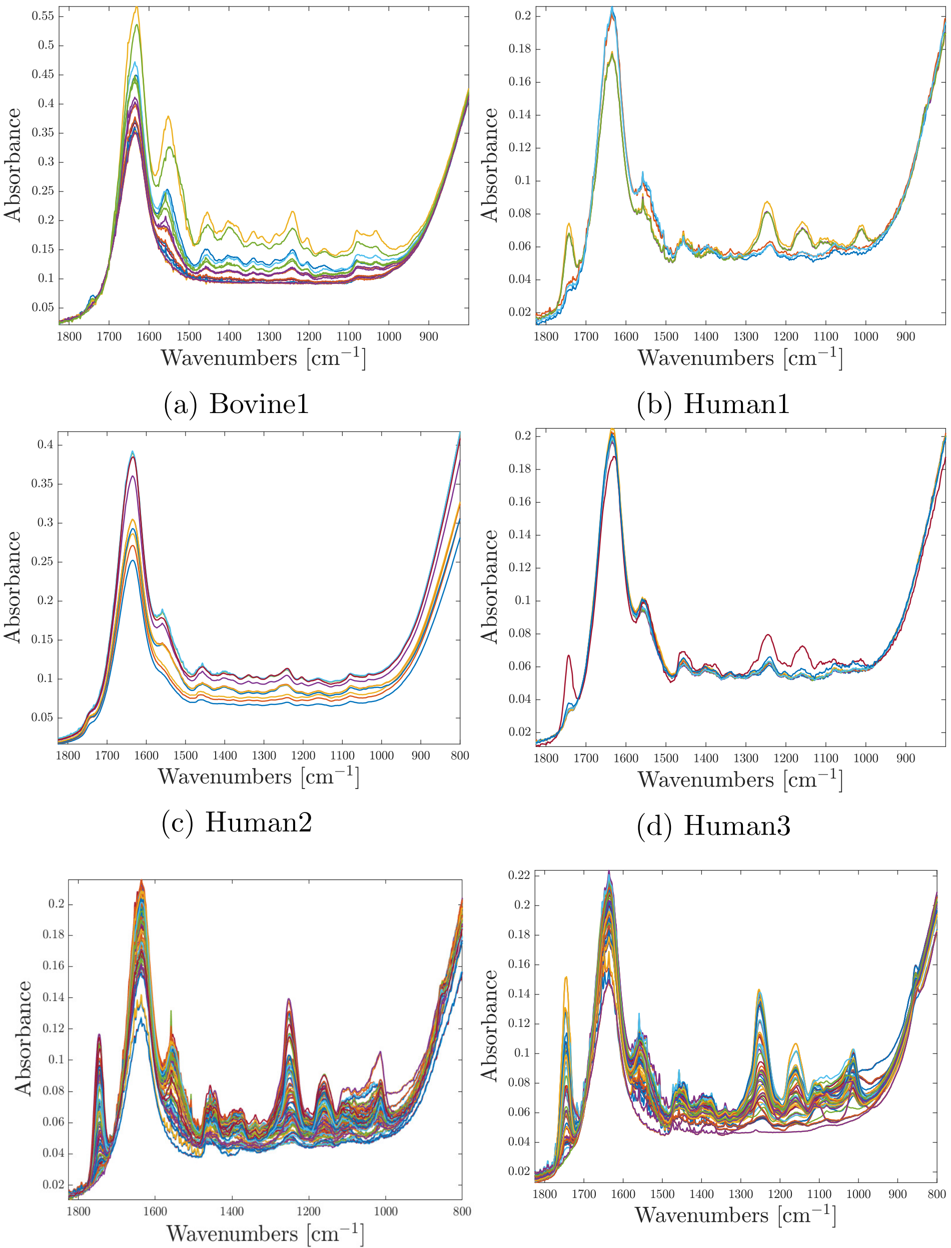
(c) Human2

(d) Human3

(e) Human11

(f) Human12

Figure 4.2: These plots show spectra which are categorised as high-noise with respect to the signal to noise criterion AmideI/noise $<$ 50, in the respective data sets.



(a) Bovine1

(b) Human1

(c) Human2

(d) Human3

(e) Human11

(f) Human12

Figure 4.3: These plots show spectra which are categorised as high-noise with respect to the signal to noise criterion AmideII/noise < 10, in the respective data sets.

dioxide is also an interferent which very likely originates from air inside the instrumentation, we do not loose any relevant information if down weighting of this region is implemented. The signal strength for carbon dioxide may depend on the instrument.

4.1.3 Classification value of spectra with no cartilage signal

In this section, spectra with no cartilage signal are commented further, and consequences of retaining such spectra for further analysis are discussed. As mentioned, one possible explanation is that spectra with low to none cartilage signal are due to non-optimal measurements when the applied probe pressure is too low. A different explanation for low cartilage signal is that such spectra are associated with high damage of cartilage. If we assume that structural changes in cartilage caused by degradation lead too a more rough surface, we would expect that water may be pooled up inside small cavities in the surface, making it more likely to obtain measurement with high content of water. While the two possible interpretations for increased water signals are so far motivated by considerations of the physical conditions of the measurement itself, we like to further discuss these interpretations by highlighting some observations from the available experimental data sets. This can give us further evidence for the one or other interpretation. Firstly, Fig. 4.4 shows a spectrum with no cartilage signal (blue) plotted together with its three technical replicates. It can clearly be seen that one of the replicate spectra is dominated by water, while all other replicate spectra are spectra with clear cartilage signals. This indicates that the experimental setup can in principal be adjusted in a way that allows to obtain spectra with high cartilage signal. While spectra used in this study are obtained under laboratory conditions, we expect that the situation when a surgeon uses the Miracle probe by hand will be comparable. It is suggested that this challenge can be solved by implementing a mechanism for maintaining constant probe pressure. In order to collect more evidence for that high water signals in spectra may be caused by the experimental setup we compare the Human2 data set (Fig. 4.1c) with Human1 (Fig. 4.1b) and Human3 (Fig. 4.1d) data sets. These data sets are measurements of the same cadavers and sample locations. It can be seen that Human2 data contains more low absorbance spectra than Human1 and Human3 data sets. This indicates as well, that differences in the water signal and the cartilage signal are due to the operator and measurement routines for the different data sets.

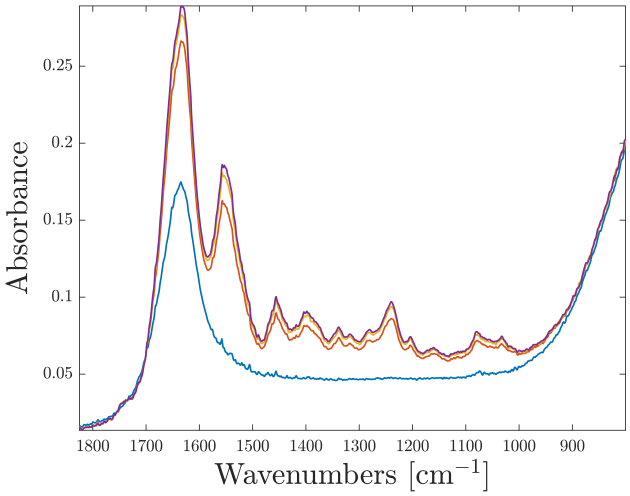


Figure 4.4: This figure shows the spectrum with no cartilage signal (blue) in the Human1 data set, plotted together with its technical replicates.

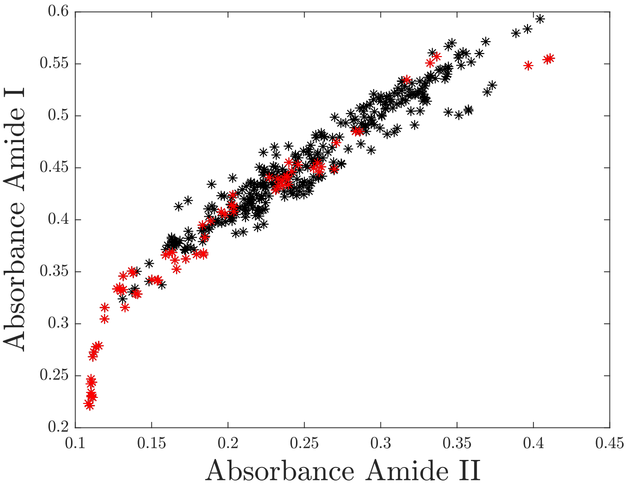


Figure 4.5: This figure shows the absorption levels of the Amide I peak plotted against absorption levels of Amide II, for the Bovine1 data set. Samples belonging to treatment group G2 (24h collagenase) is marked in red, as representatives for high degradation samples.

To investigate if it is more likely to acquire spectra with no cartilage signal for high damage cartilage samples, we have a closer look at the Bovine1 data set for which we can find a high number of low cartilage signal spectra. In Fig. 4.5 we show a plot of the absorbance level for amide I versus the absorbance level for Amide II for the Bovine1 data set. Since low-absorbance spectra are expected to have low values for both peaks, we use this plot to check if low-absorbance spectra can be associated to samples with a high degree of damage. In this plot, treatment group G2 (24h collagenase treatment) is shown in red as a representative of high damage samples. We observe that all spectra of the lowest absorbance value belong to this treatment group, which supports our hypothesis. However it should be noted that this is the most extreme of the treatment groups, and probably does not represent a very realistic damage type. We can summarize that spectra that are completely without cartilage signal, will not give any meaningful value to further classification or simulation tasks, and they should be removed. Spectra with no cartilage signal may be more likely to obtain for high degradation samples, but as seen from Fig. 4.4, membership of such a class is not guaranteed. In terms of the Miracle system, this shows that development of an automatic detection algorithm for no-cartilage-signal measurements will be vital for robust implementation.

4.1.4 Consequence of not discarding low cartilage signal spectra

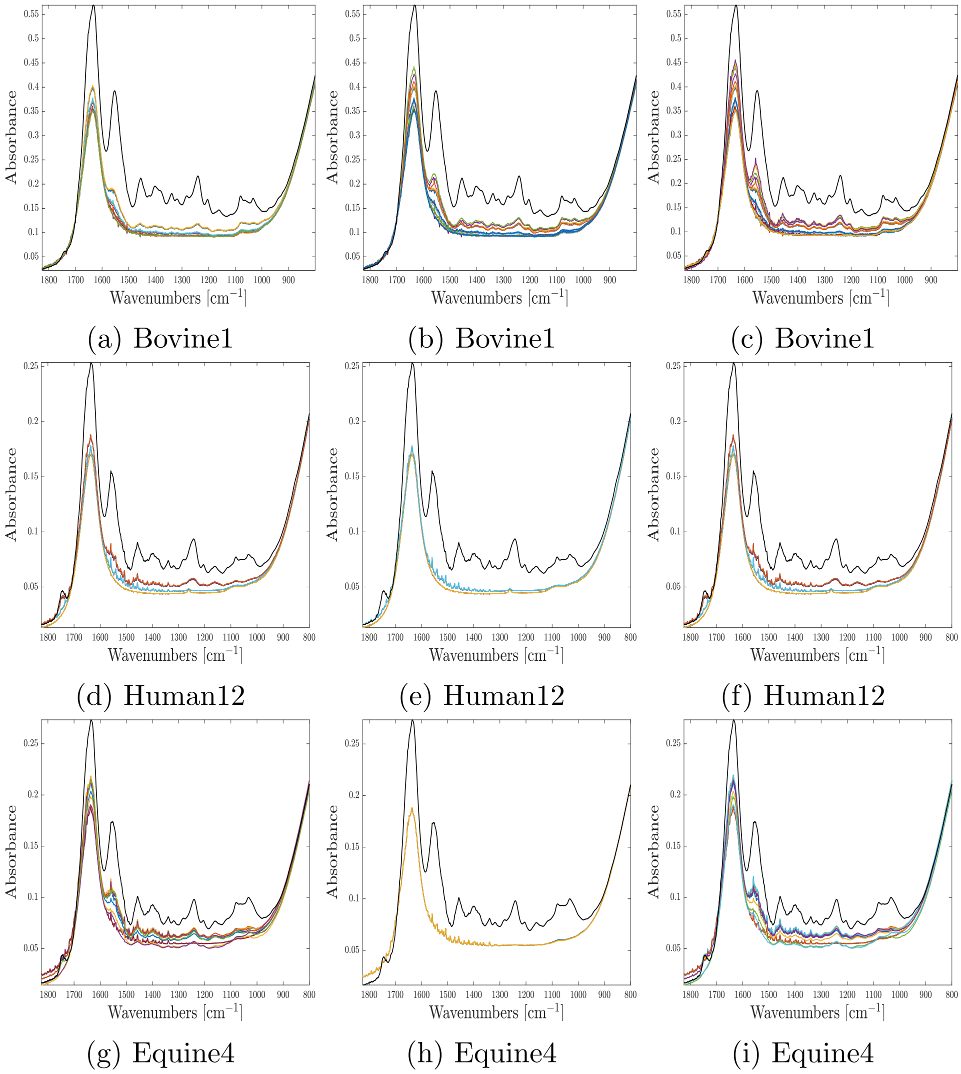
In the previous section, we found that spectra with no cartilage signal will not be of value for further classification or simulation tasks. In the following, we consider spectra with very low cartilage signals closer. We want to investigate the consequence of including such spectra in the further analysis. When low absorbance spectra are caused by too low probe pressure, there is a small space between the probe and the sample which is filled by water, i.e. synovial fluid, and consequently higher levels of absorption bands associated with water are measured. If the probe pressure is decreased further, we assume that the water-filled space between the probe and the sample becomes bigger, and that the absorption bands associated with water increase further. Simultaneously the absorption bands associated with cartilage signal will decrease because the penetration depth into the cartilage correspondingly becomes lower. For peaks which can be associated with both water and cartilage, such as the Amide I peak, these two mechanisms will mix. Consequently if the cartilage signal decreases because of lower pressure, the cartilage-contribution to Amide I levels will be lowered, but the total effect on Amide I level of this mechanism will be counteracted by the increase in water peaks which coincide with Amide I. The decrease in Amide I levels due to decreased pressure will thus be a trade off between these two mechanisms. For the peaks which are not associated with both water and cartilage, such as Amide II or any other peak in the region 1000 - 1590 cm^{-1}, there will not be a corresponding trade off. In Fig. 4.5 the absorbance levels of Amide I are plotted against the absorbance level of Amide II. Hence we plot one peak which is affected by the two trade off mechanisms against a peak which is not affected by the two trade off mechanisms. It could be argued that such a trade off mechanism may cause EMSC correction to be erroneous, by obtaining the estimation of a too low multiplicative parameter in the least squares fitting. This would mean that the multiplicative parameter does not restore the cartilage signal to its full power in the EMSC. In the analysis of this thesis, only the spectra with no cartilage signal are removed, but it is noted from this discussion, that including very low absorbance signal spectra may be a source of error, even if the spectra contain all the peaks which are associated with cartilage signal in the fingerprint region due to water disturbance.

4.2 Detection of low quality spectra

In this section, three methods for detection of spectra with low cartilage signal will be presented. The three methods are based on evaluating i) the absolute absorbance levels, ii) the derivative levels of the absorbance and iii) the residuals from an EMSC model with mean reference in comparison to an EMSC model with water spectrum reference, respectively. For more in depth description, the reader can consult section 3.3. Such an investigation across different detection strategies is also useful to motivate future ideas for how such low absorbance signals can be detected for the seven wavenumber channels data.

To illustrate robustness of each method, we set a goal for this paragraph to only detect spectra with no cartilage signal. Such spectra can be visually identified as being completely flat in the region 1000 - 1500 cm^{-1}. For applications of such detection methods, it is desirable that the cutoff value will provide an as precise separation as possible. The aim of this section is thus to tune one single cutoff limit per method for separation of flat and non-flat spectra to work across all available data sets, hence forth referred to as a global cutoff limit. Prior to running the detection algorithm, all spectra which have no cartilage signal in the data sets were manually identified. Subsequently, the global cutoff limit was tuned while inspecting results visually with the aim to detect all spectra predefined as no-cartilage-signal spectra across all data sets. Thus, the number of surplus spectra detected, presumably containing cartilage signal, can be regarded as a measure of the method's robustness and give an indication of whether automation of the detection process is feasible. We aim that the methods, for a tuned global cutoff limit, will not detect any additional spectra than the predefined no-cartilage-signal spectra.

For the three detection methods, visual tuning of global cutoff limit yielded respectively the criteria i) absolute absorbance value in region 1100-1400 cm^{-1} is less than 0.035, ii) derivative absorbance value in region 1100-1400 cm^{-1} is less than 0.00065 and iii) the difference in RMSE (of the model) in region 1100-1400 cm^{-1} between respectively an EMSC correction using a water spectrum as reference and an EMSC correction using the mean spectrum as reference is less than -0.055. In Fig. 4.6, spectra which are detected for these cutoff limits are shown for the data sets Bovine1, Human12 and Equine4 as examples. By design, all no-cartilage-signal spectra are detected. We consider the method more successful if it does not detect any extra spectra, since the global cutoff limits were tuned with this specific aim. In Fig. 4.7, we show how many non-flat spectra which were detected in additional to the completely flat spectra for each data set. We desire these to be as few as possible. Firstly we can see that none of the methods works best for all data sets simultaneously, thus all methods have some weaknesses. We can however see that across all data sets, the derivative absorbance metric (red) detect the fewest additional spectra, followed by the RMSE based method. Thus, we can conclude that the absorbance-derivative approach is the most precise and robust approach. For instance, it was observed during visual tuning of the global cutoff limit for the two other methods, that the no-cartilage-signal spectra which contained water vapor (e.g blue spectra in Human12) were the reason for the need to increase the cutoff limit, and thus leaded to detection of more non-flat spectra. The water vapor peaks are clearly observed for instance for the blue spectrum of Human12 (middle row) in the region 1350 - 1600 cm^{-1}. Evidently, the water vapor peaks can be a disturbing factor for the separation precision of the global cutoff limit. The derivative-absorbance check was thus most preferable of the three tested methods, but as mentioned none of the methods performed perfectly for all data sets, and therefore adjustments needs to be done for future data sets. All detected spectra should be confirmed by manual inspection, but the suggested methods here are useful tools to narrow down the manual inspections considerably. It should also be noted that if such an automatic detection



(a) Bovine1 (b) Bovine1 (c) Bovine1

(d) Human12 (e) Human12 (f) Human12

(g) Equine4 (h) Equine4 (i) Equine4

Figure 4.6: These plots show spectra which were categorised as spectra with no cartilage signal by three different methods, using one global cutoff limit per method. The methods used were based on absolute absorbance levels (left column), derivative absorbance levels (column in the middle) and RMSE from EMSC (right column). The global cutoff limits are tuned such that all spectra without cartilage signal are detected for all data sets. The mean spectrum is shown in black. Results are shown for only three of the data sets, including Bovine1, Human12 and Equine4.

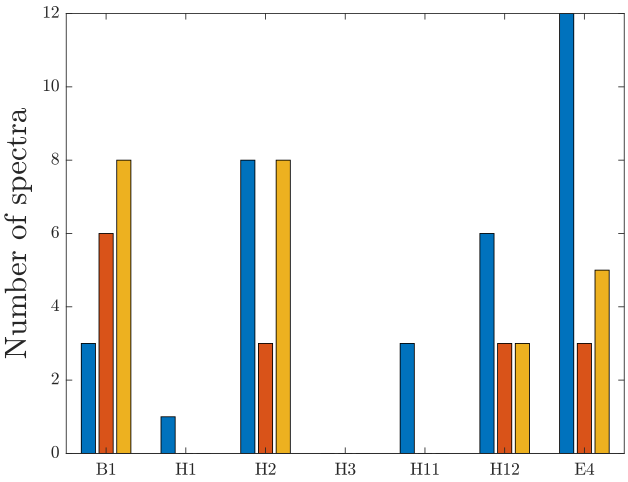


Figure 4.7: This figure shows how many spectra were detected in addition to the predefined no-cartilage-signal spectra for the global cutoff limit of the respective methods and data sets. We show results for methods based on absolute absorbance (blue), derivative absorbance (red) and RMSE from EMSC (yellow).

algorithm was to be implemented in any real system, the differences we see in absorbance levels across data sets may not be devastating, because there would be a calibration data set at hand, making sure the cutoff limits are tuned correctly for the given instrumentation.

Lastly, some notes should be made on the subject of transferability of the three methods tested for broad-band spectra to data with few wavenumber channels. Firstly, we note that the absorbance derivative approach (ii), which had the highest precision, is not a viable option for the seven wavenumber channels data because it will not be possible to calculate any derivatives based on point measurements. Furthermore, the absolute absorbance method (i) is not directly transferable either, because it is based on calculating the difference between the maximum and minimum absorbance levels in the fingerprint region. Thus, to describe the height of a peak, it relies on having one measurement point which is not situated at a cartilage peak. This is not the case for any of the seven wavenumbers chosen for the Miracle lasers. Thus, the most applicable approach for data with few wavenumber channels is the RMSE based approach (iii).

4.3 Extended multiplicative signal correction for spectral data with few spectral channels

Extended Multiplicative Signal Correction is a model based preprocessing technique utilizing Least Squares fitting of a measured spectrum to a set of model spectra including a reference spectrum and several other model components as described in section 2.3.1. For application on seven wavenumber channels data, it may be expected that the low collinearity between the quasi spectrum measurements is detrimental to the stability of the approach. How the use of a few selected wavenumbers affects the EMSC model parameter estimation compared to the situation where broad-band spectra are available is an interesting question. We performed therefore a study of the reliability of the Extended Multiplicative Signal Correction both for application to the broadband experimental data sets and for application to measurements of selected QCL wavenumbers was carried out. The objective was both to design a reliable EMSC model to use for all further corrections of the experimental spectra, and to motivate suggestions for preprocessing strategies for quasi spectra. In this section, no spectra are removed from the data sets unless otherwise specified. This is because all spectra and present spectrum characteristics represent types of readings that may occur using the Miracle probe. It is of interest to investigate these as well.

4.3.1 A simple demonstration of MSC limitations for broad-band spectra

The rationale of using infrared spectroscopy for diagnosis of cartilage damage is that healthy tissue and diseased tissue have chemical differences that show distinct chemical features in infrared spectra. Such distinct chemically different features are expected in both existing broad-band experimental data sets and in future measurements using the Miracle probe. As a simple demonstration of how preprocessing may be impacted by such chemical differences within a data set, an apparent spectrum consisting of two Lorentzian bands and a constant baseline was constructed, as shown in Fig. 4.8 (left, red). A simple multiplicative signal correction (MSC) was performed using a chemically different reference, also shown (left, black). The reference spectrum is constructed by the same Lorentzian band as the left peak of the apparent spectrum. Except for the extra Lorentzian band, the apparent spectrum differs from the reference only by a constant shift and a multiplicative constant. Using the chemically different reference, the resulting correction of the apparent spectrum is shown (right). As can be seen, the correction is not optimal.The baseline is brought below zero, thus showing that the constant baseline parameter is wrongly estimated, which implicates that the multiplicative parameter must be wrong as well. It is evident that the Least Squares method has compensated for the unmodelled chemical differences, and the correction is erroneous. This phenomena is referred to as statistic interference, and is a well known problem discussed in literature, for instance by Martens [44].

The simple example shows that with a simple correction model consisting only of a reference spectrum, the multiplicative parameter and a constant baseline, the correction of spectra which are chemically different from the reference (i.e mean in experimental data sets) may be erroneous. It possible that by increasing the model complexity, for instance with wavenumber dependent baselines, the Least-squares algorithm may fit these to the unmodelled chemical variations. In this case, more obvious disturbing effects may be introduced to the corrected spectrum and chemical information may be lost or disturbed significantly. The degree of disturbance will naturally depend on how well the model component fits to the chemical deviations from the reference, and the apparent spectrum's degree of deviation from the reference. In the next section, the degree of such behaviour is investigated carefully for the experimental data sets at hand.

4.3.2 Using EMSC approaches to correct broad-band spectra

In the previous subsection, it could be seen that statistical interference is a challenge for the multiplicative signal correction and implicitly the same issue may persist for extended versions of the EMSC preprocessing algorithm. Thus, it was desirable to map if such effects arise in the broadband experimental data sets. Literature, such as Kohler et al [4], suggests that such statistical interference can be avoided by implementing weighting of chemically inactive regions or include known absorbance bands of deviation as model components. Thus, in this section we study different EMSC-type models, and the need for weighting of chemically inactive regions was investigated as a possible remedy for statistical interference.

Firstly, EMSC correction of different complexities is run on experimental data sets. Corrected spectra are presented for the data set Human2 in Fig.

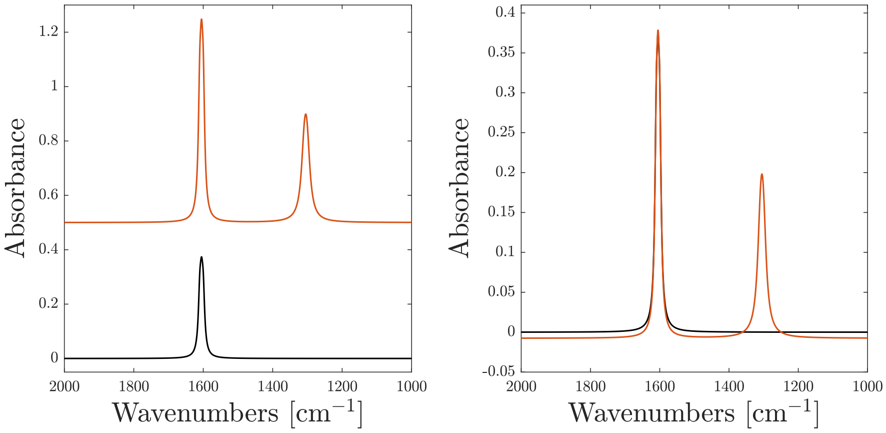
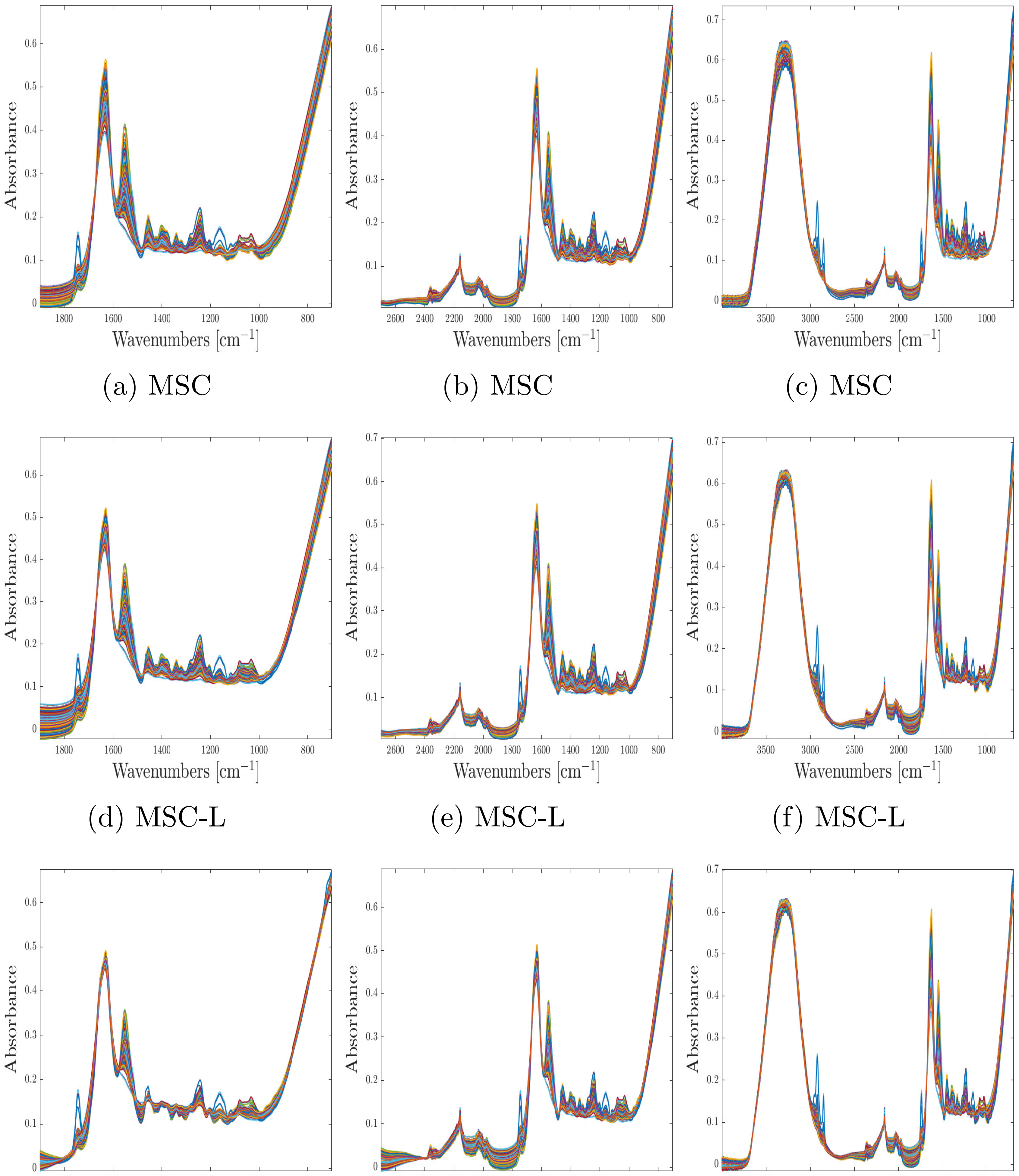


Figure 4.8: In this figure, a simulated apparent spectrum (red) and an MSC reference spectrum (black) are plotted together (left). The reference spectrum is chemically different from the apparent spectrum. The resulting MSC correction of the apparent spectrum applying this reference spectrum is shown (right).

4.9. It was also suspected that the width of the spectral range would affect the correction, thus three different ranges are tested, including 700 - 1900 cm^{-1}, 700 - 2700 cm^{-1}, and 700 - 4000 cm^{-1}. It can be seen that when using only region 700-1900 cm^{-1} spectra are not properly corrected in the region 1800 - 1900 cm^{-1}, while they should be on top of each other since this region is chemically inactive. How well the spectra are corrected depends on the EMSC model. For the EMSC model with linear and quadratic effect it is evident that the fingerprint region is not corrected as well. Particularly the correction by EMSC appears very different from the corrections by MSC and MSC with linear effect (MSC-L). It is natural to expect that by including a larger absorption free spectral region in the estimation of the EMSC parameters, the least squares method may give a better estimate for this region as well. By extending the region to 700 - 2700 cm^{-1} it can be seen that the correction works better in the chemically inactive regions for MSC and MSC-L, but for EMSC the issue is persistent. Using the full spectral range 700 - 4000 cm^{-1}, EMSC performs better. However, none of the EMSC model results in a satisfactory correction of the absorption inactive regions. Although not included in this thesis, the same was done for all available data sets, and yielded the same observations. Hence, it should be considered to implement weighting of the absorption inactive regions. It can also be made a remark that, since we see less optimal corrections for smaller wavenumber regions, we should at least expect similar effects for the measurements for the selected QCL wavenumbers in the Miracle project which are all in proximity of the fingerprint region.

Further in this paragraph, we investigate different weighting possibilities for our spectra. There are mainly two sub goals for the weighting investigation. Firstly we aim that all corrected spectra within a data set overlap in all absorption inactive regions. Secondly we aim for down-weighting of high interference regions, in essence the region 2100 - 2400 cm^{-1} which is associated with ATR crystal disturbance and carbon dioxide absorption as discussed in section 4.1.2. Therefore it is in this section studied how different weighting schemes influence the EMSC correction. Weighting up absorption inactive regions in the EMSC will allow less deviance from the reference spectrum in this region, which will promote the possibility that the corrected spectra will overlap in this region. To prevent EMSC to model interference and produce unpredictable corrections of spectra, it is desirable to weight down the region 2100 - 2400 cm^{-1}, which means that EMSC algorithm will not attempt to minimize residuals between the measured spectrum and the reference spectrum in this region. In addition to the above mentioned weighting functions, an up-weighting of the absorbance region 750 - 800 cm^{-1} is investigated. All weighting schemes are shown in Fig. 4.10. The idea of weighting up the absorbance region 750 - 800 cm^{-1} is that this may function as a standardisation of the correction across the fingerprint region of the spectrum, giving the EMSC correction of chemically different spectra firm reference points in the correction. Since region 750 - 800 cm^{-1} is part of an absorbance peak associated with water, one may argue that it would be a standardization with respect to water content, which might be advantageous since water is a source of variability in the data. To check if this indeed has value in practice, classification is run on MSC-L corrected data with and without up-weighting at 750 - 800 cm^{-1}. The confusion matrices for classification on Human2 data are shown in figure 4.12. It can be seen that up-weighting does in general not improve classification. Therefore we will not use up-weighting in the region 750 - 800 cm^{-1} in the further analysis.

It is important to note, that EMSC is also run in the quality check for the detection algorithm for no-cartilage-signal spectra. Here, the main goal is not classification, but robust detection of no-cartilage-signal spectra. Therefore, we study now closer how spectra with particularly high chemical variability such as no-cartilage-signal spectra versus spectra with strong cartilage signals are impacted by weighting of the region 750 - 800 cm^{-1}. In Fig. 4.13 one bovine spectrum with high cartilage signal (HCS) and one spectrum with



(a) MSC (b) MSC (c) MSC

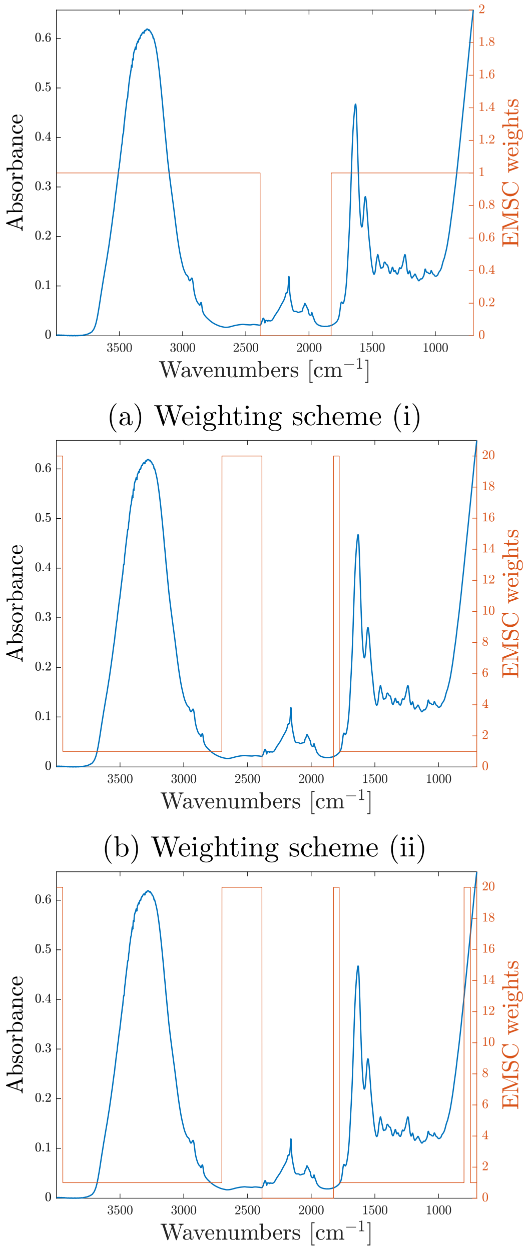
(d) MSC-L (e) MSC-L (f) MSC-L

(g) EMSC (h) EMSC (i) EMSC

Figure 4.9: These plots show the EMSC corrected Human2 data set for three different spectral regions and three different EMSC-type models. We show respectively, from left to right, regions 700 - 1900 cm^{-1}, 700 - 2700 cm^{-1} and 700 - 4000 cm^{-1}. The regions are combined with, respectively from top to bottom, correction models MSC, MSC-L and EMSC

no cartilage signal are shown using this weighting strategy. It can be seen that up-weighting of the region 750-800 cm^{-1} by scheme (iii) does not seem to make any notable differences with respect to weighting scheme (ii). Thus up-weighting of the region 750 - 800 cm^{-1} is not done for any purpose in further work. From the results in Fig. 4.13, we should note the consequences for running EMSC with down-weighting only of the region 2100 - 2400 cm^{-1} which is shown in the top row. For MSC and MSC-L, we see a erroneous elevation of the spectrum with no cartilage signal, but for EMSC it is clear that the down-weighting of this region makes it possible for the quadratic baseline effect to be fitted, introducing a large bulge in the spectra. The effect is very visible for the spectra with no cartilage signal, but it may implicate that spectra with low absorbance may also deviate enough from the mean spectrum to introduce similar effects. It is thus concluded that down-weighting should not be done alone alone. In the middle row, where up-weighting of absorption inactive regions is done simultaneous to down-weighting of region 2100 - 2400 cm^{-1}, we see that such effects are avoided, and it is concluded that for all further EMSC, weighting scheme (ii) is implemented.

Indeed we have seen in the section that a weighting scheme can solve EMSC correction challenges caused by statistical interference in the experimental broad-band spectra. We now consider which EMSC complexity to run for the final preprocessing of spectra for future classification tasks and simulation. Looking again at the middle row of Fig. 4.11, we see that when weight are implemented, the corrections across EMSC complexities visually look the same. As described in the theory section 2.2, we expect ATR spectra to have higher penetration depth for lower wavenumbers. This means that the lower wavenumber peak levels will be exaggerated, which motivates the usage of a wavenumber dependent baseline. This behaviour has been reported by other data preprocessing papers as well, such as the study by Lee [35]. Among the EMSC models we are testing, the most relevant model component accounting for such behaviour is the linear baseline. However, it is possible that the wavenumber dependence is not strictly linear. In this case, the quadratic baseline could together with a linear baseline, produce some combination baseline which in total may be more correct. To comment if this is likely, correlation plots between parameters from EMSC is included in Fig. 4.14. The correlation plot is shown for two different data sets and sample types; Human12 and Equine 4. We see that the linear (d) and quadratic (e) parameters are highly negatively correlated for both data sets, which supports the hypothesis. The quadratic baseline is also highly negatively correlated with the multiplicative parameter for both data sets. These high correlations may indicate that the quadratic baseline compete

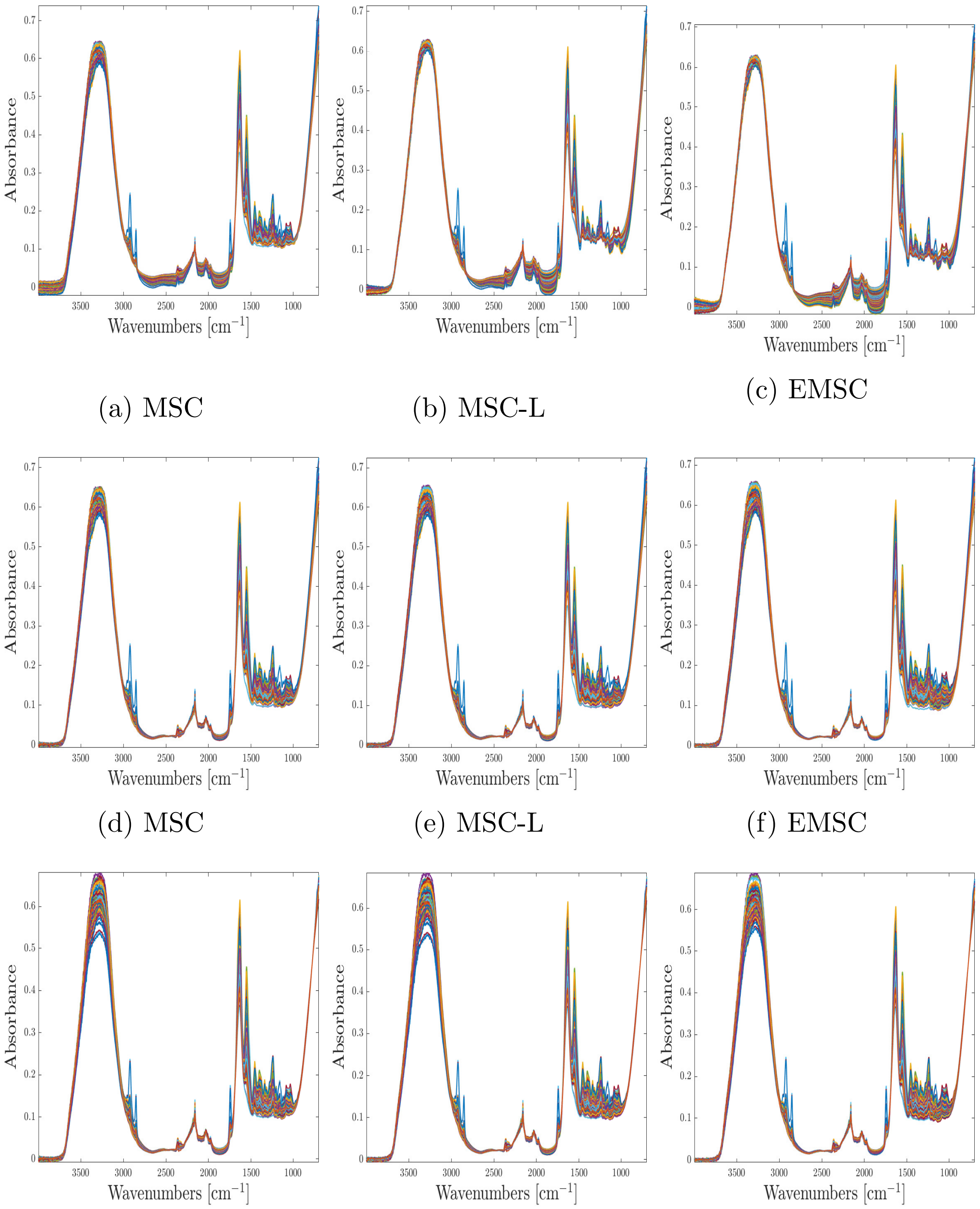


(a) Weighting scheme (i)

(b) Weighting scheme (ii)

(c) Weighting scheme (iii)

Figure 4.10: These plots show three different weighting schemes (i)-(iii) which were tested in the EMSC-type correction algorithms for broad-band spectra.



(a) MSC

(b) MSC-L

(c) EMSC

(d) MSC

(e) MSC-L

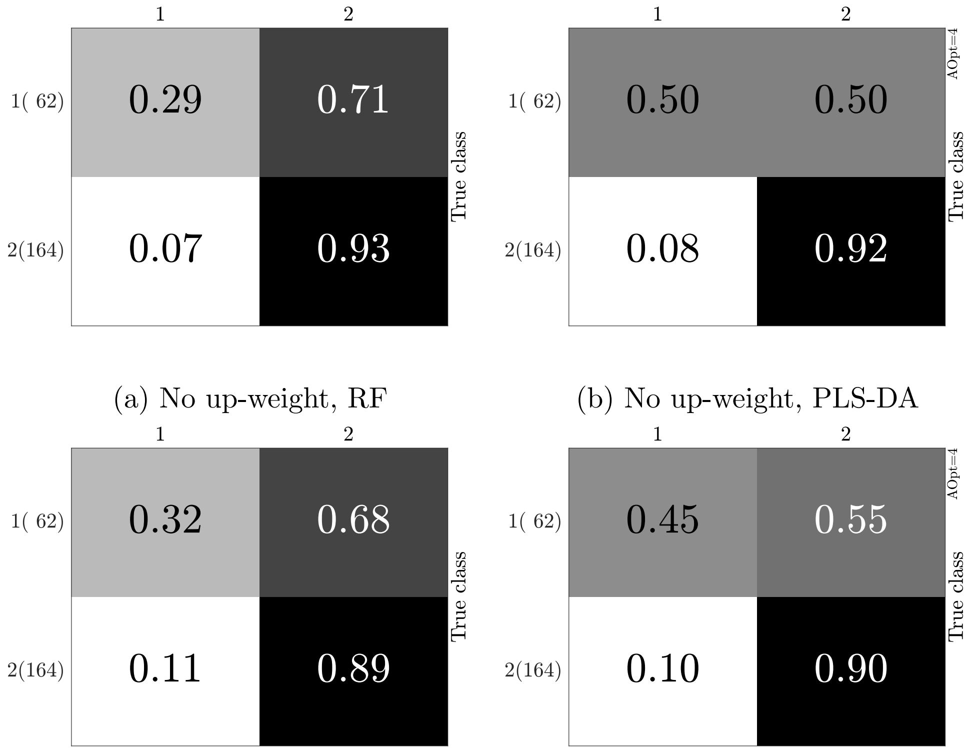
(f) EMSC

(g) MSC

(h) MSC-L

(i) EMSC

Figure 4.11: These plots show corrections of the Human2 data set for three different weighting schemes shown in figure 4.10 combined with three different EMSC models. We show, respectively from left to right, correction models MSC, MSC-L and EMSC. The models apply, respectively from top to bottom, weighting scheme (i),(ii) and (iii).



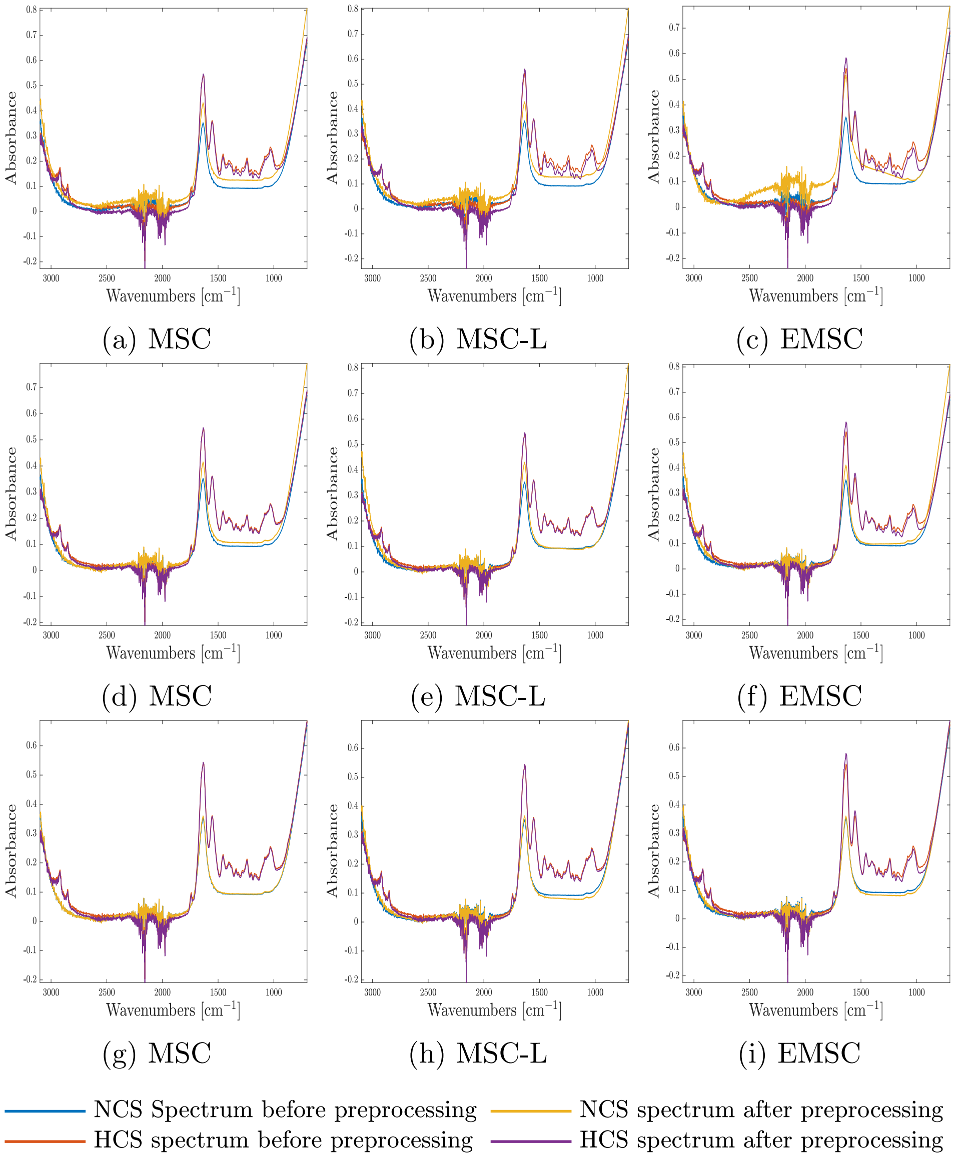
(a) No up-weight, RF

(b) No up-weight, PLS-DA

(c) Up-weight, RF

(d) Up-weight, PLS-DA

Figure 4.12: This figure shows classification impact, represented by confusion matrices, for healthy and diseased groups in Human12 by weighting up (w=20) region 750 - 800 cm^{-1}. We show confusion matrices corresponding to preprocessed data without up-weighting of the region 750 - 800 cm^{-1} (top row) and confusion matrices corresponding to preprocessed data with up-weighting (bottom row). Results are show for Random forests (left) and PLS-DA (right) for comparison. All spectra which have no cartilage signal were removed before classification.



(a) MSC (b) MSC-L (c) EMSC

(d) MSC (e) MSC-L (f) EMSC

(g) MSC (h) MSC-L (i) EMSC

Figure 4.13: These plots show different EMSC-type corrections for one high cartilage signal (HCS) spectrum and one spectrum with no cartilage signal (NCS) for weighting schemes (i)-(iii). We show corrections MSC (left column), MSC-L (middle column) and EMSC (right column), which are combined with respectively weighting schemes i (top row), ii (middle row) and iii (bottom row).

with the other effects in the Least Squares fitting in the EMSC and it is possible that this is a sign of statistical interference. This is a motivation for not including the quadratic parameter in the EMSC. Even if we consider the case that there is some trade off between the linear and quadratic effects which better explain the ATR penetration depth dependence on wavenumbers, it is desirable to avoid spreading information about the same phenomena over several parameters. For comparison, we look also at the parameter correlations for MSC with linear effect baseline included for the same two data sets. These can be seen in Fig. 4.15 that in the case the parameter correlations are very low, which indicate that our parameters are now more independent and do not explain the same phenomena. It shows that reducing the model complexity to a Multiplicative signal correction including linear baseline effect is safer. We thus conclude that for correction of the broadband spectra, we apply multiplicative signal correction with linear baseline effect.

4.3.3 Comparison of EMSC correction for broad-band spectra and 7 selected wavenumber channels

In the former sub sections, we have developed a pre-processing strategy for broad-band spectra by Extended Multiplicative signal correction. In this section we focus on preprocessing strategies for the seven selected wavenumber channels for the QCL lasers in the Miracle project. To evaluate preprocessing strategies for selected wavelength we will consider the correction of the broad-band spectra with the suggested weighting scheme of section 4.3.2 as a golden standard and compare correction strategies using only seven wavenumbers with this golden standard. For each spectrum in the broad-band data sets, we pick out the absorbance values for the seven wavenumbers. Correction by MSC, MSC-L and EMSC is subsequently run for the broad-band spectra with the suggested weighting scheme and for the corresponding seven wavenumber version of the spectrum without any weights implemented. The Root Mean Square error, $\text{RMSE}_{\text{corr}}$, between the two corrections is then calculated based on absorbance levels of the seven wavenumbers. In Fig. 4.16 (bottom) the mean value of this RMSE of correction for each data set is plotted. As seen, the correction of seven wavenumber channel data shows a clear tendency to increase the RMSE of correction for higher complexity EMSC models. For a visual aid, the correction of an arbitrary spectrum based on the full broad-band region (blue) and the corresponding correction of seven wavenumber data (red) is shown (top row), for MSC, MSC-L and EMSC. As seen, none of the EMSC complexities give satisfactory results for the seven wavenumber data.

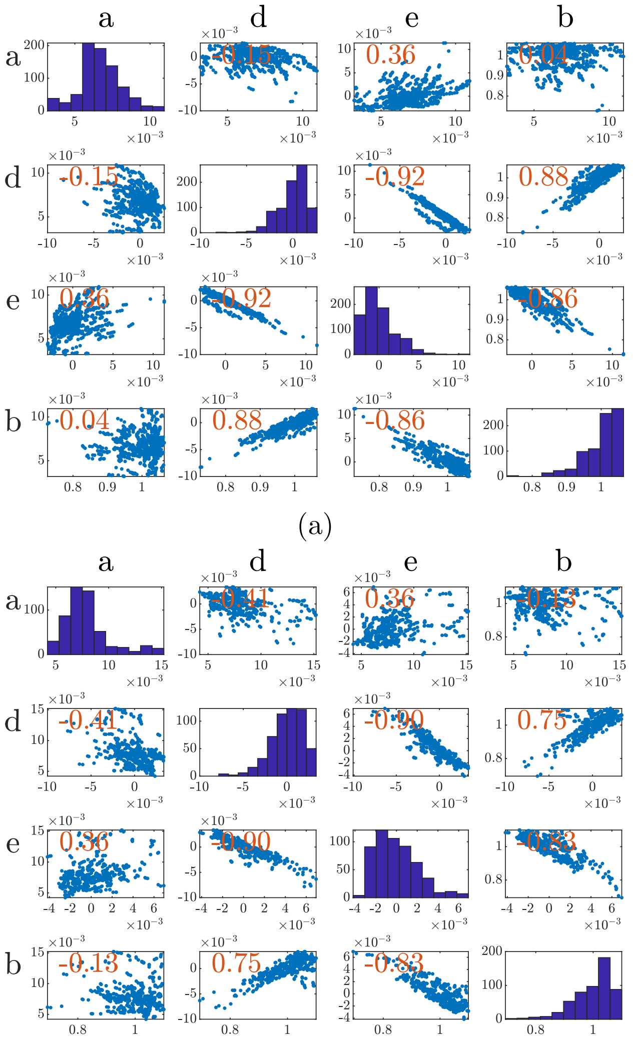
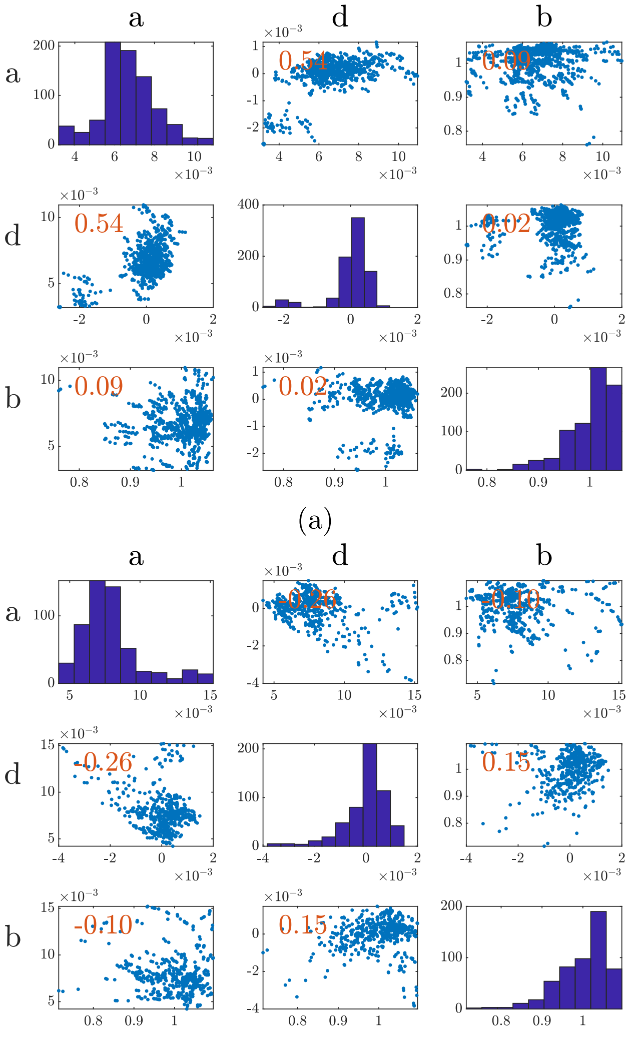


Figure 4.14: This figure includes correlation plots between estimated EMSC parameters for data sets Human12 (top) and Equine4 (bottom). Spectra with no cartilage signal are removed prior to correction.



(a)

(b)

Figure 4.15: This figure includes correlation plots between estimated MSC-L parameters for data sets Human12 (top) and Equine4 (bottom). Spectra with no cartilage signal are removed prior to correction.

At this point, some comments about the EMSC parameters as describors of physical phenomena in the sample should be made. In general, diseased cartilage has different morphology than healthy cartilage. For instance softening of the tissue is associated with diseased cartilage [45]. We can hypothesise that such difference in morphology leads also to optically different properties and physical effects in the spectra , e.g. due to variations in the penetration depth of the infrared radiation. The physical effects are expected to result in discriminative information in the spectra. For classification tasks, it may thus be desirable to exploit this. However, as concluded from the previous paragraph, we can not guarantee that correction of 7 wavenumber channel data to be as accurate with respect to retrieving pure chemical information as for broad-band spectra. Implicitly, the estimated physical baseline effect parameters from the EMSC correction of seven wavenumber channel data does most likely not exclusively describe physical phenomena in the sample, but may in stead express a trend in the relationship between absorbance levels for the 7 wavenumbers. Nevertheless, it may be of value to exploit such a trend in classification tasks of healthy and diseased cartilage.

4.3.4 Suggestions for preprocessing strategies for 7 selected wavenumber channels data

In the sections leading up to this point, EMSC correction for broad-band spectra and the corresponding corrections for seven wavenumber channels data was discussed. In this section we exploit these observations to motivate preprocessing strategies for spectral data with the seven wavenumber channels for which QCL lasers are being developed in the Miracle project.

We have seen that corrections are more comparable between the two types of data with simpler EMSC complexities. MSC is considered more correct than MSC-L, and correspondingly MSC-L is considered more correct than EMSC. In addition, it was argued that even though EMSC corrections of seven wavenumber channels data is not guaranteed to be as accurate as for broad-band spectra, it may identify important trends in the data. Naturally, different trends can be identified by using different EMSC models, which means that even though MSC is concluded to be more accurate, it is still of value to investigate all EMSC models. By including EMSC parameters as extra variables in classification, we can make sure the classifier will have the opportunity to use them separately from the chemical information. Further-

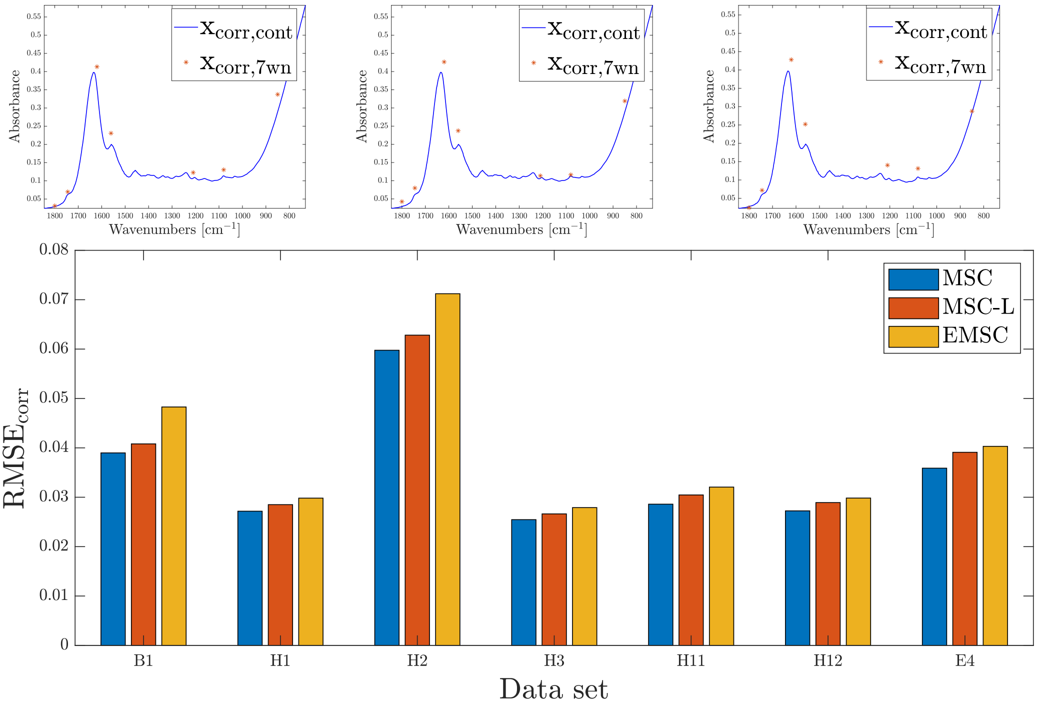


Figure 4.16: This figure summarises the comparison of EMSC-type correction models applied for broad-band spectra and 7 wavenumber channels data. The top row shows the corrections for i) an arbitrary broad-band spectrum, $x_{corr,cont}$ (blue line) and ii) the corresponding correction, $x_{corr,7wn}$, based on the 7 wavenumbers only (orange). We show, respectively from left to right, corrections by MSC, MSC-L and EMSC. The bottom row compares the mean RMSE difference between the corresponding corrections in all given data sets.

more, using raw data without any further consideration may be dangerous since physical effects may lead to uncontrollable interferences. When physical effects as estimated by EMSC model parameters and chemical effects are used for a classifier,differences in the actual values of the data may bias the model if the classifier is not scaling-invariant. In this case, it may be advantageous to standardise the variables (by subtracting mean and dividing by standard deviation) before classification. Based on these observations, the following suggestions for preprocessing and classification strategies for the seven wavenumber channel data were evaluated.

1. Establishment of a classifier with non-preprocessed data.
2. Establishment of a classifier with constant baseline corrected data, by using background measurement at 1800cm^{-1}.

3. Establishment of a classifier with MSC/MSC-L/EMSC corrected data.
4. Establishment of a classifier with non-preprocessed data and estimated MSC/MSC-L/EMSC parameters for weighting of trends
5. Establishment of a classifier with MSC-corrected spectra as a best possible correction, and estimated MSC/MSC-L/EMSC parameters for weighting of trends.

All suggested strategies were tested with and without standardizing variables. In order to validate the suggested preprocessing strategies we establish a simulated data set. In the next section we consider and discuss the establishment of this simulated data set, before the validation is run in section 4.5.

4.4 Simulation of spectra

In this section we discuss the simulation of cartilage spectra, and present the resulting data sets. The simulation of the data set is based on a PCA simulation approach which is described in more detail in the method section 3.4. In total 1000 spectra were simulated, of which 50% are spectra corresponding to healthy cartilage and 50% are cartilage spectra corresponding to diseased cartilage. The data sets were designed such that they contain the same variability as present in the healthy and diseased groups of the Human12 broad-band data set. We first simulate pure absorbance spectra by running PCA on an EMSC corrected broad-band spectra and recombine the principal components as described in the methods section. As described in section 4.3.2, broad band spectra are corrected by Multiplicative Signal Correction with the inclusion of a linear baseline in the model. This resulting corrected data set is considered a pure absorbance spectral data set, and is the basis for the simulation of the pure absorbance data set. Subsequently the simulated data set is perturbed by physical effects according to the variability of physical effects present in the healthy and diseased groups of the broad-band spectra. Thus the result is a data set with group specific chemical variations and group specific perturbation of physical effects, which is needed in order to validate the preprocessing-and-classification strategies suggested in section 4.3.4.

For perturbing the simulated data set with white noise, which represents random fluctuations in a spectrometer, random values is uniformly drawn from the interval 0 - 0.005. This level was chosen by inspecting the absorbance values in the absorption inactive region 2500-2550 cm^{-1}, in all the available data sets. The white noise levels range from 0.00078 for the Human12 data set to 0.016 for the Bovine data set, with Bovine having a particularly high noise level in comparison to the other data sets. Thus the level of 0.005 was chosen as it is within this range. Moreover, it was necessary to choose a cutoff value in the cartilage damage grade, represented by OARSI grading, to define which spectra are in the category healthy and which are in category diseased. The cutoff chosen was OARSI grade 1.5. This choice was motivated by the fact that the Miracle project aims at detecting early stage degradation of the articular cartilage tissue. As the OARSI grades range from 0 to 6, a cutoff at 1.5 means that we aim at identifying diseased samples from an early stage degradation on the strongly degrated cartilage. However it should be noted that for this cutoff, the healthy and diseased groups are not balanced with respect to each other. This is in general an unfavorable situation, but what is most important for simulation purposes is that the size of each group in the broad-band spectra is considerable enough to give a realistic group variation. If there are too few spectra in one group, the corresponding standard deviation of the spectral data has a high error. This was one of the reasons why the large data set size for the Human12 data was prioritized instead of choosing a data set with higher signal to noise ratio in section 4.1.1. There are 248 healthy and 534 diseased spectra with the chosen cutoff for the Human12 data set. Thus, there is still a reasonable amount of spectra in the healthy group, and it is assumed that this is a satisfactory amount to get a realistic group variation. In the following sub sections, some further simulation specifications are considered.

4.4.1 Window function for weighting out irrelevant interferents

For simulation, we do not desire to recreate the absorption peaks associated with CO2 or the water combination band in region 1780 - 2600 cm^{-1}. This absorption is a source of variation in the data set, but we do not expect it to carry any discriminative information for healthy and diseased cartilage. Moreover, because these absorbance peaks are located in an otherwise absorption free region, we have the possibility to apply a window function to filter it out, without disturbing any informative spectral variations. To achieve this, a function for smoothly filtering out the peaks in region 1780 - 2600 cm^{-1} was constructed based on the the Tukey window function [43]. The Tukey window function, also called the tapered cosine function, is for

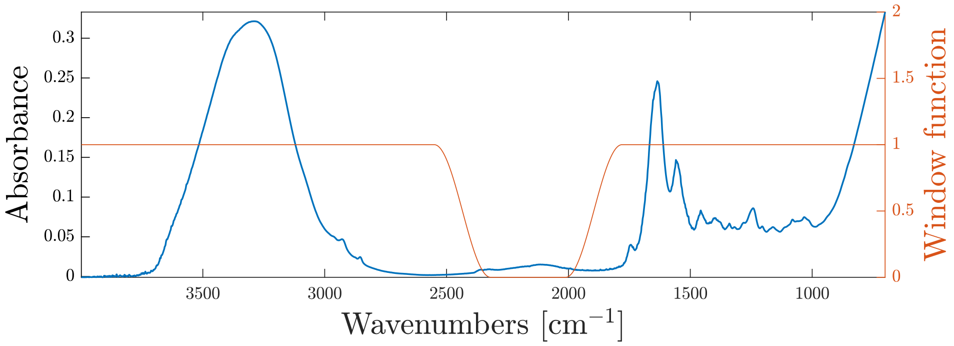


Figure 4.17: This figure shows the window function (orange) used to weight out water combination band and CO2 band in the simulation, together with the mean spectrum (blue) for the Human12 data set.

this purpose augmented to the resulting function shown in Fig. 4.17 (red). It should be noted that the function is applied after the MSC-L correction, to avoid disturbance of the parameter estimation, since realistic values of the parameters are important for the validation of the suggested preprocessing strategies for seven wavenumber channels data.

4.4.2 Selection of principal components

In the simulation approach used in this thesis, the principal components (loadings) for the PCA model are calculated and recombined into new spectra for the construction of a healthy cartilage tissue group and a diseased cartilage tissue group. The recombination of loadings was done separately for healthy and diseased groups. In this section, we evaluate how many principal components to include in the simulation model. To this end, PCA is run on the MSC-L corrected broad-band Human12 data set, and the calculated loadings are investigated. We use influence plots as an extra quality check for the data on which the simulation model is built, and motivate the further removal of some spectra. The goal is to include the components which contain information about the between-class spectral variation as well as variability that is common for the two classes, without introducing too many irrelevant artefacts.

In the PCA simulation approach, we run PCA on healthy and diseased groups separately, identifying the spectral variability in the data set within each of these groups. We thus obtain two separate sets of loadings, shown respectively in Fig. 4.18 and Fig. 4.19. Firstly, we note that the three first components appear smooth and free from physical and other interference effects for both healthy and diseased groups. By including the three first principal components, we explain 94,2% of the spectral variability for group healthy and 95,2 % for group diseased, which we in general consider an acceptable amount for the purpose of simulation. However, whether to include more than the first three components should be considered in more detail. As we see for both the healthy and diseased group loadings, the 4th and 5th components show signs of interference effects in the regions 3700-4000 cm^{-1} and 1700-2000 cm^{-1}, which are attributed to water vapor rotational transitions (see section 2.2, Fig. 2.4). Indeed component 5 accounts mainly for water vapor interferences. Therefore we must consider whether water vapor is something we want to include in the simulation. This is discussed in the following paragraph. Although the 6th principal component also contains some water vapor features, it is discarded because it explains only 0.8 % of the variance in both groups.

We evaluate now if water vapor should be included in the simulation or not. Firstly, the presence of water vapor in spectra is often an indicator that there has been water vapor in the air inside the instrumentation during measurements. To investigate if water vapor is associated with a limited number of spectra, or if it is a common occurrence in most spectra, the 5th principal component from PCA is used, since this component contains almost only signals from water vapor. In order to remove other possible contributions, we set the regions not associated with water vapor to zero. We recombine the the scores of the 5th components, t_5, with the augmented water vapor component $\tilde{p_5}$ by,

$$X_{wv,centered} = t_5 \tilde{p_5}' \tag{4.1}$$

, where $X_{wv,centered}$ is water vapor contributions (with respect to the mean spectrum) for each spectrum. $X_{wv,centered}$ is shown in Fig. 4.20, where we have zoomed in at the water vapor absorbance peaks in region 1300 - 1900 cm^{-1}. It can clearly be seen that many spectra contain more water vapor than the mean spectrum. This motivates us to include the principal component identifying water vapor variations in the simulation, since it is clearly an interference which is always present, at least for the instrumentation used for data set Human12. This may also be the case for the final Miracle probe instrumentation, unless instrumental precautions are made. For classification tasks such as in the following validation section, it should thus be desirable that classifiers are able to handle these variations. We concluded that water vapor components 4 and 5 should be included in the PCA simulation for

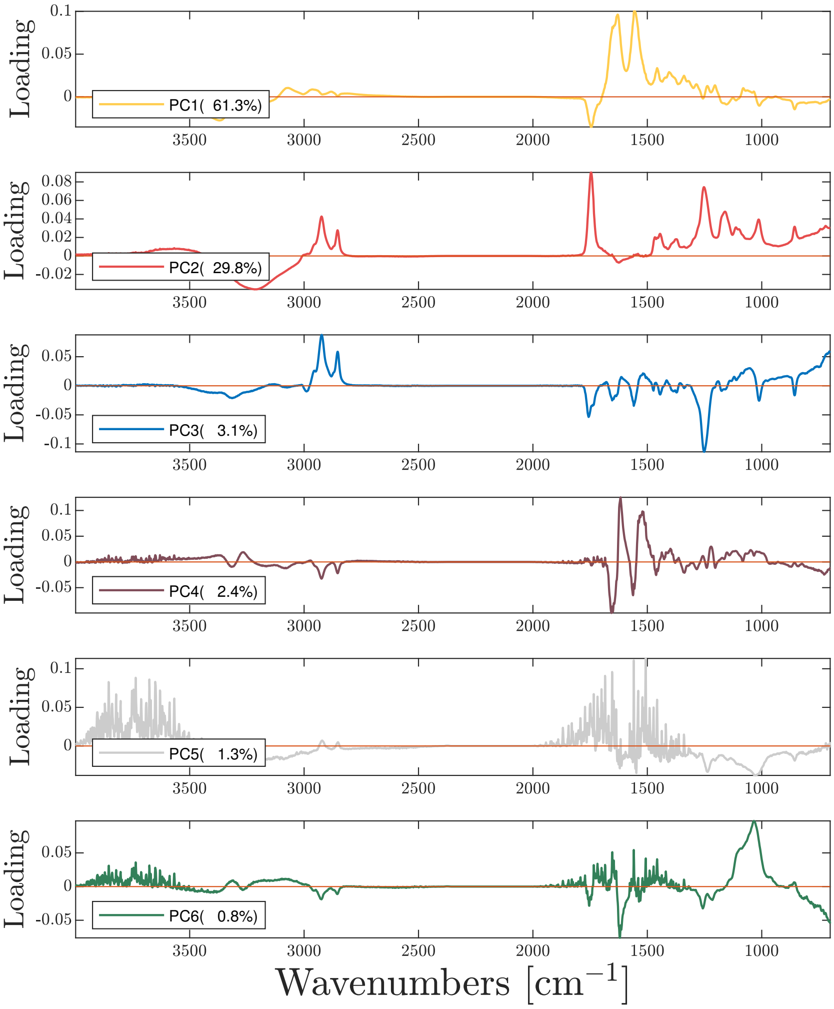


Figure 4.18: The figure includes the six first PCA loadings (PCs) for the healthy group in the broad-band data set Human12. The explained variance by each component is marked in the legend.

healthy and diseased groups even though they represent interference effects. Further, in section 4.6, we will see how classification results are impacted by including versus not including water vapor in the simulation, for further discussion on this topic.

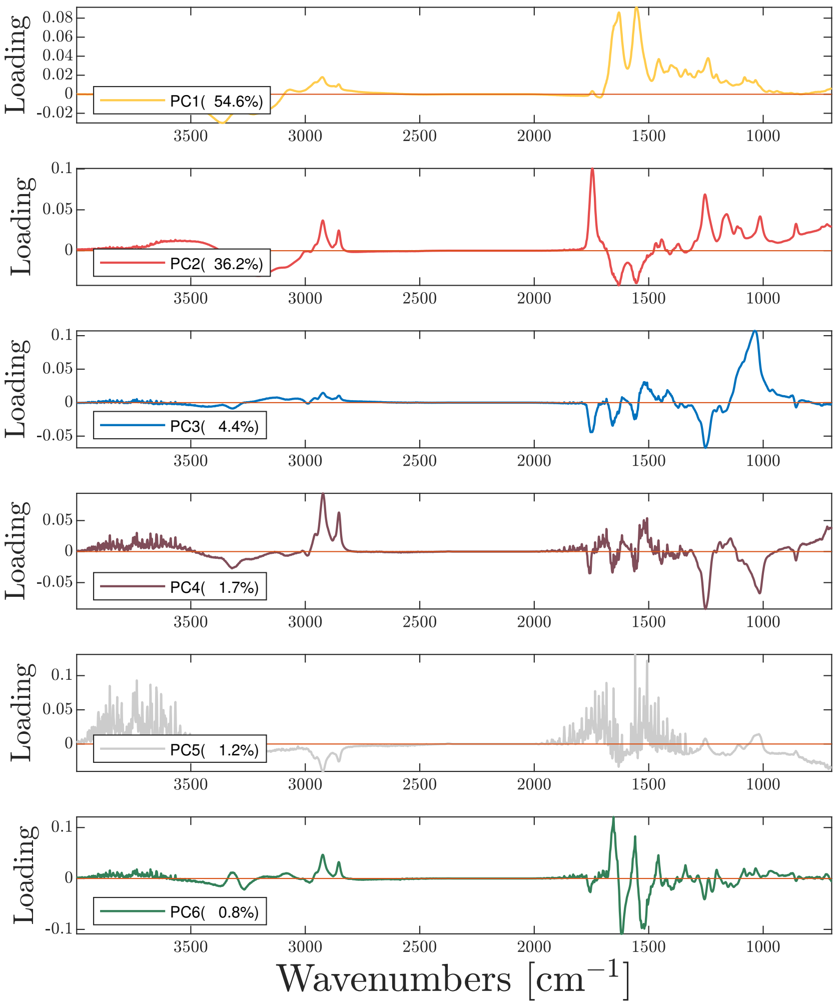


Figure 4.19: The figure includes the six first PCA loadings (PCs) for the diseased group in the broad-band data set Human12.The explained variance by each component is marked in the legend.

4.4.3 Thorough quality check for the data used for the simulation

In this section we use the residuals and leverage plots for the final model established for PCA and an additional PLS-DA model as an extra quality check of the data on which the simulation will be based. We identify additional spectra with unsatisfactory cartilage signal. In Fig. 4.21 we present influence plots for the final PCA simulation models for healthy (top) and diseased (middle) cartilage spectra groups, as well as for the additional PLS-DA model (bottom). For the PLS-DA model we use the three first PLS

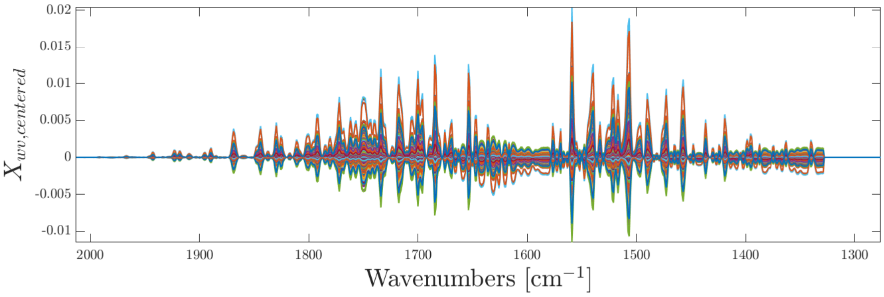


Figure 4.20: This figure shows the water vapor contribution calculated by equation 4.1 for all spectra in Human12 experimental data set.

loadings. We consider first the influence plot for the PCA simulation model for the healthy group. The three replicates with number 162 - 164 are characterized by having both high leverage and high residuals. This means that they considerably affect the model, but in addition are not well explained by the model. This motivates us to investigate these measurement further. The three high leverage and high residual spectra are shown in Fig. 4.22 (top row, left column). The ID tags for the sample are also included for later comparison across PCA and PLS-DA influence plots, since the numbers in the influence plot refer to row number within each group data block (i.e block healthy, block diseased or full data block). As we see, this sample shows considerable different features than what we expect from cartilage tissue, as we see by qualitatively comparing the spectra with Fig. 2.2. The absence of an Amide II peak in the region 1500 - 1600 cm^{-1} is particularly notable. Consequently, it is concluded to discard this sample from the data set before simulation. Furthermore, we can inspect the influence plot for the PCA simulation model of the diseased group in Fig. 4.21 (middle). The three replicates with number 1 - 3 clearly stands out from the rest of the spectra, having both high leverage and high residuals. In addition, replicate sets 55 - 57, 204 - 206 and 124 - 126 have higher residuals than the majority of the spectra. All of these replicate sets are plotted inf Fig. 4.22 (row 1, columns 2-3 and row 2,columns 1-2). It can be seen that all these replicate sets, except 55 - 57, show high deviance from what we expect from cartilage signal and are thus excluded from the data set. Replicates 55-57, however, is kept.

Lastly, we consider the the influence plot for the PLS-DA model in Fig. 4.21 (bottom). It is observed that several replicate sets have both high residual and high leverage, including 4 - 6, 354 - 356 and 533 - 535. Checking replicates 4 - 6 in Fig. 4.22 (row 2, column 3), we can readily see from the ID tags that these replicates are the same as the replicates 1 - 3 identified from the PCA simulation model (row 1, column 2). They were therefore already decided to be excluded. This is also the case for replicates 533 - 535 (row 3,column 3) as can be seen by comparing ID tags with replicates 162 - 164 from PCA simulation model (row 1, column 1). The replicates 354 - 356 (row 3, column 2) show the same types of features as the ones we already removed. Therefore, we remove them as well. In addition, the two spectra 375 and 376 have high leverage, while their residual is at the same order as it is for the majority of the spectra. This means that the PLS-DA model highly weight the spectral features seen in these spectra and successfully account for these variations. We inspect the full replicate set 374 - 376 to check if the spectral features seen for these spectra is something we want to account for. As seen in Fig. 4.22, the sample is characterised by having high peaks in the region 1000 - 1100 cm^{-1} in comparison to the healthy cartilage spectrum (Fig. 2.2). This is clearly not a healthy cartilage spectrum. However, it may represent a diseased cartilage spectrum and it is thus important to keep it. For comparison, the main absorbance band for bone in this region is associated with a phosphate peak at 1010 cm^{-1} [21]. Since our diseased group consists of cartilage samples with both middle and high degeneration grades (OARSI 1.5 - 6), the highest peaks in our simulated spectra for this region may originate from bone. However, upon closer inspection of the peaks in the Human12 data set, the most dominant peak in the carbohydrate region is not located at 1010 cm^{-1}, but closer to 1032 cm^{-1}, which is an expected peak for cartilage, as seen from table 2.1. By inspecting the raw Human11 and Human12 data sets shown in Fig. 4.1 in section 4.1, it is apparent that there are many spectra which show a strong band at 1032 cm^{-1}, which may indicate that it is indeed an important characteristic to include in the simulation. We kept therefore the replicates 374 - 376 in the simulation.

4.4.4 Simulation results

In this section, we present the simulation results. We discuss the quality of the simulated data set, and link the apparent impairment to some simulation drawbacks. Results from the simulation are included in Fig. 4.23. The most apparent difference between healthy and diseased groups are the absorbance values in the region 950 - 1125 cm^{-1}. Consulting table 2.1 in section 2.1.4, we see that absorbance bands in this region are mainly associated with collagen and proteoglycans for cartilage tissue, suggesting that healthy and diseased cartilage can be discriminated from each other mainly based on absorbance in this area. However, by inspecting the mean difference between

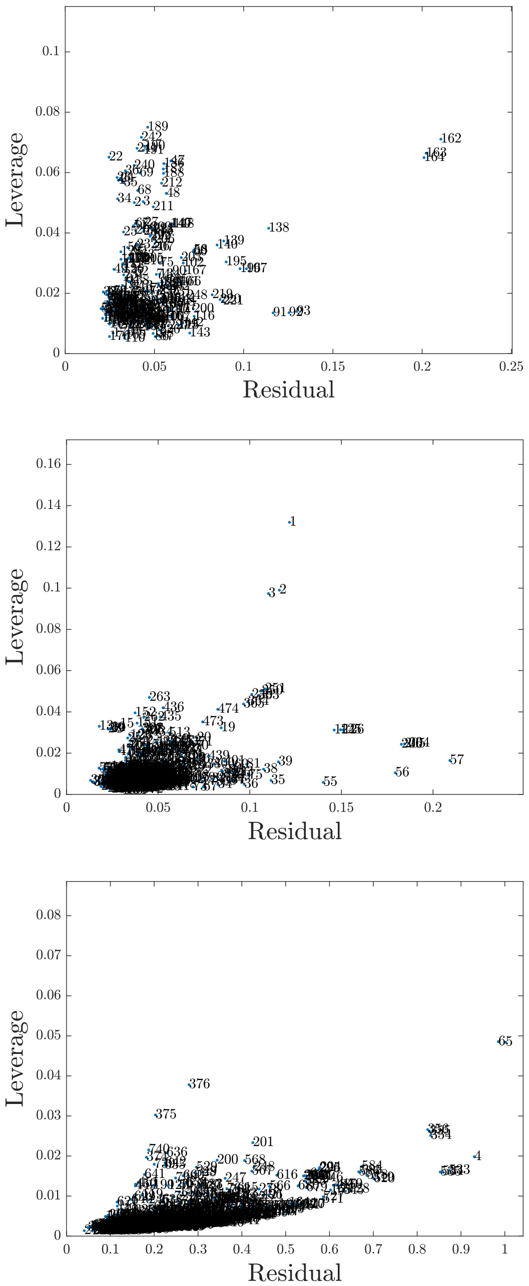
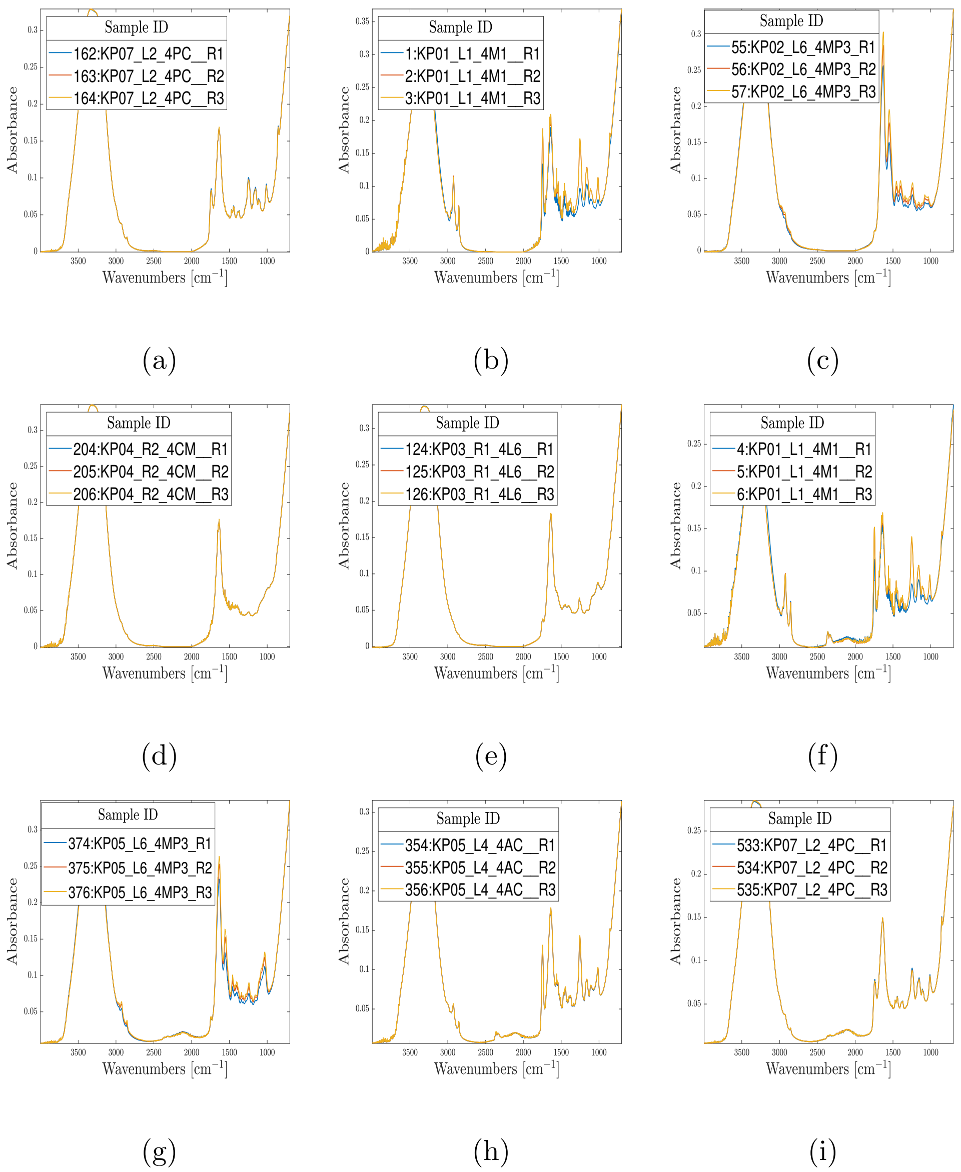


Figure 4.21: This figure shows influence plots for the final PCA simulation model of healthy group (top) and diseased group (middle) and the additional PLS-DA model (bottom). Some spectra with high residuals and high leverage can be seen.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.22: In this figure, we show the replicate sets of spectra that were identified in Fig. 4.21 as having particularly high residual or high leverage for the PCA based simulation model for respectively healthy and diseased cartilage spectra as well as for the additional PLS-DA model. Replicates in (a)-(b) were found by PCA in healthy group. Replicates in (d)-(f) were found by PCA in diseased group, and replicates in (g)-(i) were found by PLS-DA in the full data set including both healthy and diseased samples.

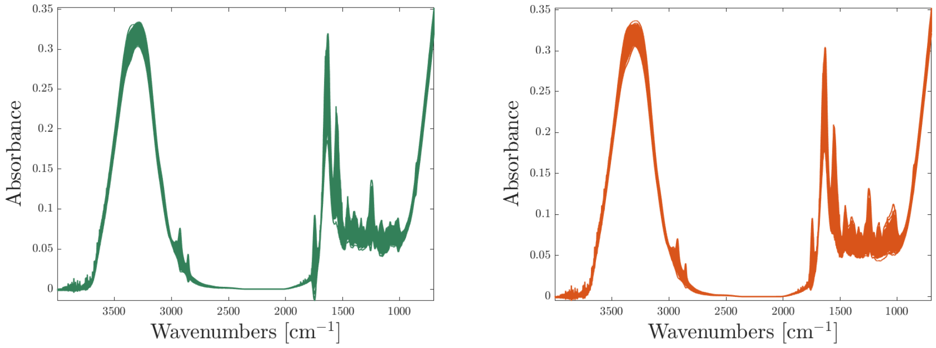


Figure 4.23: These plots show the simulation results for the unperturbed healthy (left) and diseased (right) groups.

the healthy and diseased groups from Fig. 4.24, we see that the difference in average spectra of the healthy (green) and diseased (red) groups in region 950 - 1125 cm^{-1} is not as big as some of the simulated spectra suggests. This may indicate that there is a large variety of signal strength in this region for the diseased group, and furthermore that spectra with very high absorbance levels in the collagen and proteoglycan region may belong to particular high degradation cartilage as also discussed in section 4.4.3.

An artefact in our simulated spectra is the occurrence of below-baseline features in the region 1720 - 1780 cm^{-1}, seen particularly for the simulation of healthy cartilage group seen in Fig. 4.23 (left) . This artefact is also apparent in the diseased groups to some extent. This may be linked with the simulation method itself. The simulation is based on drawing scores and perturbation parameters from a normal distribution with mean and standard deviation calculated from the experimental data set. Drawing scores for each principal component independently in this manner will most likely create some unrealistic combinations of components. In fact, the assumption that scores from PCA and PLS-DA are exactly normally distributed for healthy and diseased cartilage spectra may be erroneous. For a quick check of this, we present the distribution of the first three principal components from PCA on Human12 data set in Fig. 4.25 plotted together with the fitted normal distribution. As exemplified by the 2nd principal component (PC2) for a PCA run on the healthy spectrum group (middle row, right column), some components exhibit distributions which can not simply be explained by a normal distribution defined from the mean and standard deviation parameters. The histogram of the 2nd principal component for healthy group shows

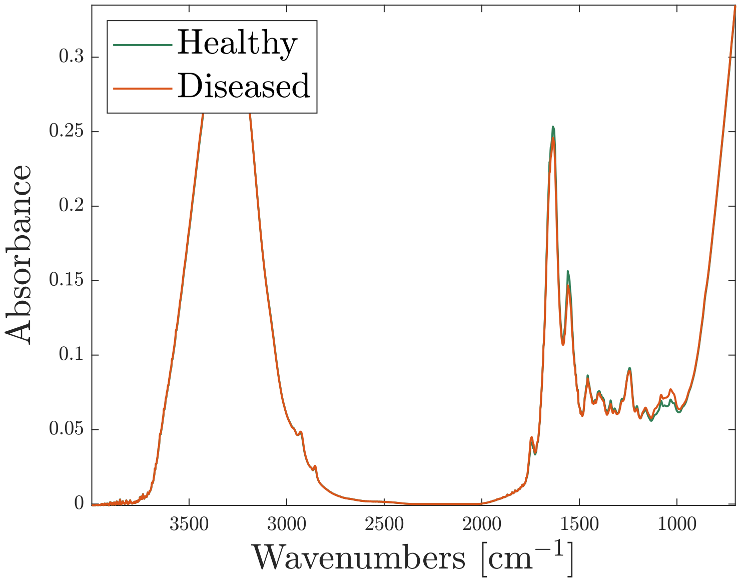
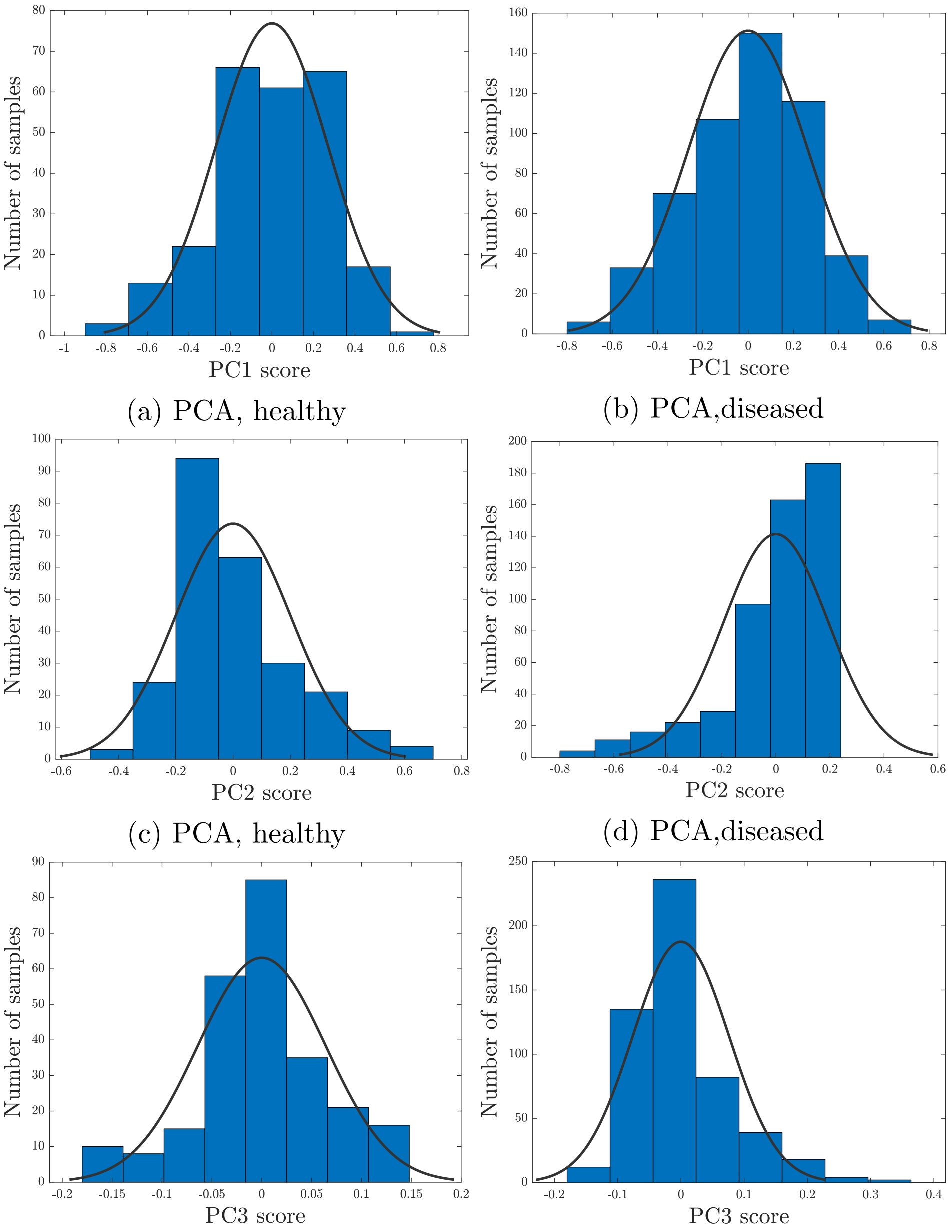


Figure 4.24: This plot shows the mean spectra for the unperturbed healthy (green) and diseased (orange) groups from simulation.

a non-symmetrical shape, which means that there are no samples contained in the right tail of the fitted normal distribution. Thus, the real distribution may be skewed. When this is not accounted, it may ultimately lead to the below-baseline artefacts in the region 1720 - 1780 cm^{-1} which are seen in the simulated data set. For future work, this is thus an issue which should be considered further. For the upcoming validation section, the spectra with such artefacts were removed by a simple absorbance level criteria at 1750 cm^{-1} (Abs < 0.011), for which 20 spectra are discarded from the data set.



(a) PCA, healthy

(b) PCA,diseased

(c) PCA, healthy

(d) PCA,diseased

(e) PCA, healthy

(f) PCA,diseased

Figure 4.25: Distributions of the first three components of PCA scores in the experimental data set Human12 are shown for healthy (left column) and diseased (right column) groups. The fitted normal distributions for the scores are shown by the black line.

4.5 Validation of preprocessing strategies

In this section, we validate the preprocessing and classification strategies for seven wavenumber channels data which were suggested in section 4.3.4. This is achieved by applying a spectrum of classifiers to the simulated data, namely Random Forest (RF), Partial Least Squares Discriminant Analysis (PLS-DA), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). We tested 11 different variants of preprocessing strategies in combination with these classifiers. The different preprocessing strategies are

1. Non-preprocessed data
2. Constant baseline corrected data by subtraction of absorbance at 1800 cm^{-1}
3. MSC corrected data
4. MSC-L corrected data
5. EMSC corrected data
6. Raw data and MSC parameters added
7. Raw data and MSC-L parameters added
8. Raw data and EMSC parameters added
9. MSC correction and MSC parameters added
10. MSC correction and MSC-L parameters added
11. MSC correction and EMSC parameters added

The 11 preprocessing strategies were in addition combined with standardisation of all classification variable inputs, yielding in total 22 different preprocessing strategies. We present the classification accuracy results from the exhaustive search among all suggested preprocessing-and-classification strategies, using the simulated data set, in table 4.2. When we applied standardisation of all classification variable inputs, we denoted results in table 4.2 by (*). We see that results vary across classifiers and preprocessing strategies. The Support Vector Machine (SVM) classifier is not scale-invariant, and thus it does not perform well on data for which the variables are not standardized. This is readily observed in our table by comparing SVM accuracy for the standardized strategies (1* - 11*) with the non-standardised approaches (1 - 11). For the results using non-standardised variable approaches, we thus ignore the SVM. By inspecting the results for the non-standardised strategies 1-11, the best preprocessing approach across all classifiers is apparently a simple MSC with the estimated MSC parameters added as additional variables for the classifiers (green row). However, a simple MSC correction without the inclusion of estimated parameters as extra variables (blue row) leads to a comparable accuracy. The inclusion of extra MSC parameters had most effect on the ANN classifier with 3,2 % increase in accuracy, and the difference is marginal for RF (+0.6 %.) and PLS-DA (+0.1 %). Comparing these observations with the corresponding standardised strategies, the same pattern is seen. Random Forest gave the highest accuracy of all tested classifiers, and in comparison to no preprocessing (1), marked in grey, we achieved a classification accuracy increase of 5,6 % for a simple MSC and 6,2 % for an MSC correction with MSC parameters included as additional variables (i.e weighting of trends in data). As Random Forest is a scaling invariant method, it performs equally on standardised and non-standardised data.

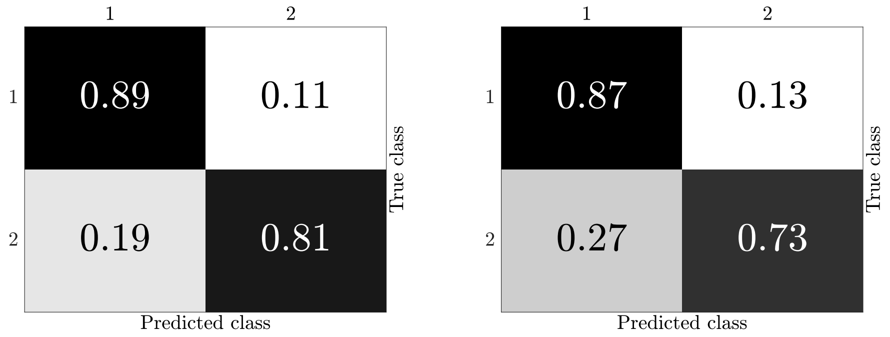
We consider the impact of standardization of the input variables for the classifiers further, by comparing non-preprocessed data (1) and standardized data (1*). Across the four different classifiers, we do not see a general improvement in classification accuracy due to standardisation. However, the effect varies. While Random Forests yields insignificant difference in accuracy, PLS-DA shows a marginal increase in accuracy of 1,8 % and for ANN we see a marginal decrease of 1,0 %. The impact is, as mentioned earlier, significant for SVM due to the classifier's sensitivity to scales, and the accuracy increases by 21 %. By correspondingly comparing the non-standardised version of highest accuracy preprocessing approach (9) with the standardised version (9*) (both marked in green), we observe that the marginal differences that was seen by comparing (1) and (1*), becomes even smaller. For ANN and RF, there is no difference in accuracy and for PLS-DA there was an accuracy increase of 0,5 %. Thus, standardisation had less impact for the MSC preprocessed data than for the raw data. In the appendix we include, correspondingly to table 4.2 for RF as a representative of the highest accuracy classifier in our case, other conventional classification metrics, to double check that all metrics show the same pattern, and for instance that specificity or sensitivity are not critically affected for any preprocessing strategies. From this table, we can see that none of the metrics (accuracy, true negative rate, precision, recall or F1-score) are critically different than the others for the respective strategies. This is in accordance with what we would expect, since we created a nearly perfectly balanced simulated data set. We can summarise from this section that preprocessing by conventional MSC was the most valuable preprocessing technique in combination with the Random Forests classifier. The inclusion of the estimated MSC parameters as extra input variables to the classifier led to further increase in accuracy, though marginal of 0,6 %. Standardisation is not necessary in the case of Random Forests.

Table 4.2: Table showing different classifiers' accuracy using the simulated data for the main variants of preprocessing strategies (1-11) and the corresponding versions with standardisation of variables (*).

Strategy	**RF**	**PLS-DA**	**ANN**	**SVM**
1	0.749480	0.735967	0.764033	0.573805
2	0.749480	0.748441	0.758836	0.567568
3	0.805613	0.747401	0.775468	0.515593
4	0.778586	0.745322	0.785863	0.515593
5	0.696466	0.699584	0.735967	0.515593
6	0.744283	0.734927	0.761954	0.608108
7	0.743243	0.738046	0.740125	0.613306
8	0.738046	0.732848	0.759875	0.618503
9	0.811850	0.746362	0.807692	0.604990
10	0.807692	0.750520	0.781705	0.607069
11	0.804574	0.741164	0.803534	0.610187
1*	0.747401	0.752599	0.753638	0.786902
2*	0.749480	0.749480	0.758836	0.801455
3*	0.805613	0.751559	0.775468	0.792100
4*	0.778586	0.737006	0.785863	0.796258
5*	0.696466	0.676715	0.735967	0.739085
6*	0.744283	0.737006	0.761954	0.786902
7*	0.743243	0.755717	0.740125	0.791060
8*	0.738046	0.745322	0.759875	0.770270
9*	0.811850	0.751559	0.807692	0.800416
10*	0.807692	0.743243	0.781705	0.800416
11*	0.804574	0.745322	0.803534	0.786902

4.6 Impact of water vapor interference on the classification results

One interferent which is clearly present in cartilage spectra is water vapor, which is associated with air inside the instrumentation. In this section, we exploit the clear separability of water vapor, which was seen for the PCA loadings in section 4.4.2, to investigate the impact of water vapor on classification. To achieve this, we simulated two data sets. One of the data sets was constructed only from loading 1-4, which mainly contain non-interferent features, and the other data set included in addition the 5th loading which contains mainly water vapor. For this study, noise was not added to the



(a) No water vapor: ACC=85% (b) With water vapor: ACC=80%

Figure 4.26: This figure shows the Random forests classification results for the simulated dataset, where the 5th loading, which contains almost only water vapor information, is (a) included and (b) not included in the simulation. The Accuracy (ACC) when not including the water vapor component is 85 % and the accuracy when including the water vapor component is 80 %.)

perturbed spectra. Subsequently, Random Forest classification was run on these data sets for performance comparison of a data set nearly free of water vapor and a data set including water vapor. In figure 4.26, confusion matrices for the two classifications are shown. It is seen that water vapor has an impact on classification results, and there is a 5% decrease in classification accuracy for the data set including water vapor. Thus, water vapor has a significant effect on classification of healthy and diseased cartilage for the seven wavenumber channels data. It is recommended that instrumental precautions are made to try and minimize this classification impairment. For instance, a purging mechanism can be implemented.

Chapter 5

Conclusion

The aim of this thesis was to (i) explore interferent and measurement variability in broad-band spectra, (ii) establish routines for detection of low quality broad-band spectra, (iii) use only selected wavelengths from the broad-band spectra (the wavelengths that were selected for the QCL lasers) and investigate preprocessing strategies based on only few wavelengths, (iv) to suggest preprocessing strategies for data with few wavelength channels, and finally (v) to simulate a data set based on the knowledge about interference effects from broad-band spectra and use the simulated data set for validation of the suggested preprocessing strategies.

In broad-band spectra of cartilage, several interference and measurement variations were identified from the raw data, including variations in water vapor, carbon dioxide, noise and cartilage signal strength. Spectra that did not show cartilage signals at all could also be identified. We suggested that this was due to the high degradation of cartilage in these samples. However, it was shown that it is difficult to classify these samples based on the spectral fingerprint and therefore it was concluded that such spectra will not give any meaningful value to further classification tasks, and should be removed. In terms of the Miracle probe system, this is an important observation because it shows that development of an automatic detection algorithm for no-cartilage-signal measurements will be vital for robust implementation. Three approaches for detection of spectra without cartilage signal was tested for the broad-band spectra. The most robust approach for broad-band spectra was calculating the difference between maximum and minimum of the absorbance derivative in the fingerprint region. However, such an approach will not be applicable to the Miracle probe data consisting only of seven wavenumber channels. Of the tested approaches, an approach based on calculating the residuals from an EMSC model with mean reference in comparison to an

EMSC model with water spectrum reference, respectively is the most applicable one for seven wavenumber channels data. This approach was shown to successfully identify all low absorbance spectra in broad-band spectra, however not as precise as the two other approaches.

We suggested that spectral features that are due to physical effects can carry discriminative information about healthy and diseased cartilage for broad-band spectra. EMSC-type correction methods could successfully separate the physical features and the absorption features. However, for the seven wavenumber channel data, corresponding EMSC-type correction methods were not as accurate as for broad-band spectra in separating physical and chemical information. Due to the low number of variables, they could not be separated completely since absorption features were modelled by the EMSC model functions for physical effects. This problem increased with the complexity of the EMSC model. Therefore, it is concluded that the estimated physical effects described by the EMSC model in seven wavenumber channels data most likely do not correctly describe physical phenomena in the sample. Based on this, 11 EMSC type preprocessing strategies for seven wavenumber channels data were suggested to test. To validate the suggested preprocessing strategies, a simulated data set of healthy and diseased cartilage spectra was established by exploiting broad-band spectra variability and using Principal Component Analysis. After an exhaustive search among the suggested preprocessing strategies, the best performance across all tested classifiers was obtained by using conventional MSC. The inclusion of the estimated MSC parameters as extra input variables to the classifier led to further increase in accuracy, although the improvement was marginal. In combination with the Random Forests classifier, the maximum accuracy of 81,2 % was achieved, which represented an increase of 6,2 % with respect to classification based on raw data. In conclusion, the preliminary study based on simulated data done in this thesis, suggests that application of MSC for preprocessing is the most promising approach for the seven wavenumber channels data which will be acquired by the Miracle probe.

As an additional test, the simulation approach was used to investigate how water vapor impact classification accuracy. By adding water vapor signals to the simulated data set for the seven wavenumber channels data in a level which was adopted from the broad-band spectra, a decrease of 5 % in classification accuracy was observed. Based on this, it is recommended that instrumental precautions are made to try and minimize this classification impairment. For instance, the possibility of implementing a purging mechanism may be investigated.

New interesting questions arose during the thesis which were outside the scope of the thesis. Firstly, this thesis tested only detection algorithms for low cartilage signal data for broad-band spectra. It is suggested that detection methods for seven wavenumber channels data are considered in future research. Secondly, it is suggested that the established simulation framework, which provides a controlled environment for testing algorithms, is used further to investigate how the suggested preprocessing and classification approaches in general, react to noise and interferents.

This thesis has shown how viable the use of EMSC type correction methods are for preprocessing of IR data with few wavenumber channels, such as the data which will be acquired by the Miracle probe. The Miracle system aims for an in-situ application, where the goal is real-time evaluation of cartilage, and therefore all data processing must be automatic. In this situation it will be particularly important that implemented preprocessing approaches are reliable and promote high classification performance to make in-surgery decisions safer.

Chapter 6

Appendix - Additional figures and tables

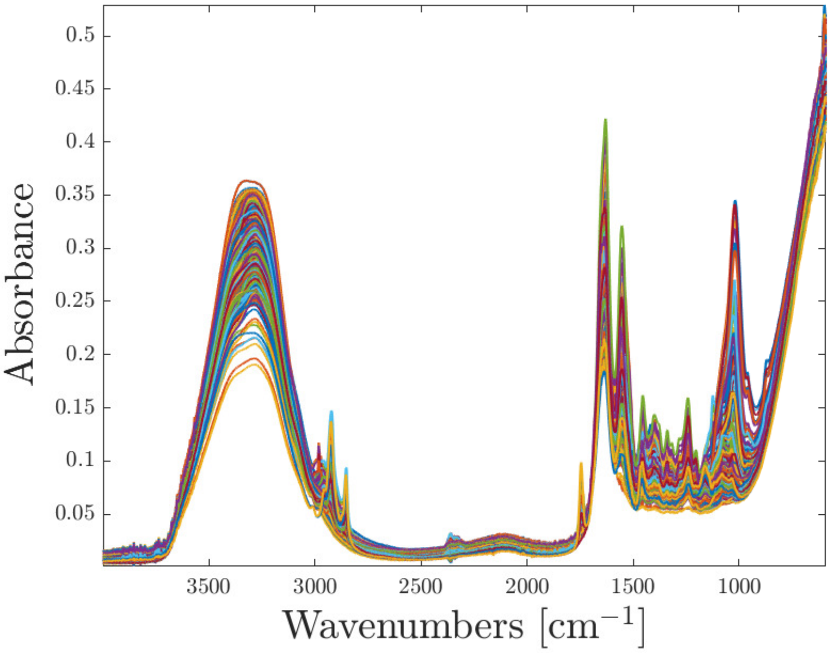


Figure 1: This figure shows the raw spectra of data set Equine4.

Table 1: This table shows classification metrics from Random Forest validation of the suggested preprocessing strategies 1 - 11, using the simulated data. The metrics included are accuracy (ACC), true negative rate (TNR), precision (PRE), recall (REC) and the F1-score.

Strategy	**ACC**	**TNR**	**PRE**	**REC**	**F1**
1	0.749480	0.751086	0.752769	0.751086	0.749283
2	0.749480	0.751021	0.752523	0.751021	0.749311
3	0.805613	0.806365	0.806327	0.806365	0.805613
4	0.778586	0.779766	0.780444	0.779766	0.778546
5	0.696466	0.697987	0.699242	0.697987	0.696244
6	0.744283	0.745591	0.746513	0.745591	0.744193
7	0.743243	0.744583	0.745572	0.744583	0.743143
8	0.738046	0.738829	0.738895	0.738829	0.738045
9	0.811850	0.812543	0.812434	0.812543	0.811848
10	0.807692	0.808251	0.808037	0.808251	0.807682
11	0.804574	0.805097	0.804865	0.805097	0.804560
1*	0.747401	0.748745	0.749751	0.748745	0.747303
2*	0.749480	0.751021	0.752523	0.751021	0.749311
3*	0.805613	0.806365	0.806327	0.806365	0.805613
4*	0.778586	0.779766	0.780444	0.779766	0.778546
5*	0.696466	0.697987	0.699242	0.697987	0.696244
6*	0.744283	0.745591	0.746513	0.745591	0.744193
7*	0.743243	0.744583	0.745572	0.744583	0.743143
8*	0.738046	0.738829	0.738895	0.738829	0.738045
9*	0.811850	0.812543	0.812434	0.812543	0.811848
10*	0.807692	0.808251	0.808037	0.808251	0.807682
11*	0.804574	0.805097	0.804865	0.805097	0.804560

Bibliography

[1] Lyn March, Emma U.R. Smith, Damian G Hoy, Marita J Cross, Lidia Sanchez-Riera, Fiona Blyth, Rachelle Buchbinder, Theo Vos, and Anthony D Woolf. Burden of disability due to musculoskeletal (MSK) disorders. *Best Practice & Research Clinical Rheumatology*, 28(3):353–366, jun 2014.

[2] Tatiana Konevskikh, Rozalia Lukacs, and Achim Kohler. An improved algorithm for fast resonant Mie scatter correction of infrared spectra of cells and tissues. *Journal of Biophotonics*, 2018.

[3] A. Köhler, J. Sulé-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. Van Pittius, G. Parkes, and H. Martens. Estimating and correcting Mie scattering in synchrotron-based microscopic fourier transform infrared spectra by extended multiplicative signal correction. *Applied Spectroscopy*, 2008.

[4] A. Kohler, C. Kirschner, A. Oust, and H. Martens. 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. *Applied Spectroscopy*, 2005.

[5] Harald Martens and Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 1991.

[6] J. L. Ilari, H. Martens, and T. Isaksson. Determination of particle size in power by scatter correction in diffuse near-infrared reflectance. *Applied Spectroscopy*, 1988.

[7] Peter R. Griffiths. Fourier transform infrared spectrometry, 1983.

[8] Reeta Davis and Lisa J. Mauer. Fourier transform infrared (FT-IR) spectroscopy: a rapid tool for detection and analysis of foodborne pathogenic

bacteria. *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology. A. Méndez-Vilas (Ed.)*, 2010.

[9] Anna Tinti, Vitaliano Tugnoli, Sergio Bonora, and Ornella Francioso. Recent applications of vibrational mid-infrared (IR) spectroscopy for studying soil components: A review. *Journal of Central European Agriculture*, 2015.

[10] Zanyar Movasaghi, Shazza Rehman, and Ihtesham U. Rehman. Raman spectroscopy of biological tissues, 2007.

[11] David W. Ball. Michelson Interferometer. In *Field Guide to Spectroscopy*. 2009.

[12] J Fahrenfort. Attenuated total reflection. *Spectrochimica Acta*, 17(7):698–709, jan 1961.

[13] Roland W. Frei and Harry Zeitlin. Diffuse Reflectance Spectroscopy. *C R C Critical Reviews in Analytical Chemistry*, 1971.

[14] Michael P. Fuller and Peter R. Griffiths. Diffuse Reflectance Measurements by Infrared Fourier Transform Spectrometry. *Analytical Chemistry*, 1978.

[15] Bruce Hapke. Specular reflection. In *Theory of Reflectance and Emittance Spectroscopy*. 2012.

[16] Melanie M. Beasley, Eric J. Bartelink, Lacy Taylor, and Randy M. Miller. Comparison of transmission FTIR, ATR, and DRIFT spectra: Implications for assessment of bone bioapatite diagenesis. *Journal of Archaeological Science*, 2014.

[17] Barbara Louise Mojet, Sune Dalgaard Ebbesen, and Leon Lefferts. ChemInform Abstract: Light at the Interface: The Potential of Attenuated Total Reflection Infrared Spectroscopy for Understanding Heterogeneous Catalysis in Water. *ChemInform*, 2011.

[18] Jože Grdadolnik. ATR-FTIR spectroscopy: Its advantages and limitations. *Acta Chimica Slovenica*, 2002.

[19] Alice J. Sophia Fox, Asheesh Bedi, and Scott A. Rodeo. The basic science of articular cartilage: Structure, composition, and function. *Sports Health*, 2009.

[20] Adele Boskey and Nancy Pleshko Camacho. FT-IR imaging of native and tissue-engineered bone and cartilage, 2007.

[21] Ioannis Kontopoulos, Samantha Presslee, Kirsty Penkman, and Matthew J. Collins. Preparation of bone powder for FTIR-ATR analysis: The particle size effect. *Vibrational Spectroscopy*, 2018.

[22] Cyril Petibois and Gérard Déléris. Chemical mapping of tumor progression by FT-IR imaging: towards molecular histopathology, 2006.

[23] Cyril Petibois, Gilles Gouspillou, Katia Wehbe, Jean Paul Delage, and Gérard Déléris. Analysis of type i and IV collagens by FT-IR spectroscopy and imaging for a molecular investigation of skeletal muscle connective tissue. *Analytical and Bioanalytical Chemistry*, 2006.

[24] Erik Goormaghtigh, Jean Marie Ruysschaert, and Vincent Raussens. Evaluation of the information content in infrared spectra for protein secondary structure determination. *Biophysical Journal*, 2006.

[25] Heinz Fabian and Dieter Naumann. Methods to study protein folding by stopped-flow FT-IR. *Methods*, 2004.

[26] Nancy P. Camacho, Paul West, Peter A. Torzilli, and Richard Mendelsohn. FTIR microscopic imaging of collagen and proteoglycan in bovine cartilage. *Biopolymers - Biospectroscopy Section*, 2001.

[27] Michael Jackson, Lin P.ing Choo, Peter H. Watson, William C. Halliday, and Henry H. Mantsch. Beware of connective tissue proteins: Assignment and implications of collagen absorptions in infrared spectra of human tissues. *BBA - Molecular Basis of Disease*, 1995.

[28] A. Kohler, D. Bertrand, H. Martens, K. Hannesson, C. Kirschner, and R. Ofstad. Multivariate image analysis of a set of FTIR microspectroscopy images of aged bovine muscle tissue combining image and design information. *Analytical and Bioanalytical Chemistry*, 2007.

[29] Michael Jackson, Michael G. Sowa, and Henry H. Mantsch. Infrared spectroscopy: A new frontier in medicine. In *Biophysical Chemistry*, 1997.

[30] R. Servaty, J. Schiller, H. Binder, and K. Arnold. Hydration of polymeric components of cartilage - An infrared spectroscopic study on hyaluronic acid and chondroitin sulfate. *International Journal of Biological Macromolecules*, 2001.

[31] Tatiana Konevskikh, Arkadi Ponossov, Reinhold Blümel, Rozalia Lukacs, and Achim Kohler. Fringes in FTIR spectroscopy revisited: Understanding and modelling fringes in infrared spectroscopy of thin films. *Analyst*, 2015.

[32] Heather J. Gulley-Stahl, Sharon B. Bledsoe, Andrew P. Evan, and André J. Sommer. The advantages of an attenuated total internal reflection infrared microspectroscopic imaging approach for kidney biopsy analysis. *Applied Spectroscopy*, 2010.

[33] Susanne W. Bruun, Achim Kohler, Isabelle Adt, Ganesh D. Sockalingum, Michel Manfait, and Harald Martens. Correcting attenuated total reflection-fourier transform infrared spectra for water vapor and carbon dioxide. *Applied Spectroscopy*, 2006.

[34] Åsmund Rinnan, Lars Nørgaard, Frans van den Berg, Jonas Thygesen, Rasmus Bro, and Søren Balling Engelsen. Chapter 2 - Data Preprocessing. In *Infrared Spectroscopy for Food Quality Analysis and Control*. 2009.

[35] Loong Chuen Lee, Choong Yeun Liong, and Abdul Aziz Jemain. A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum, 2017.

[36] Ghazal Azarfar, Ebrahim Aboualizadeh, Nicholas M. Walter, Simona Ratti, Camilla Olivieri, Alessandra Norici, Michael Nasse, Achim Kohler, Mario Giordano, and Carol J. Hirschmugl. Estimating and correcting interference fringes in infrared spectra in infrared hyperspectral imaging. *Analyst*, 2018.

[37] Sebastian Raschka and Vahid Mirjalili. Learning best practices for model evaluation and hyperparameter tuning. In *Python machine learning*, chapter 6, pages 185–217. Packt Publishing, 2 edition, 2017.

[38] Valeria Tafintseva, Evelyne Vigneau, Volha Shapaval, Véronique Cariou, El Mostafa Qannari, and Achim Kohler. Hierarchical classification of microorganisms based on high-dimensional phenotypic data. *Journal of Biophotonics*, 2018.

[39] Achim Kohler, Mohamed Hanafi, Dominique Bertrand, El Mostafa Qannari, Astrid Oust Janbu, Trond Møretrø, Kristine Naterstad, and Harald Martens. Interpreting several types of measurements in bioscience. In

Peter Lasch and Janina Kneipp, editors, *Biomedical vibrational spectroscopy*, chapter 15, pages 333–256. John Wiley & Sons, Inc., Hoboken, New Jersey, 2008.

[40] Kenneth P.H. Pritzker, S. Gay, S. A. Jimenez, K. Ostergaard, J. P. Pelletier, K. Revell, D. Salter, and W. B. van den Berg. Osteoarthritis cartilage histopathology: Grading and staging. *Osteoarthritis and Cartilage*, 2006.

[41] Pierre Mainil-Varlet, Boudewijn Van Damme, Dobrila Nesic, Gunnar Knutsen, Rita Kandel, and Sally Roberts. A new histology scoring system for the assessment of the quality of human cartilage repair: ICRS II. *American Journal of Sports Medicine*, 2010.

[42] B.O GmbH. OPUS Spectroscopic Software: reference manual, 2004.

[43] David S. Stoffer and Peter Bloomfield. Fourier Analysis of Time Series: An Introduction. *Journal of the American Statistical Association*, 2000.

[44] Harald Martens. The informative converse paradox: Windows into the unknown. *Chemometrics and Intelligent Laboratory Systems*, 2011.

[45] Guiyang Li, Mary Thomson, Edward Dicarlo, Xu Yang, Bryan Nestor, Mathias P.G. Bostrom, and Nancy P. Camacho. A chemometric analysis for evaluation of early-stage cartilage degradation by infrared fiber-optic probe spectroscopy. *Applied Spectroscopy*, 2005.

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

Postboks 5003
NO-1432 Ås
Norway