Anomaliedetektion in hyperspektralen Daten am Beispiel von Olivenöl

Masterarbeit
von

Nisse Knudsen

Fraunhofer Institut für Optronik, Systemtechnik und Bildauswertung

Fakultät für Informatik
Karlsruher Institut für Technologie

Verantwortlicher Betreuer:	Prof. Dr.-Ing. Jürgen Beyerer
Betreuender Mitarbeiter:	M.Sc. Julius Krause

Tag der Anmeldung:	01.12.2017
Tag der Abgabe:	31.05.2018

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissenschaftlicher Praxis in der zum Zeitpunkt der Anfertigung dieser Arbeit gültigen Fassung beachtet habe.

Karlsruhe, den 31. Mai 2018

Zusammenfassung

Olivenöl gehört weltweit zu den meist verfälschten Nahrungsmitteln. Aufgrund seiner Popularität unter westlichen Verbrauchern bildet es ein beliebtes Ziel für Lebensmittelbetrug. NIR-Spektroskopie ist eine Methode, um die Molekülkonzentration in Materie über ihre Vibrationen zu messen. Diese Molekülkonzentration in verschiedenen Wellenlängenregionen gibt bei Olivenöl einen Hinweis auf den Fettsäuregehalt und andere Substanzen. Während die Schätzung des Fettsäuregehalts bisher in vielen Arbeiten behandelt wurde, ist für die Detektion von unbekannten Substanzen bisher keine Standardmethode entstanden. Leider sind die unbekannten Substanzen diejenigen, die unerwartete Schäden für die Konsumenten verursachen können. Daher wird ein Algorithmus vorgeschlagen, der auf der Wavelet-Transformation basiert. Um für die Detektion von Absorptionsspitzen in NIR-Spektren geeignet zu sein, musste die klassische Wavelet-Transformation angepasst werden, unter anderem durch Unterstützung für nichtperiodische Wavelets, Normalisierung für eine vergleichbare Wavelet-Transformation unabhängig von der Sensorintensität, und durch Analyse optimaler Fenstergrößen, um die Unschärfe hinsichtlich Position und Ausmaß der Absorptionspitzen zu verringern. Der Algorithmus wurde auf Olivenöl- und anderen pflanzlichen Öl-Proben getestet, und seine Performance mit der Partial-Least-Squares-Methode (PLS) verglichen. Die PLS zeigte bessere Ergebnisse als die Wavelet-Analyse mit linearer Regression, kann aber keine unbekannten Absorptionen in einem Spektrum erkennen. Für eben diese Absorptionen wurde eine Kombination aus Wavelet-Analyse und Entfernungs-basierten Algorithmen vorgeschlagen. Das Ergebnis ist ein Wavelet-Analyse-Algorithmus, der Absorptions-Korrelationen aus jedem Spektrum extrahieren kann, die in anderen Algorithmen mit hochentwickelten Anomalie-Erkennungstechniken verwendet werden können.

Abstract

Anomaly detection in hyperspectral data by the example of olive oil

Olive oil is amongst the most adulterated foods world-wide. Its popularity amongst western consumers makes it a valuable target for food fraud. NIR spectroscopy is a method to measure molecule concentrations in matter due to its vibrations. For olive oil, this molecule concentration at different wavelength regions gives an indication about fatty acid content and other substances. While the estimation of a fatty acid content has been researched by many, for the detection of unknown substances no standard method has emerged so far. Unfortunately, the unknown substances are the ones which can cause unexpected harm to consumers. Therefore, an algorithm is proposed based on wavelet transformation. To be suitable for absorption detection in NIR spectra, the classical wavelet transformation had to be adapted by including support for non-periodic wavelets, normalization for a comparable wavelet transformation independent of sensor intensity and by analyzing optimal window sizes to decrease uncertainty regarding position and scale of absorption peaks. The algorithm was tested on olive and other vegetable oil samples, and its performance compared to partial least squares (PLS). The PLS performed better than the wavelet analysis linear regression, but can not detect unknown peaks in a spectrum. For those cases a combination of wavelet analysis and distance-based algorithms was proposed. The result is a wavelet analysis algorithm which can extract absorption correlations from any spectrum, which can be used in other algorithms for sophisticated anomaly detection techniques.

Contents

1. Introduction

Hyperspectral data have been used to detect anomalies in food for several years. The definition of anomaly in this context contains, amongst others, non-food elements, other-food elements or different origins, and can be generally subsumed as quality parameters. ElMasry and Nakauchi [7] give an overview of publications in which authors classified or predicted quality parameters in various foods, based on near-infrared (NIR) absorption spectra. The majority of works in this field uses partial least squares (PLS) method for prediction, since it is a more robust predictor than multiple linear regression, or linear discriminant analysis (LDA). The most commonly used method for feature reduction is principal component analysis (PCA). Analyzed food products include meat, fish, wheat, vegetables and fruits.
Both PLS and LDA are supervised learning methods, which require normal and anomalous at the same time. By setting the problem space to a known set of anomalies, most works find well performing models for those. It can be assumed that the real set of possible anomalies is infinite, and that the models described above do not perform well. In most research papers, the detection of unknown anomalies has been neglected in favor of precisely predicting contents of ingredients in a stable environment, with respect to sample size, temperature, light source, sensor etc. Applying any of the resulting models which were created in these perfect environment to other environment is hardly given. By simply choosing a sensor with a different bandwidth or lower resolution, most of the regression models will not work as precisely as they were developed to be. Therefore, a more robust method is required which can extract required information from a NIR spectrum, ideally semi-supervised. The normal state or region should be known for a product sample, while any other anomalous behavior can be immediately detect.
One potential new method to extract information about anomalies is the wavelet transformation. By convolving a wavelet over a signal, it can detect location and scale of a wavelet function. The continuous and discrete wavelet transformations have been defined to detect periodic signals. Therefore, an adaptation is required that works with non-periodic signals, for example Lorentz or Gauss functions.
In this thesis it will be evaluated, which information from a NIR absorbance spectrum is required for a successful anomaly detection. By the use of the wavelet transformation, which will be adapted to the needs and characteristics of NIR spec-

tral data, these information will be extracted and an exemplary anomaly detection algorithm will be implemented. The work towards a wavelet analysis algorithm is based on patent application *DE102017220103.7* (Krause [23]).
The food products used for experiments and model evaluation are two extra virgin olive oils (hereinafter referred to as olive oils) and four other vegetable oils. Olive oil has homogenous properties, which makes measuring the transmittance in NIR spectra reliable, especially across different sensors and measuring points within the sample. Additionally, the adulteration of olive oil is such a major business in food fraud, that legislators create special regulations for olive oil (European Commission [8]) in contrast to any other cooking oil. Since these regulations are in place without any obligatory quality analysis for products on the market, the majority of olive oils still don't meet these quality standards (Stiftung Warentest [32]). The anomalous space of olive oil experiments will include mixtures of olive oil with all other selected oils.
The result of this thesis will be an informed analysis how to create a well-performing algorithm for anomaly detection near-infrared spectral data by the example of olive oil. Ideally, this algorithm is generalized and detects anomalies in other food products in a sufficient way.

2. Related Work

Armenta et al. [2] detail why the detection of quality (or vice-versa anomalies) in olive oil industry products is an important topic. Despite olive oil being considered a provider of healthy unsaturated fats, it is also one of the most regulated cooking oils in the European Union (European Commission [8]). According to that regulation, they identify three types of virgin olive oils: *extra virgin*, *virgin* and *lampante*[1] *virgin*. Additionally, olive oils that are mixed with different oils or processed using solvents can not be labeled *virgin*. They list the following as main physiochemical parameters for olive oil classification and provide a short description for each (Armenta et al. [2][p. 571]):

Free acidity "Acidity is considered as the percentage of free fatty acids (FFA), measured by weight of free oleic acid that it contains."

Ultraviolet Absorption "The extinction coefficient value at the wavelength 232 nm [...] is a spectrophotometric feature for detecting oxidative reactions and mixtures with other olive oils. Meanwhile, absorption at 270 nm [...] is a parameter related to the detection of oils previously treated with alumina or other unusual compounds."

Peroxide value "The index of peroxide expresses the quantity of peroxide contained in 1 kg of oil. It is a quality parameter which permits to measure oxidative alterations, thus indicating possible alterations of some nutritionally interesting components [...]."

Moisture and volatile matter "Moisture and volatile matter is the loss in mass undergone by the product on heating at 103 °C ± 20 °C [...]."

Insoluble impurities "Insoluble impurities content is the quantity of dirt and other foreign matter insoluble in n-hexane or light petroleum [...]."

[1] **lampante** refers to olive oil which is generally not considered suitable for human consumption due to its oleic acidity over 3.3 %(w/w). It is a result of bad olive fruits or thermal processing.

Trace metal "Trace metal parameter concerns the amount, in $\mu g\,kg^{-1}$, of copper and iron in all types of olive oils, including contaminants from bleaching earth and/or catalysts [...]"

Afterwards they review and compare different publications that worked on detecting one or more of these characteristics using NIR based methods. Most works analyzed olive oil in combination with other edible cooking oils, and applied *Partial Least-Squares* method (PLS) *Principal Component Analysis* (PCA), *Factor Analysis* (FA) or a combination of them. The NIR bandwidth ranges from 850 to 2200nm; the sample sizes from 126 to 216. For some references there was no information about the bandwidth or samples size. With respect to acidity and peroxide index regression, the results from different works showed high determination and low errors.
In Armenta et al. [1] 69 edible oils (olive, sunflower, seed, maize) are used for calibration, validation and prediction of the acidity / peroxide quantification using PLS. In addition, hierarchical clustering was applied to classify edible oils into origins based on NIR spectra. The ground truth was determined by chemical analysis, the NIR transmittance measurement by using a *Bruker* FT-NIR spectrometer with a range from 800 to 2500nm. In Armenta et al. [1][section 2.3], the nominal resolution is given as $4\,cm^{-1} = 2\,500\,000\,nm$, which appears to be an error in units; therefore no assessment regarding the resolution can be made. They found that while spectra are generally very similar among edible oils, olive oil spectra show the most differences to other edible oils. By analyzing the absorption, the authors identified peaks at or around 1207, 1391, 1408, 1715 and 1734nm, which they attribute to the second and first overtones of CH molecule combinations stretching vibrations. The (visually) most significant differences between olive oils and the rest of the edible oils were found to be at wavelengths 1668, 1715 and 2144nm. Regarding acidity quantification, the best performing model for olive oil was a PLS model with 13 factors and linear removal preprocessing; the resulting RMSE in prediction (RMSEP) was 0.034 %(w/w). For predicting acidity in the other edible oils, another PLS model performed best, i.e., PLS with 10 factors and RMSEP of 0.037 %(w/w). A combined model applicable to all edible oils required 20 factors and resulted in a RMSEP of 0.083 %(w/w). They conclude that "for PLS-NIR acidity determination in edible oils the use of specific models as a function of the oil nature provides a clear enhancement as compare with models created from a pooled sample population of all types considered [...] and, in all the cases, it involves a significant reduction of the number of factors required." (Armenta et al. [1][section 3.5]).
Similarly, models were trained and evaluate to predict the peroxide index in olive and other edible oils. The results were analogous to the acidity prediction: training and using two separate PLS models for peroxide index prediction yielded better results than one shared model.
To validate if the trained models could also be used for acidity and peroxide index determination in a new edible oil set, they used 67 olive oils, 6 sunflower oils and 9 seed oils from another season and predicted the contents. Fig. 2.1 shows the coefficient of determination (R^2) for acidity than peroxide index prediction in the new season oils. The R^2 decreased significantly for all test cases, except acidity in olive oil. It should be noted that between both season's oil measurements, the NIR light sources were replaced. Since the correlation graphs in Fig. 2.1 and the original data samples have "statistically comparable slopes" (Armenta et al. [1][section 3.9]), they conclude that the proposed models are robust across light sources and seasonal

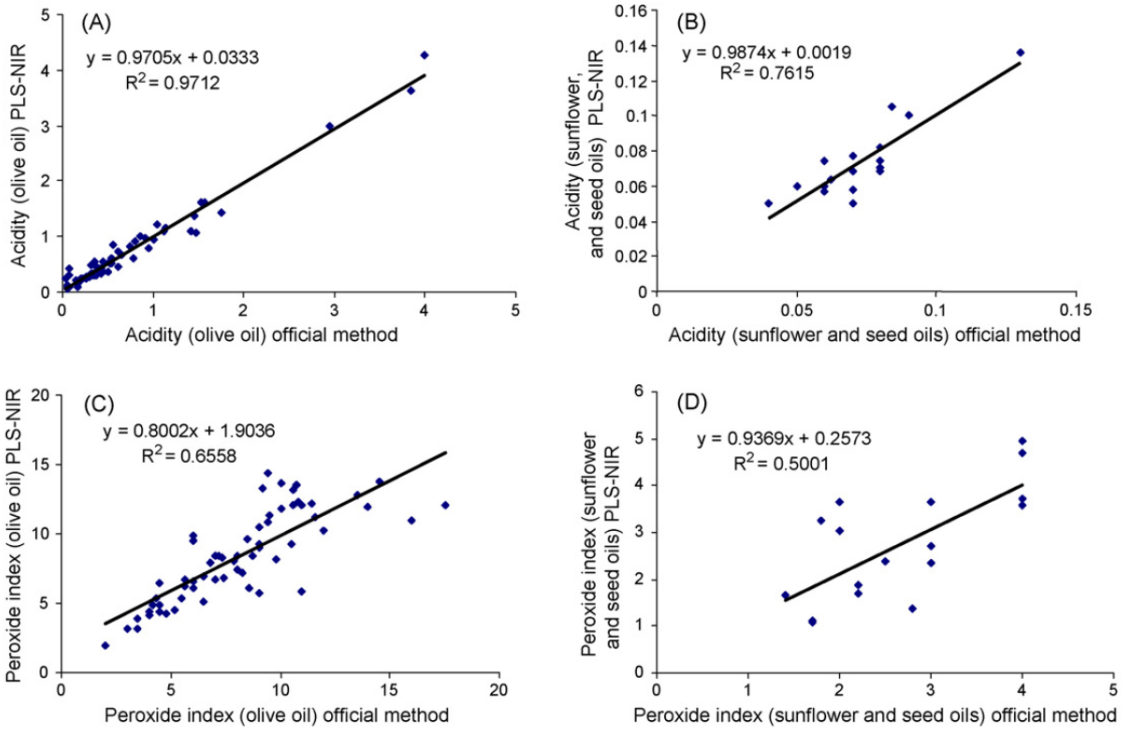


Figure 2.1: Correlation plots for (A) acidity of olive oils, (B) acidity of sunflower and seed oils, (C) peroxide index of olives oils, (D) peroxide index of sunflower and seed oils. Adapted from Armenta et al. [1][p. 336] with permission.

products. There is no information on the regional origin of the oils. Due to the fact that (a) Spain is by far the largest olive oil producing country in the world (cf. Fig. 2.2), (b) the work was published by researchers of *Universitat de València* and (c) the oil samples were provided by the *Laboratorio de Salud Pública de la Generalitat Valenciana*, it can be assumed that the majority of the oils have their origins in Spain. Therefore, any transferability of the models to products of other regions (e.g., Italian or Greek olive oil) and seasons remains to be validated.

The author's effort to cluster the edible oils regarding their flower / seed / fruit sources was done by *Hierarchical Cluster Analysis* (HCA) using the *Ward linkage* method. The incrementally merged partitions can be visualized in a dendrogram, as shown in Fig. 2.3. On the one hand, Fig. 2.3a shows that olive oil can be distinctively classified when compared to the other edible oils. On the other hand, Fig. 2.3b shows that for the analyzed sunflower, seed and maize oil no clear distinction is possible, because partitions are mixed in a very early clustering stage. This allows the conclusion that olive oils possess characteristics that none of the other oils do, but leaves open the question how mixtures of olive and non-olive oils are separated and what the limit of detection with respect to mixing ratio is. Additionally, the authors don't provide any information what the chemical interpretation of the similarity inside partitions is, e.g., acidity, peroxide index or something different within the spectrum.

Woodcock et al. [40] used NIR spectra of extra virgin olive oil samples to clas-

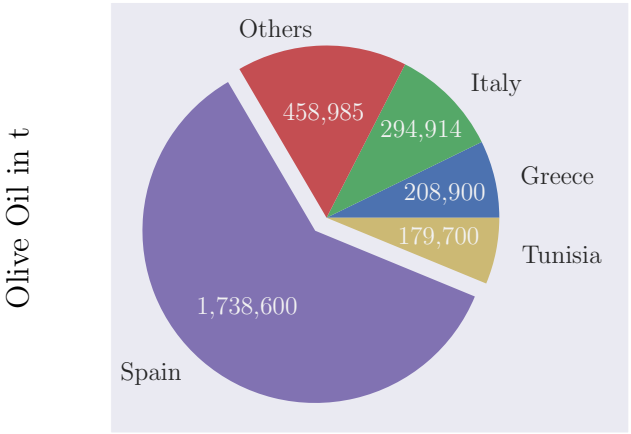
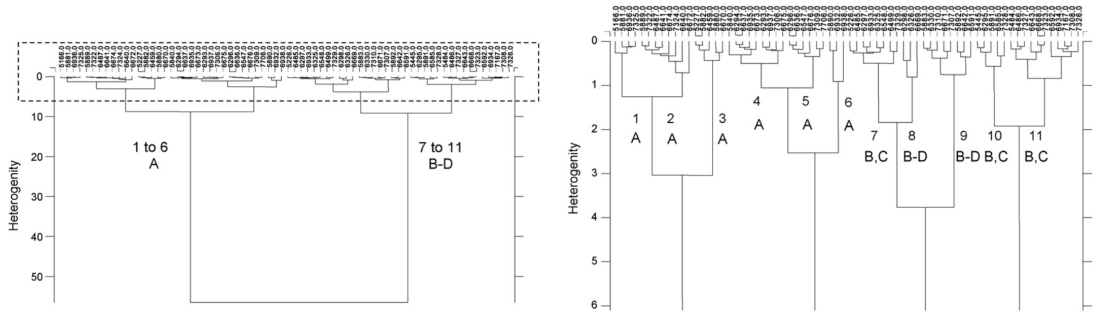


Figure 2.2: Worldwide virgin olive oil production in 2014 (Food and of the United Nations [10])



(a) Clear dissimilarity between A and B-D. Dashed area shown in Fig. 2.3b.

(b) Pure partitions for olive oil, yet, mixed partitions for other oils.

Figure 2.3: Hierarchical Clustering partitions of edible oils. Numbers indicate cluster numbers, letters classes of oils: (A) olive, (B) sunflower, (C) seeds, (D) maize. Adapted from Armenta et al. [1][p. 332] with permission.

sify if a sample had its origin in Liguria[2] or in another region. It's important to note that this was only a binary classification; the authors did not distinguish the other regions, but simply labeled them as *non-Ligurian*. They compiled an overview of functional group assignments for the olive oil spectra in the range from 1100 to 2498nm, as can be seen in Table 2.1. Using PCA, they identified the most important peaks at 1211, 1727, 1761, 2310 and 2350nm in the first component. These findings and assignments are in accordance with the peak findings from Armenta et al. [1], and provides a valuable source for interpretation of absorption peaks in this thesis.

The classification was trained using the *Partial Least-Squares Discriminant Analysis* (PLSDA), regressing a dummy predictor variable (1: Ligurian; 0: Non-ligurian), which was thresholded at 0.5. Every model was trained using raw data, first deriva-

[2]**Liguria** is a coastal region in North-west Italy.

Table 2.1: Functional Group Assignments in Olive Oil Spectra. Most important absorption peaks in bold print (adapted from Woodcock et al. [40][p. 11522]).

wavelength (nm)	functional group	assignment
1168	CH3—	C—H stretch second overtone
1211	**—CH2—**	**C—H stretch second overtone**
1391	CH3—	2C—H str + C—Hdef
1414	—CH2—	2C—H str + C—Hdef
1664	cis R1CH=CHR2CH3—	cis CH
1727	**—CH2—**	**C—H first overtone**
1761		**C—H first overtone**
1901	C=O str	second overtone
1931	C=O str	second overtone ester
2124	—COOR	C—H str + C=O str
2145	—HC=CH—	=CH str+ C=O str
2176	—HC=CH—	CH assym str + C=C str
2310		**CH combinations and deformation**
2350		**CH combinations and deformation**

tives, second derivatives and *Standard Normal Variate* (SNV) normalized data. Afterwards, the relatively correctly classified oils for both classes were used as a performance measure. The results showed that if training one model with the full data set of $n = 913$ samples from 3 seasons and a ratio of 25% Ligurian vs. 75% non-Ligurian oils, it has a bias toward the non-Ligurian class and classifies at maximum 50% of the Ligurian oils correctly. If using a model that was trained using 50% Ligurian and non-Ligurian oils, the correct classification of Ligurian oils was significantly better of up to 93%. For both models the results were best when not using raw or SNV data, but using the second derivative.
The authors conclude that "NIR spectroscopy coupled with chemometric analysis of data provides a promising tool for geographical classification of olive oils", but also recognize that a "Transition of this technique into an industrial setting would require the establishment of a larger database that would take into account greater variability in factors such as weather conditions at the time of harvest" (Woodcock et al. [40][p. 11524]). Similarly to Armenta et al. [1], the trained model works well on the provided samples, but is tailored to regress or classify on oils from specific origin or even season. A generalization of the methods to other oils is missing.

When searching for hyperspectral detection algorithms, a lot of results are rooted in *Hyperspectral Imaging* (HSI) with applications for aerospace engineering, e.g., drone or satellite imagery of planets' surfaces, in mind. This is why almost all of the "traditional" methods for anomaly detection require conditions which are not present in food sensing, and even less in oil sensing. Manolakis et al. [24] asked the question "Is there a best hyperspectral detection algorithm?". They introduce the term *target* as synonym for a spectral signature of certain interest. Target detection algorithms, which can be separated into *Spectral Anomaly Detection Algorithms* and *Spectral Matching Detecting Algorithms*. The former detects unknown targets (i.e., spectral signatures), the latter when the signature of a target is known, e.g.,

extracted from a spectral library as the *USGS Spectral Library* (Kokaly et al. [22]). They observed that spectral matching algorithms work better than anomaly detection algorithms, because the target is already known. Additionally, they mentioned that to successfully detect targets, they have to be distinguishable from background data. Applying this requirement to the measurement of homogeneous matter or single-point measurements, these algorithms will not work.
One of the most prominent and improved algorithms is the *Reed-Xiaoli Algorithm* (RX) (Reed and Yu [30]). The algorithm is simple to use by only taking an image as input. The image is then segmented into smaller windows, of which the covariance matrix is calculated and the to-be-tested target is the center pixel. The classification if a pixel is an anomaly or not is done using a *Constant False Alarm Rate* (CFAR), which is a parameter that needs to be set or trained. This algorithm could be used to detect anomalies in 2D food HSI, like apples with bad spots, but does not return the type of anomaly, and requires spectral differences to trigger the CFAR not too often as false alarm, but also to not miss any (possibly dangerous) anomalies.

Kim et al. [21] show the benefits of knowing the target anomaly before-hand and how it can be used to train a supervised anomaly detector. They cite fecal (from cows, deer and humans) contaminated apples as a major source for *E. coli*, after these had been processed to apple juice and consumed by children in the U.S.. After selecting four cultivars of apples, these were artificially contaminated with thick and thin patches of dairy cow feces. For measurement they used visible and near-infrared reflectance imaging in a range of 450 to 851nm using 110 channels. Analyzing the reflectance at specific wavelengths, they found that for some wavelengths, the results differed significantly depending on if the side of the apple was sun-exposed (higher reflectance) or shaded (lower reflectance). Especially at lower wavelengths this difference was visible, but decreased with higher wavelengths. Using PCA on the HSIs, PC-3 and PC-2 showed the highest performance in detecting fecal contamination on the surface, even as very thin layers. The classification was done using a thresholded mask.
The authors do not apply other methods for classification, e.g., HCA (as seen in Armenta et al. [1]), and do not provide any performance results regarding incorrect pixel classifications. Yet, they recognize in their conclusion that “the reflectance imaging method detected only a fraction of the thin feces spots, indicating a lack of sensitivity” (Kim et al. [21][p. 2037]).

3. Anomaly Detection in NIR Spectroscopy

In this chapter, the requirements for a successful anomaly detection in NIR spectra are described. First, the basic principles of NIR spectroscopy are introduced. Afterwards, domain-independent types of anomalies are introduced and outlined which are relevant for NIR spectroscopy, followed by a combination of NIR fundamentals and anomaly definitions to evaluate what information are required to theoretically detected anomalies in NIR data.

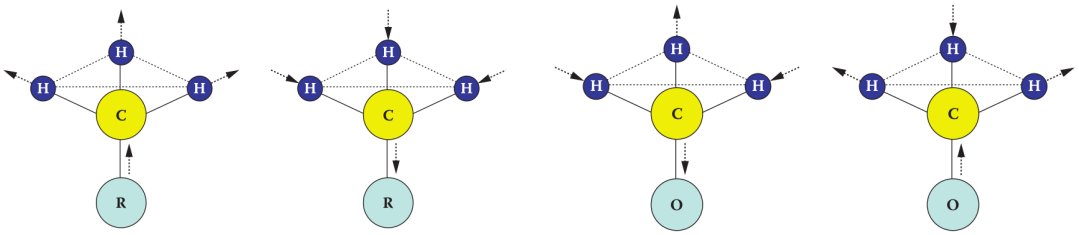
3.1 Fundamentals of NIR Spectroscopy

NIR spectroscopy is the analysis of matter by measuring the absorbance at specific wavelengths in the *near-infrared* (NIR) spectrum. The NIR region ranges from 280 to 2500nm (cf. Workman and Weyer [41, table 1.3]). The measured absorbance is due to "the periodic motions (or vibrational modes) of atomic nuclei within their respective molecules.[...] The energy level in a molecule is described as the sum of the atomic and molecular motions due to translational, rotational, vibrational, and electronic energies. Translational energy has no effect on the molecular spectra, whereas the other motions do affect the spectral characteristics"(Workman and Weyer [41, p. 2]). Fig. 3.1 shows a simplified view how stretching of molecules work. For the symmetrical case, all atoms move towards either the same upper direction or lower direction. In contrast, the asymmetrical movement can be seen as a vertical shrinkage with horizontal expansion and vice versa. For both exemplary molecules, the C atom does not stretch but is rather the focal point of all movements. In general is the NIR region most active for X—H bonds, e.g., C—H, according to Workman and Weyer [41, p. 8], which includes alkanes and alkenes that saturated and unsaturated fatty acids belong to. These fatty acids are an essential nutrition component of olive oil, which is to be analyzed later in this work.
NIR is neighbored to the *(mid-)infrared* (MIR) region, which is defined from 2500 to 25 000nm. Between both regions exists an important relationship: if molecules start vibrating in the MIR region, it is defined as the *fundamental molecular vibration* ν. This vibration has several *overtones* $\{2\nu, 3\nu, 4\nu, 5\nu\}$, which have their absorbance

Table 3.1: "Relative Band Intensities: MIR versus NIR for C—H Stretch". Adapted from Workman and Weyer [41, table 1.2]

Region	Band	Wavelength Region	Relative Intensity
MIR	Fundamental (ν)	3380-3510 nm	1
NIR	First overtone (2ν)	1690-1750 nm	0.01
NIR	Second overtone (3ν)	1127-1170 nm	0.001
NIR	Third overtone (4ν)	845-878 nm	0.0001
NIR	Fourth overtone (5ν)	690-770 nm	0.00005

wavelength in the NIR region. Since these absorbances are only overtones, their relative intensity compared with the fundamental vibration is decreasing. Table 3.1 shows how the relative intensity decreases with every overtone. One can imagine how that exemplary C—H stretch vibration is much more difficult to detect in NIR than in MIR region, just by the intensity. Additionally, the table also reveals how a molecular vibration does not necessarily occur only at a specific wavelength, but rather in a region.



(a) "Methyl symmetrical stretching of CH_3". Adapted from Workman and Weyer [41, Fig. 1.1] with permission.

(b) "Asymmetric stretching of O—C—H". Adapted from Workman and Weyer [41, Fig. 1.6] with permission.

Figure 3.1: Exemplary symmetric and asymmetric stretching vibrations.

According to *Beer's law*, the measured absorbance A is proportional to the concentration c of a molecule in the matter (cf. Workman and Weyer [41, p. 2]). Furthermore, the length of the light beam within the sample holder l and the molecular vibration absorptivity ϵ need to be known. While A and l can be measured, the absorptivity needs to be calculated experimentally by "careful measurements of the absorbance of a compound" (Workman and Weyer [41, p. 2]). Afterwards, Beer's law $A = \epsilon cl$ can be used to calculate ϵ.

In applied NIR spectroscopy, the absorbance is not measured directly, but either the reflectance R or the transmittance T of light from/through the matter. Fig. 3.2 outlines how transmittance T is measured based incident light through a sample.

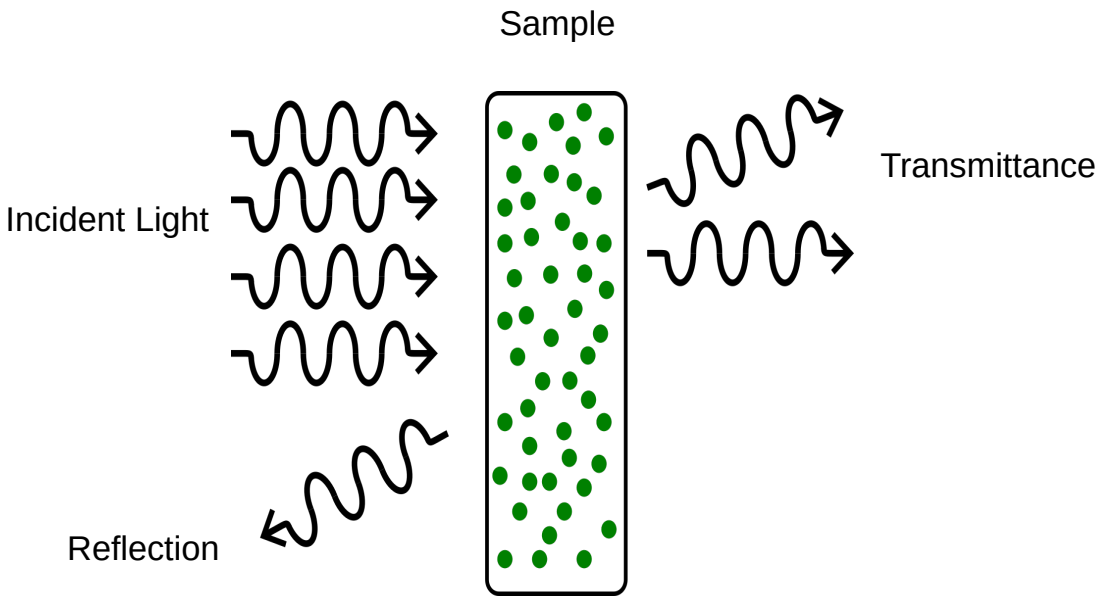


Figure 3.2: Outline of light transmittance through sample.

Fortunately, there are simple operations to convert either of them to an absorbance value (implicitly using Beer's law):

$$T = \frac{I}{I_0} = 10^{-\epsilon c l} \Longrightarrow A = -\log_{10}\left(\frac{I}{I_0}\right) = -\log_{10} T = \epsilon c l \tag{3.1}$$

and

$$R = \frac{I}{I_0} = 10^{-\epsilon c l} \Longrightarrow A = -\log_{10}\left(\frac{I}{I_0}\right) = -\log_{10} R = \epsilon c l \ . \tag{3.2}$$

Workman and Weyer [41, p. 6] recommend specific path lengths l for optimal measurement results in the NIR region. For the region from 800 to 1100nm they recommend $5\text{cm} \leq l \leq 10\text{cm}$, and for 1100 to 2500nm $1\text{mm} \leq l \leq 20\text{mm}$.

Harmonic and Anharmonic Oscillator Models

The vibrations of molecules can be modeled as an harmonic or anharmonic oscillator. The underlying model for both oscillators is that two atoms in a molecule are a dipole. If the molecule is *infrared active*, the atoms start vibrating with a certain frequency ν. Fig. 3.3 visualizes the movement between two atoms. Without any IR light, the atoms have a relaxed distance of x_{equib}. When they start vibrating, they move towards and apart from each other. The displacement is limited by x_{min} and x_{max}. The frequency ν for a diatomic molecule can be modeled using a simple mass-spring model, according to Workman and Weyer [41, eq. 1.12]:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{3.3}$$

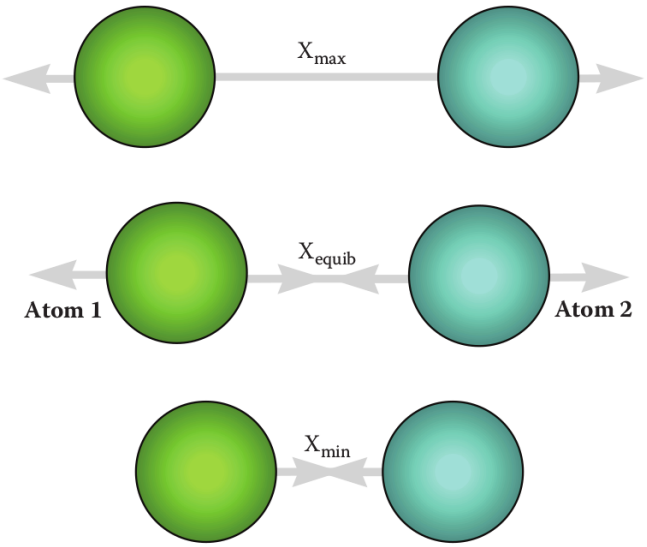


Figure 3.3: "Model of infrared active molecule as a vibrating dipole between two atoms". Adapted from Workman and Weyer [41, fig. 1.12] with permission.

with m being the masses of two atoms and k being "a force constant that varies from one bond to another" (Workman and Weyer [41, p. 10]). Therefore, a greater mass yields a lower frequency and vice versa. When the atoms are outside their relaxation state, the energy of the molecule increases. This energy function can be modeled using the previously mentioned harmonic or anharmonic oscillator. The harmonic oscillator models the energy equally for x_{min} and x_{max}. Fig. 3.4a shows the energy as a function of displacement. Function A and B differentiate by bond strength, with B having a stronger bond. As a result, the x_{min}, x_{max} are smaller and the energy increases faster than for A. In contrast, the anharmonic oscillator in Fig. 3.4b expects that bonds can break when displacement too further away from each other. Additionally, if atoms are too close to each other, they are modeled to repel each other.
According to Demtröder [5, sec. 7.4], the resulting spectral line (or the absorption line) in the spectrum has approximately the shape of a Lorentz function:

$$L(x) = \frac{2s}{\pi(s^2 + 4(x_0 - x)^2)} \tag{3.4}$$

with

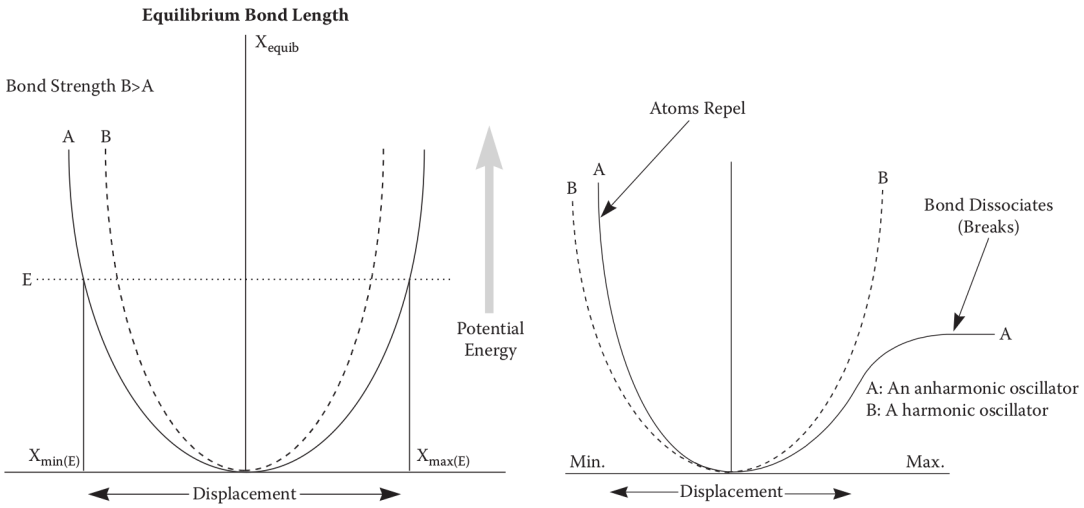
$$\int_{-\infty}^{\infty} L(x)\mathrm{d}x = 1 \tag{3.5}$$

and s being the scale parameter. The maximum value of a Lorentz function is determined by

$$L(x_0) = \frac{2}{\pi s} \ , \tag{3.6}$$

having half maximum value at

$$x = (x_0 \pm \frac{s}{2}) \ . \tag{3.7}$$



(a) Energy between atoms depending on their displacement. x_{max} is a limit and no bonds are modeled to break. Adapted from Workman and Weyer [41, Fig. 1.13] with permission.

(b) The anharmonic oscillator models that bonds can break if displacement becomes too large. Adapted from Workman and Weyer [41, Fig. 1.15] with permission.

Figure 3.4: Harmonic and anharmonic oscillator in comparison.

Therefore, its *full width at half maximum* (FWHM) is s. This property makes any handling of Lorentz functions very intuitive. Knowing the Lorentz shape of an absorption peak in the spectrum will be used in the wavelet transformation in Ch. 4. Eq. 3.8 and 3.9 show the first and second derivative of the Lorentz function, which can be used on derived spectral lines.

$$\frac{\mathrm{d}L(x)}{\mathrm{d}x} = \frac{16s(x_0 - x)}{\pi(s^2 + 4(x_0 - x)^2)^2} \tag{3.8}$$

$$\frac{\mathrm{d}^2L(x)}{\mathrm{d}x^2} = \frac{16s(-s^2 + 12(x_0 - x)^2)}{\pi(s^2 + 4(x_0 - x)^2)^3} \tag{3.9}$$

3.2 General Anomaly Detection

Chandola et al. [4, chapter 1] define *Anomaly Detection* as "the problem of finding patterns in data that do not conform to expected behavior". To decide when expected (or *normal*) behavior is not present, first we need to define what normal behavior is.

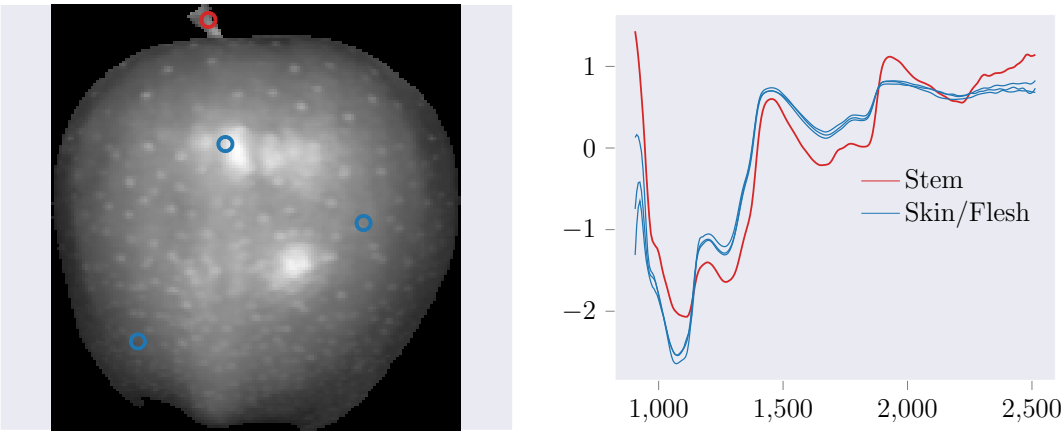
A naive approach is defining of a normal region, and classifying all observation outside of that region as an anomaly. This will most likely result in unsatisfactory classification results, due to challenges that are in most cases not covered by a fixed border between normal and anomalous patterns (adapted from Chandola et al. [4, section 1.2]):

- In real applications, a fixed border is often too precise. A pattern close to the border could be a normal behavior, but due to binary classification fall into the anomalous region.
- Attacks to circumvent the detection will be designed to appear like normal behavior, thus requiring higher complexity from the detection method.
- Normal behavior is an evolving definition, and needs to be extended or updated over time.
- The notion of anomaly depends on the applied domain. Some anomalies simply result in lower-than-expected quality, while some result in danger or damage.
- Depending on the detection method, labeled training data is required. While the normal space is finite, the anomalous space is infinite and therefore requires a broad distribution in labeled training data.
- The presence of noise is inevitable when measuring data. While noise is explicitly not considered an anomaly, it makes the detection of latter more difficult. Noise can be smoothed by applying appropriate filters, but requires parameters, which do not remove anomalies and make them undetectable.

These problems need to be incorporated into the design for any anomaly detection algorithm. Several of these points will be discussed during this work, e.g., while adapting the wavelet transformation method or when detecting peaks from a wavelet surface.

3.2.1 Types of Anomalies

According to Chandola et al. [4, section 2.2] anomalies can be classified as either *Point Anomaly*, *Contextual Anomaly* or *Collective Anomaly*. Transferring the definition of these classes to the scope of generic anomaly detection in hyperspectral data, only the contextual anomaly is relevant. While point anomalies can point to anomalously measured data points, they do not consider any context within the data instance itself, e.g. relation to other data points. A simple example shows why keeping the context is very important: Assuming the measurement of an apple and vegetable oil. The spectral signature of the vegetable oil has a very small first overtone absorption at 1900 nm wavelength, while the apple has a very high absorption at that wavelength. Increasing absorption at this wavelength is due to increasing H2O content. While a high absorption due to water is highly unusual for vegetable oil, it is perfectly normal for a (fresh) apple. Simply regarding data points as point anomalies does not suffice in this domain.
Collective anomalies expect to have a data set which a single data instances can be compared to. This requires the recording and labeling of several data instances, which also have to be very similar. A possible application is in the series production of goods, where a high percentage of goods have normal attributes, and only a few rare instance are anomalous. For example, Fig. 3.5 shows signatures of an apple stem inserted into a series of apple. The stem is an anomaly which has a different absorption than the flesh, which can be used to remove such an object from the production line (e.g., in apple juice production).



(a) Location of Spectra in Reflectance Image

(b) 2x Savitzky-Golay filtered, SNV spectra

Figure 3.5: Anomaly introduced through non-edible apple parts

Where point anomalies fall short, and collective anomalies require too many data, the contextual anomaly thrives. It requires a relatively small data set of normal examples and considers relations between data points within instances. Chandola et al. [4, subsec. 2.2.2] state that "the anomalous behavior is determined using the values for the behavioral attributes within a specific context", which fits the vegetable oil vs. apple example from above. In that case the behavioral attribute is the absorption due to water, but seen in the context of regarding an apple or vegetable oil. A contextual anomaly consists of two elements: *contextual attribute* and *behavioral attribute*. The contextual attribute is the location where an anomaly appears, e.g., the wavelength (region) in a NIR spectrum. The behavioral attribute is the value at a contextual attribute, e.g., the absorbance intensity at a given wavelength.

3.3 Anomalies in NIR Spectra

Combining the knowledge from Sec. 3.1 and 3.2, requirements for an anomaly detection in NIR spectra can be derived. Due to the fact that absorbance intensity are correlated with molecular concentration inside the sample, any changes in the spectrum could be classified as an anomaly. Basically, three different changes can occur for every wavelength region inside the spectrum:

Changed Absorption The spectrum has a defined absorbance at a certain wavelength λ, which has suddenly a different intensity. It might be greater or less intense, but a change in molecular concentration has occurred.

New Absorption In a normal state, there was no absorption peak defined at λ. In the anomalous sample, a new peak is added where none should be.

Removed Absorption This is the extreme case of *changed absorption*: a normal peak is not existent in the measured spectrum anymore and therefore no concentration of the associated molecules is present in the sample.

The works presented in Ch. 2 use methods like PLS or PCA which identify the highest variances in the training data. On the one hand, this is an excellent way to reduce high dimensional data into a subspace which has fewer features, and then have a regression or classification algorithm learn these features. On the other hand, it (a) requires data that are very similar to the training data with regards to resolution, any preprocessing, intensity etc. and (b) discards any information that were not present in the training data. Considering that new absorptions can be the most severe anomalies, e.g., industrial solvents in olive oil, these methods do not sufficiently work for anomaly detection. Of course, any anomaly could be represented in the training data and be learned by any algorithm, but the space of anomalies, e.g. impurities or adulteration for olive oil, is practically endless. Therefore, a misrepresentation exists in the training data. One extreme but to the point example is described in Ch. 6 by the example of *Spanish toxic-oil syndrome*: until today there is no exact knowledge about which chemical substance specifically caused the outbreak of the disease, only that rapeseed oil was de-denaturated and some substance created during that process. Such an anomaly can never be represented in any training data.
To detect all three types of absorption anomalies in NIR spectral data, three information are required for every absorbance peak:

- Position [nm]
- Scale [nm]
- Intensity [a.u.]

The absorbance intensity is measured in *arbitrary units* [a.u.], due to its conversion from transmittance T or reflectance R (cf. Eq. 3.1 and 3.2). The $\log_{10}$ requires the previous scaling of the data to $[0, 1]$, which loses any information about intensities in any other unit. Additionally, if even any information were kept regarding the original transmittance or reflectance intensity, these would differ from sensor to sensor and prohibit any generic anomaly detection algorithm from working properly. Therefore, instead of using the absolute value of any intensity to detect anomalies, the relation between each of the values could serve as a measure. This would underline the concept of a contextual anomaly: if absorption at λ_1 is usually two times higher than absorption at λ_2, a measured relation of $\frac{\lambda_1}{\lambda_2} = 1.5$ can be defined as anomalous. In case a new peak is present, λ_{new} will be compared against λ_1, which was before zero, but is now greater zero, i.e., an anomaly.

In the following chapter an approach is presented how the previously listed three information position, scale and intensity can be evaluated for any spectrum, without the requirement for any training data.

4. Wavelet Transformation for NIR Absorption Spectra

In the previous chapter, required information for a successful anomaly detections were outlined. These were defined as *position*, *scale* and *intensity* of any absorbance peak in a NIR spectrum. To acquire these information regardless of the sensor resolution, maximum intensity or measured matter, the use of a method based on wavelet transformation is proposed. After introducing basic wavelet transformation definitions, a formal definition of the adapted wavelet transformation is provided in Sec. 4.2.

4.1 Fundamentals of Wavelet Transformation

Wavelet transformation is a method that is more location-aware compared to Fourier transformation when detecting frequencies in a signal. While the Fourier transformation does not specifically locate the occurrence of a certain frequency in the time-domain (except when using derived methods, e.g., Short Time Fourier Transform, which transforms smaller windows from time- to frequency-domain), the wavelet transform tries to do it by using finite oscillation wavelets. Therefore, a wavelet oscillates only in a certain part of the signal, which can be interpreted as a window under test. The term *wavelet* itself "corresponds to the French term 'ondelette,' which means 'small wave'" (Beyerer et al. [3, p. 728]). The following definitions are taken from Beyerer et al. [3] and any alterations are made according to their nomenclature, as long as applicable.

4.1.1 Continuous Wavelet Transformation

The most fundamental definition of the wavelet transformation is done using continuous functions, i.e., *Continuous Wavelet Transformation* (CWT). Given a continuous signal $g(x)$ and a generic wavelet $\psi_{s,\tau}(x)$, where s denotes the scale parameter of the wavelet function (e.g., the FWHM of a Lorentz function) and τ denotes the shift along the x-axis, the wavelet transformation is defined as

$$\Gamma_\psi(s,\tau) \coloneqq \int_{-\infty}^{\infty} g(x)\psi^*_{s,\tau}(x)\mathrm{d}x \tag{4.1}$$

with

$$\psi_{s,\tau}(x) := \frac{1}{\sqrt{s}}\psi\left(\frac{x-\tau}{s}\right), \quad s,\tau \in \mathbb{R},\ s > 0. \tag{4.2}$$

Fig. 4.1 shows shifted and scaled instances of the mother wavelet $\psi(x) = e^{-\frac{x^2}{2}}\sin(2\pi x)$ (cf. Beyerer et al. [3, eq. 15.17]). The shift parameter can be used to detect correlations between signal and wavelet at specific positions.

The inverse CWT is defined using

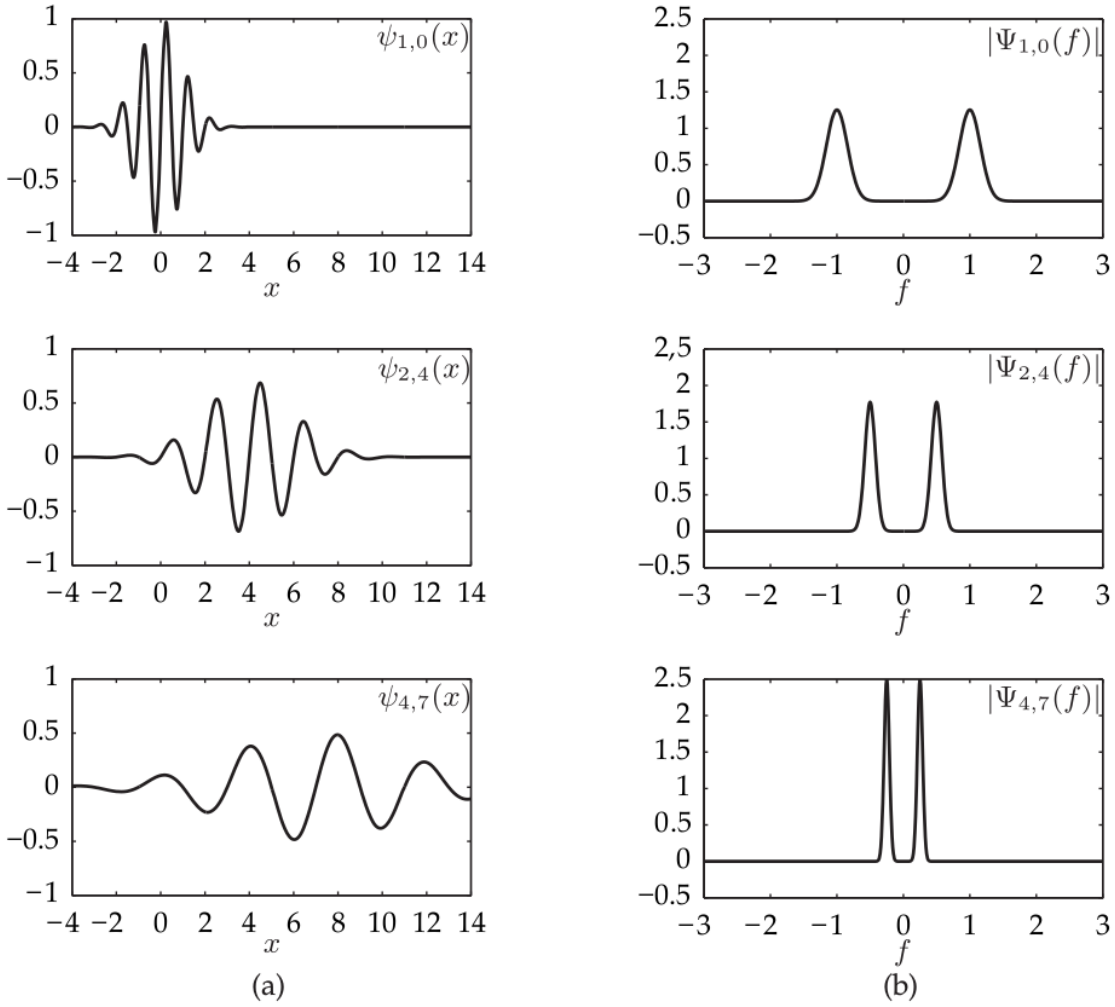


Figure 4.1: Three wavelets with different s and τ and their respective Fourier transforms. Adapted from Beyerer et al. [3, p. 730] with permission.

$$g(x) = \frac{1}{C_\psi}\int_0^\infty \int_{-\infty}^\infty \Gamma_\psi(s,\tau)\frac{\psi_{s,\tau}(x)}{s^2}\mathrm{d}\tau\mathrm{d}s \tag{4.3}$$

with

$$C_\psi = \int_{-\infty}^\infty \frac{|\Psi(f)|^2}{|f|}\mathrm{d}f \quad , \tag{4.4}$$

where

$$\Psi(f) = \mathcal{F}\{\psi(x)\} \qquad . \tag{4.5}$$

Beyerer et al. [3, p. 729] note that "the mother wavelet must satisfy the admissibility condition $C_\psi < \infty$. This implies that the DC component of the wavelet is zero". While this condition holds for basic periodic functions as sine or cosine, the DC component of a Lorentz wavelet is greater than zero.

4.1.2 Discrete Wavelet Transformation

Often, the signal itself is not represented by a continuous function but by discrete data, e.g., output of a sensor as an NIR spectrometer. In such cases the *Discrete Wavelet Transformation* (DWT) can be applied. It is, given by the discretization of Eq. 4.1 (cf. Beyerer et al. [3, eq. 15.20])

$$\Gamma_\psi(s,l) = \frac{1}{\sqrt{s}} \sum_n g_n \psi^* \left(\frac{(n-l)\Delta x}{s} \right) \tag{4.6}$$

with

$$g_n \coloneqq g(n\Delta x) \tag{4.7}$$

where Δx is the size of a discrete step, $n \in \mathbb{N}$ the step number. Additionally, the shift parameter is discretized by $\tau = l\Delta x$, where $l \in \mathbb{Z}$. Beyerer et al. [3] propose that the scale parameter s follows a logarithmic discretization of $s = 2^p$, $p \in \mathbb{Z}$. Afterwards, the logarithmic scale parameters is linked to the shift parameter:

$$l \coloneqq s\nu = 2^p \nu \quad \Leftrightarrow \quad \tau = 2^p \nu \Delta x \tag{4.8}$$

The result is a discretization that is fine for higher frequencies and coarse for lower frequencies, as seen in Fig. 4.2. "For a doubling of the spatial frequency (halving the scale) the number of sampling points is halved with respect to the coordinate of the spatial frequency" (Beyerer et al. [3, p. 732]).

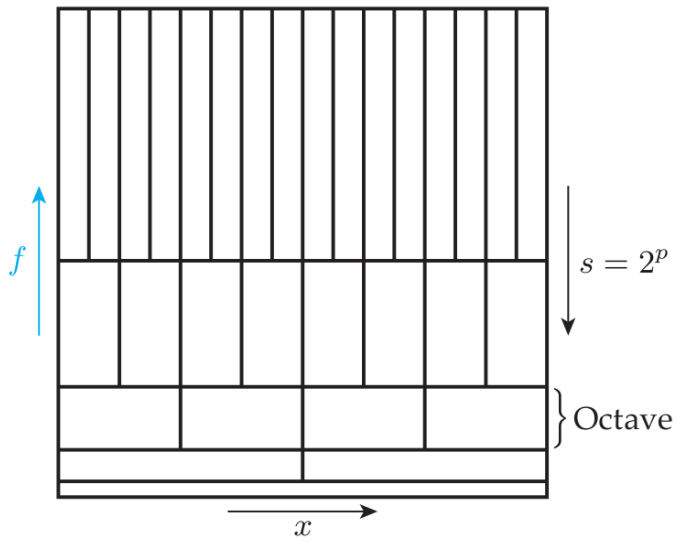


Figure 4.2: Finer location awareness with increasing frequency and vice versa. Adapted from Beyerer et al. [3, p. 732] with permission.

4.2 Wavelet Transformation for NIR spectroscopy

In Sec. 4.1 the fundamentals of continuous and discrete wavelet transformations were summarized. Based on the core idea of convolving a target function, i.e., wavelet, with a signal and receiving a scalar which quantifies the similarity between both functions, an adapted transformation method was developed. It is based on patent application *DE102017220103.7* (Krause [23]).
The wavelet transformation has the property to identify the existence of wavelet-shaped signal segments based on the wavelet shift and scale parameters. As a result, the absolute spatial location can be recovered and an estimation can be made as to which function is present. Relating back to Sec. 3.1, a NIR spectrum can be described as a superposition of Lorentz-shaped absorption peaks. Based on the location of a peak, a conclusion can be drawn which functional group and therefore which ingredient is responsible for the absorption (e.g., saturated fatty acids, when thinking about NIR spectroscopy for oil analysis). Additionally, knowing the scale of that absorption function and its absorbance intensity can yield estimations about the amount of that ingredient, solely based on the spectral information. To apply a wavelet transformation to NIR spectral data, some adaptations are proposed in the following subsections.

4.2.1 Windowed Signals

The original definition of CWT and DWT assumes that the wavelet is finite in a limited region, as in Fig. 4.1a for different scales. The wavelet is zero outside this region and factors any signal through the product inside the integral or sum. For NIR spectra, the mother wavelet is a Lorentz function, which converges towards zero asymptotically. As a result, signals far outside the center of the wavelet still contribute to the wavelet transformation result. Therefore, the window size described by discrete samples w is added as an additional parameter to the wavelet function:

$$\psi_{s,l,w}(n) \coloneqq \begin{cases} \psi\big((n-l)\Delta x, s\big) & , \ -\lfloor \frac{w}{2} \rfloor \leq (n-l) \leq \lfloor \frac{(w-1)}{2} \rfloor \\ 0 & , \text{otherwise} \end{cases} \tag{4.9}$$

with

$$w \in \mathbb{N}, \quad w \geq 2. \tag{4.10}$$

Eq. 4.9 is designed to handle even window sizes, e.g., $w = 10$, by distributing the discretized steps symmetrically around l, but adding the one remaining step to the left side of the symmetrical center (cf. Fig. 4.4b). The additional condition guarantees a window size greater than one, since that would simply compare single data points that do not have any information regarding the shape of the wavelet. As a visual example, Fig. 4.3 shows two discretized wavelets with $s = 3$ and $\Delta x = 1$ based on a Lorentz function as their mother wavelet. The discrete range is $n \in \{0, 1, \ldots, 40\}$, and the shift parameter is $l = 20$. The blue wavelet has its window parameter set to the maximum range, i.e., $w = 41$, while the orange wavelet is limited to $w = 11$. Therefore, only signal values with an discretization index that satisfies $20 - \lfloor \frac{w}{2} \rfloor \leq n \leq 20 + \lfloor \frac{w-1}{2} \rfloor$ are factored into the transformation result.

Instead of using a single window size w in the wavelet transformation, a range of window sizes are possible. In Section 4.3, elaborations on different reasons for analyzing a signal using different window sizes are provided. Regarding a formal

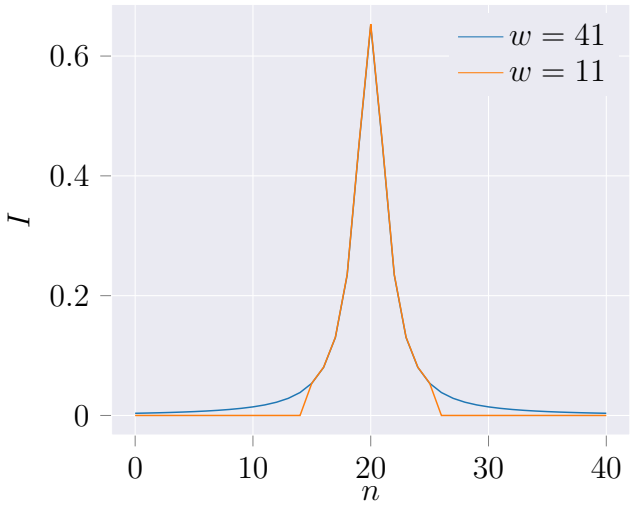


Figure 4.3: One wavelet with window size $w = 11$, the other one with maximum range $w = 41$.

definition, a new parameter is added to the wavelet function and represented by a fourth index. The previous window size parameter w becomes w_{min} and is complemented by the new parameter w_{max}. The design assumption is an increment of 1 discrete unit inside the range. Let $\mathcal{W}$ denote the set of window sizes w, that satisfy

$$w \in \mathbb{N},\ w_{min} \leq w \leq w_{max}\ . \tag{4.11}$$

Due to the range of window sizes, the wavelet transformation yields multiple results, i.e., exactly the cardinality $|\mathcal{W}| = w_{max} - w_{min} + 1$. Since it is only of interest to store the correlation between the wavelet and signal with regards to position and scale, the results from wavelet transformations using multiple window sizes need to be reduced to one. Admittedly, the choice of which transformation result to keep could be a complex algorithm that considers different parameters, especially the *uncertainty* as it is described in Sec. 4.3. In this thesis, the most simple approach is used: accepting the result from any w that yields the highest transformation result:

$$\begin{aligned}\Gamma_\psi(s,l) &= \sum_n g_n \psi_{s,l,w_{min},w_{max}} \\ &= \max\left(\left\{\sum_n g_n\psi_{s,l,w_{min}}, \sum_n g_n\psi_{s,l,w_{min}+1}, \ldots, \sum_n g_n\psi_{s,l,w_{max}}\right\}\right)\ .\end{aligned} \tag{4.12}$$

4.2.2 L^2-normalized Signals

The classical wavelet transformation does not necessarily normalized either the signal or the wavelet. The result is the scalar $\Gamma_\psi \in \mathbb{R}$. When thinking about two signals g_1 and g_2, where $g_2 := ag_n$, $a > 1$ which are transformed using the same wavelet, the results will differ. Even though g_2 is simply an amplified instance of g_1, the transformation result will be a greater scalar than for g_1. It is difficult to make a decision which signal was a better fit for the wavelet. Therefore, this adapted wavelet transformation for NIR spectroscopy proposes a normalization step using

the L^2 norm, also known as Euclidean norm.
Golub and Loan [12, p. 52] define a vector norm as follows: "a vector norm on $\mathbb{R}^n$ is a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ that satisfies the following properties:"

$$\begin{aligned} f(\boldsymbol{x}) &\geq 0 & \boldsymbol{x} &\in \mathbb{R}^n \\ f(\boldsymbol{x}+\boldsymbol{y}) &\leq f(\boldsymbol{x}) + f(\boldsymbol{y}) & \boldsymbol{x},\boldsymbol{y} &\in \mathbb{R}^n \\ f(\alpha\boldsymbol{x}) &= |\alpha| f(\boldsymbol{x}) & \alpha \in \mathbb{R},\ \boldsymbol{x} &\in \mathbb{R}^n\ . \end{aligned} \tag{4.13}$$

A norm is generally defined by the usage of double bar notation, e.g., $f(x) = \|x\|$, while a subscript at the end provides information which specific norm is used. Golub and Loan [12, eq. 2.2.1] attribute the generic *p-norms* as "a useful class of vector norms":

$$\|\boldsymbol{x}\|_p = (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}}, \quad p \geq 1\ . \tag{4.14}$$

The L^2 norm is a specific instance of the p-norms, where $p = 2$. Insert $p = 2$ into the generic equations shows immediately why one of the alternative names is the *Euclidean norm*. It also has the special property, that

$$\|\boldsymbol{x}\|_2 = (|x_1|^2 + \ldots + |x_n|^2)^{\frac{1}{2}} = (\boldsymbol{x}^T\boldsymbol{x})^{\frac{1}{2}}\ . \tag{4.15}$$

Gabor [11, p. 53] also state that "a unit vector with respect to the norm $\|\cdot\|$ is a vector $\boldsymbol{x}$ that satisfies $\|\boldsymbol{x}\| = 1$". Using the *Cauchy-Schwarz inequality*, as defined in Fischer [9, p. 275] as

$$|\langle \boldsymbol{x}\,, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\| \tag{4.16}$$

with

$$|\langle \boldsymbol{x}\,, \boldsymbol{y}\rangle| = \|\boldsymbol{x}\|\|\boldsymbol{y}\|, \quad \text{if } \boldsymbol{x} \parallel \boldsymbol{y}\ , \tag{4.17}$$

it can be concluded that the scalar product of two L^2 normalized vectors is always $-1 \leq \langle\cdot,\cdot\rangle \leq 1$. Applying this normalization to the scalar product in the wavelet transformation, the result is a correlation (or *similarity*) measure between signal and wavelet. Since the absorbance, transmittance or reflectance intensity is always $I \geq 0$, the wavelet transformation will be $0 \leq \Gamma_\psi \leq 1$. When normalizing the signal, the complete spectrum will be used as a normalization factor and therefore negatively influence the cross-correlation with $\psi_{s,l,w}$. As a consequence, the signal itself needs to be windowed, too. Only then can a perfect match between wavelet and underlying signal window be represented by $\Gamma_\psi(s,l) = 1$. Instead of defining the discretized signal as in Eq. 4.7, it is re-defined to reflect a normalization within the location and window size provided by $\psi_{s,l,w}$:

$$g_{l,w}(n) \coloneqq \frac{g(n\Delta x)}{\|g(\tilde{n}\Delta x)\|_2}, \quad l - \lfloor\frac{w}{2}\rfloor \leq \tilde{n} \leq l + \lfloor\frac{w-1}{2}\rfloor,\ \tilde{n} \in \mathbb{Z}\ . \tag{4.18}$$

Similarly, the wavelet is re-defined as L^2 normalized:

$$\psi_{s,l,w} \coloneqq \frac{\psi_{s,l,w}(n)}{\|\psi_{s,l,w}\|}\ . \tag{4.19}$$

Although the whole signal g is normalized by the L^2 norm of a smaller interval, there is no resulting negative influence on the wavelet transformation. Since the wavelet was already set to 0 for all values outside the specified window in Eq. 4.9, the influence of the signal outside the window will also be 0.
After implementing the L^2 normalization for both the signal and the wavelet, the wavelet transformation represents the correlation between these and can be interpreted as a measure of confidence that a signal with the shape of wavelet ψ with scale s is present at position l.

4.2.3 Removal of Scalings

In Eq. 4.2 the mother wavelet's intensity is scaled by factor of $\frac{1}{\sqrt{s}}$ and its position by $\frac{1}{s}$. This scaling is not necessary for L^2 normalized Lorentz wavelets. The removal was already implicitly done in Eq. 4.9 and its L^2 normalized adaptation in Eq. 4.19. The DWT, described in Subsec. 4.1.2, uses a shift parameter that is a function of the scale of the signal (cf. Eq. 4.8). This function is also removed for the proposed wavelet transformation. Independent of the scale of the Lorentz wavelet, the occurrence should be detected for every shift l. Instead, the implemented windowing parameter can limit broad wavelets to a narrow spectrum.

After describing the proposed changes for a wavelet transformation algorithm that is adapted to fit the characteristics of NIR spectra, the resulting equation is

$$\Gamma_\psi(s,l) = \sum_n g_{l,w}(n)\psi_{s,l,w}(n) \ , \tag{4.20}$$

which can be simplified to the scalar product notation

$$\Gamma_\psi(s,l) = \langle \boldsymbol{g}_{l,w} \, , \boldsymbol{\psi}_{s,l,w} \rangle \ . \tag{4.21}$$

Parameter w is not passed as a parameter of function Γ, because in Ch. 4.3 it will be proposed that $w = f(s)$.

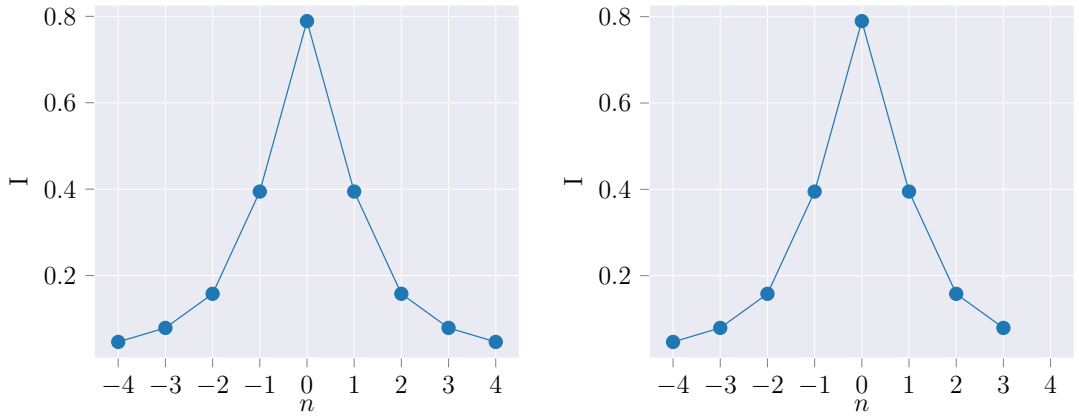
4.3 Evaluation of the Wavelet Window Size

In the previous section, an wavelet transformation was introduced, that was adapted to the needs of detecting peaks inside NIR spectra. The result was a scalar product of windowed and normalized signal and wavelet. For comparing two different spectra using wavelet analysis, the position and scale of the signal (in the domain of NIR absorptions approximately a Lorentz function (cf. Sec. 3.1)) must be known as exactly as possible. This requirement is independent of any subsequent peak detection algorithm which might discard some peaks later on as insignificant.
A recorded spectrum is represented as discrete data points in a sequential order, namely ascending by wavelength. This spectrum can have multiple absorption peaks which need to be detected. Therefore, a sliding window with a target wavelet is an appropriate way to match signal windows with the target function, as described in Eq. 4.9. Yet, the question of the appropriate window size remains unanswered. One can imagine that trying to detect a Lorentz function with $s = 4\,\text{nm}$ using the full width of a spectrum ranging from 900 to 1800nm with a resolution of 1 nm as a window will not yield any reliable results. But simply choosing any small window will also not produce satisfying results, e.g., the most extreme example of selecting a window of two discrete units (in the previous example that would be a window of length 2 nm, covering 3 data points). Gabor [11] was one of the first works to acknowledge that "the frequency of a signal which is not of infinite duration can be defined only with a certain inaccuracy, which is inversely proportional to the duration, and vice versa"(Gabor [11][p. 429]). They applied *Heisenberg's uncertainty principle* to signal processing, more specifically to the application of Fourier transformations. While the content of their work diverges from this work's application, the core idea remains an interesting factor: when using a sliding window to detect

Lorentz shaped absorptions in a spectrum, the windowed signal is of finite duration and any detection result becomes *uncertain*. This observation has two opposite directions:

- applying larger windows creates more certainty regarding the **scale** of the Lorentz function
- applying smaller windows creates more certainty regarding the **position** of the Lorentz function

Due to their diverging character, the question remains what an optimal window size should look like to satisfy both parameters. As stated in this chapters first paragraph, both the position and scale are important information about absorption peaks.
To find an answer to that question, several experiments were conducted using pure Lorentz signals and different window sizes. When referring to *odd* or *even* window sizes, the terms relate to the number of data samples captured by that window. Odd windows are symmetric around the center, while even windows are defined as having one data point more to the left of the center. Fig. 4.4b visualizes how an even number of points is handled by an imbalance on the left side. This is a clear difference to other works, e.g., Harris [15, Fig. 3], that use *even components* to describe that there is a balance of data points to either side of the center.



(a) Window with an odd number of $n = 4 + 1 + 4 = 9$ data points

(b) Window with an even number of $n = 4 + 1 + 3 = 8$ data points

Figure 4.4: Comparison between odd and even number of n data points

4.3.1 Optimal Window Size

For several experiments, the "fit" between two functions is used as a performance measure. The fit is defined as the *correlation coefficient* r between both functions. For two discrete vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ the coefficient is defined as

$$r_{\boldsymbol{x},\boldsymbol{y}} = \frac{\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{n\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}} \tag{4.22}$$

with

$$\sigma_{\boldsymbol{x}}, \sigma_{\boldsymbol{y}} \coloneqq \text{standard deviation of } \boldsymbol{x}, \boldsymbol{y}, \quad -1 \leq r \leq 1 \tag{4.23}$$

(cf. Han et al. [14, eq 3.3]). If $r = 1$, then both vectors are positively correlated, if $r = -1$ they are negatively correlated, and if $r = 0$ then both vectors are orthogonal.

4.3.1.1 Scale

The experiment was designed to find if and how the scale uncertainty decreases with increasing window size. First, a Lorentz function with a fixed scale s was defined. Afterwards, a window around the center of the function for every window size $w \in \{1, 2, \ldots, 200\}$ was multiplied with every Lorentz function with $s \in \{1, 2, \ldots, 60\}$. The resulting cross-correlation was recorded in a data frame for subsequent analyses. For the purpose of a more concise notation, the original definition of cross-correlation in Eq. 4.22 was changed in this experiment:

$$r_{w,s_1,s_2} = \sum_{n=0}^{w} L_{w,s_1}(n) L_{w,s_2}(n) \tag{4.24}$$

with

$$\begin{aligned} L_{w,s}(n) &\coloneqq \frac{2s}{\pi(s^2 + 4((n-l)\Delta x^2)} \\ l &= \lfloor \frac{w+1}{2} \rfloor \ , \\ \Delta x &= 1 \ . \end{aligned} \tag{4.25}$$

Fig. 4.5 shows the results for a Lorentz signal with $s = 10$. Different window sizes

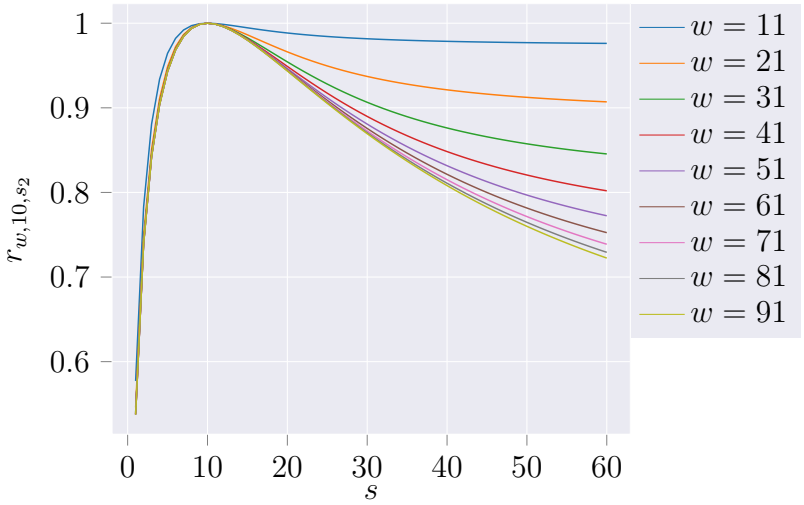


Figure 4.5: Cross-correlation r_{w,s_1,s_2} for $s_1 = 10$, $s_2 \in \{1, 2, \ldots, 60\}$, $w \in \{11, 21, 31, \ldots, 91\}$

are shown as individual lines. There are several observations that can be made:

1. The highest cross-correlation is found exactly at the target scale $s = 10$, which is a requirement for a successful analysis.

2. Using a smaller window size, neighboring values of $s = 10$ have also high cross-correlation. Using a larger window size, the difference between target s and neighboring values becomes more apparent and a *peak* emerges.

3. The curves appear to converge towards a limit with increasing window size, which could be used as an upper bound when selecting appropriate window sizes.

4. Functions with larger s have a higher cross-correlation than functions with a small s. Especially when using a small window size (cf. $w = 11$ in Fig. 4.5), the cross-correlation nearly plateaus for $s > 10$. This could lead to ill-detection of the peak scale parameter.

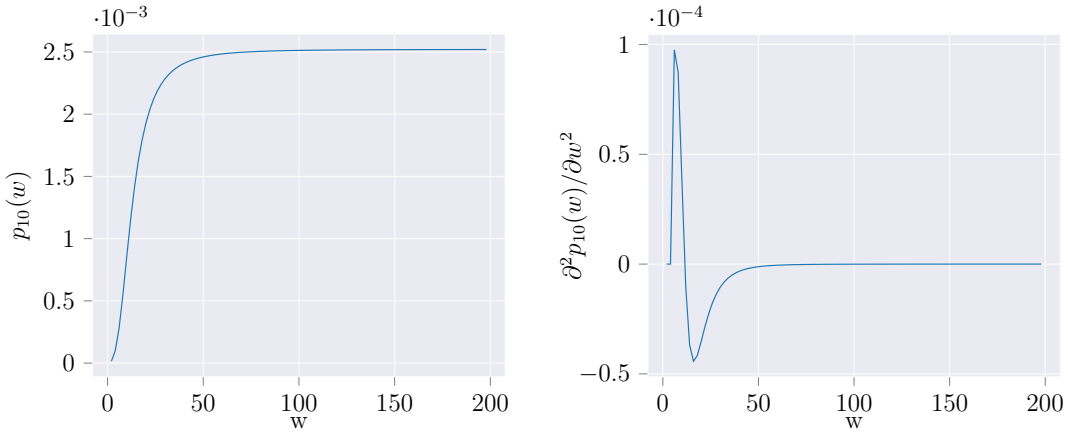
Following to these first observations, a performance measure needed to be defined, so relations between target s and different w could be quantified. Due to observation no. 1, one simple measure was defined as follows:

$$p_s(w) = \sum_{k=1}^{N} \left| r_{w,s,(s-k-1)} - r_{w,s,(s-k)} \right| + \left| r_{w,s,(s+k-1)} - r_{w,s,(s+k)} \right| , \quad N \in \mathbb{N} . \tag{4.26}$$

Depending on how much the "long-tail" of the curve should be factored in, a higher N can be selected. In the following analysis, the most simple version of Eq. 4.26 with $N = 1$ was selected, which could be reduced to

$$p_s(w) = \left| r_{w,s,s} - r_{w,s,(s-1)} \right| + \left| r_{w,s,s} - r_{w,s,(s+1)} \right| . \tag{4.27}$$

The general interpretation of $p_s(w)$ is that a larger value is based on a more significant peak at target s, and therefore has a decreased uncertainty when detecting the scale. Fig. 4.6 shows how the performance measure p increases with larger window



(a) Underived function (b) Second derivative

Figure 4.6: $p_{10}(w)$ over different window sizes w

size, but converges to a certain p_{limit}. The absolute values of p were different for any

tested s, but the shape of the curve always was similar to a *sigmoid* function. The typical "S"-shape was stretched with increasing s.
Relating back to the central question of this sub-section, the simple answer is that any larger window size yields less uncertainty about the scale of the function that is to be detected. Yet, using this naive approach is unbounded without resulting in significantly better performance after a certain point. Therefore, a cut-off value should be defined where the convergence towards p_{limit} is considered to be reached. During this experiment several methods for detecting convergence were tried, but the most robust method was to simply define the maximum window w_{max} where values of p grow larger than

$$p(w_{max}) = p_{max} - 0.1(p_{max} - p_{min}) \ . \tag{4.28}$$

$p(w_{max})$ is the value at 90 % of the range between the minimum and maximum value of p.
After defining a rule to select a maximum window size, a rule for a minimum window size was needed. Certainly, one window size could be considered enough, but only if the underlying signal is in shape of a perfect (infinite) Lorentz function. When thinking about real NIR spectra, peaks often overlap or are cut off by another absorption in the neighborhood. Therefore, a window range that is used in a wavelet analysis can provide some advantage.
Regarding the function in Fig. 4.6 as a sigmoid function, there must be an inflection point. Before that point there must be an exponential growth in the performance measure, after that an exponential decay until asymptotic behavior. Thus, selecting a window size *before* the inflection point will be significantly worse than any of its successors. Any window size after the inflection point will still improve, but at a decaying rate until the previously defined w_{max} is reached. Finding the inflection point is simple differential calculus. Fig. 4.6 shows the second derivative of the performance function with its minimum representing the inflection point in the original function. Following, the minimum window w_{min} can be defined as

$$w_{min} = \arg\min \frac{\partial^2 p}{\partial w^2} \ . \tag{4.29}$$

After having defined functions for both w_{min} and w_{max}, these were applied to all experiment results, i.e., Lorentz functions with $s \in \{1, 2, \ldots, 60\}$. The aggregated data showed that there was a difference if an odd or even window was used for signal detection. Fig. 4.7 shows the minimum and maximum window sizes separated by odd and even window size. While the odd and even window data points lie on a function with the same slope, combining them creates a noisy function. Therefore, they are displayed as separated but *entangled* functions. To smooth both window functions, a simply algorithm was applied:

$$\begin{aligned} w_{min}(s) &= \min(w_{min,odd}(s), w_{min,even}(s)) \\ w_{max}(s) &= \max(w_{max,odd}(s), w_{max,even}(s)) \end{aligned} \tag{4.30}$$

The result of taking the minimum window or maximum window for w_{min} or w_{max}, respectively, is shown in Fig. 4.8. The functions were labeled as *smoothed*, due to their less noisy shape. By visually inspecting both functions, the minimum window function appears to be linear, while the maximum window function shows some

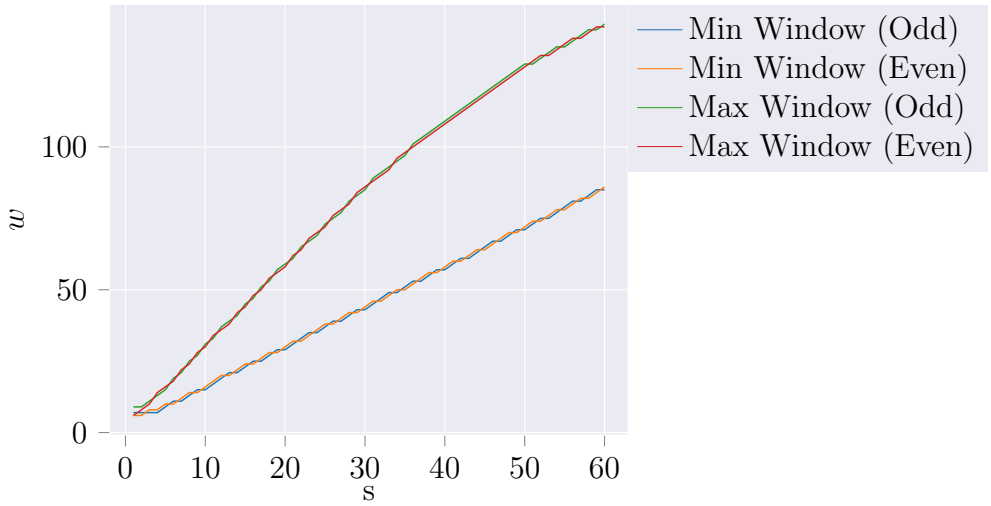


Figure 4.7: $w_{min}(s)$, $w_{max}(s)$ for $s \in \{1, 2, \ldots, 60\}$, separated by odd and even values for w

concave trend. Considering the methods which were used to define both values in Eq. 4.28 and 4.29, these differences can be explained. When defining w_{min}, only the position of the inflection point was required. As long as the inflection point is within the sampled window sizes ($w \in \{2, 3, \ldots, 200\}$), the value stays the same for any window size. In contrast, w_{max} was defined by using p_{max}, which is a value that behaves asymptotically, yet, if window size w is not sufficiently large for a scale parameter s, the value will be further away from its asymptotical limit. To keep the computation time for the experiment within a reasonable range, the tested window size was not increased to a much larger value, but rather values of $s \in 31, 32, \ldots, 60$ were disregarded for the following function fit.
Both smoothed curves could be used to fit a linear function into each of them, to create a generic rule to set windowing frames for any scale parameter s. Fig. 4.9 shows the fitted functions and the actual smoothed values as a baseline in the background. It is a visual confirmation that selecting values of $s \in \{1, 2, \ldots, 30\}$ for fitting the maximum window function was a right decision. For $1 \leq s \leq 30$, there is clearly linear behavior, while it soon decays afterwards. For fitting the data points to a linear function, a non-linear least squares methods implementation by Jones et al. [19, scipy.optimize.curve_fit] was used. The resulting functions were

$$\begin{aligned} w_{min}(s) &= 1.358s + 2.382 \\ w_{max}(s) &= 2.785s + 3.264 \end{aligned} \tag{4.31}$$

and could be used for implementation in a wavelet analysis algorithm for selecting window sizes backed by a analytical approach. It needs to be mentioned again, that the previous analysis was only based on the reduction of uncertainty in scale detection. The next sub-section will try to find a recommended window size (range) which allows a reduction in uncertainty regarding position detection.

4.3.1.2 Position

To have comparable results between *Scale* and *Position* uncertainty reduction methods, again, Lorentz functions with $s \in \{1, 2, \ldots, 60\}$ were tested with windows sizes

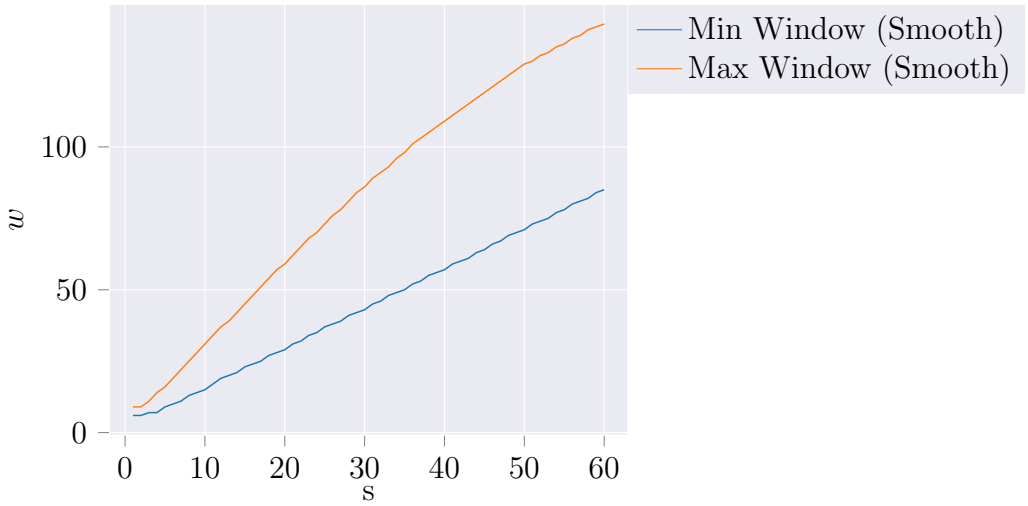


Figure 4.8: w_{min}, w_{max} for $s \in \{1, 2, \ldots, 60\}$

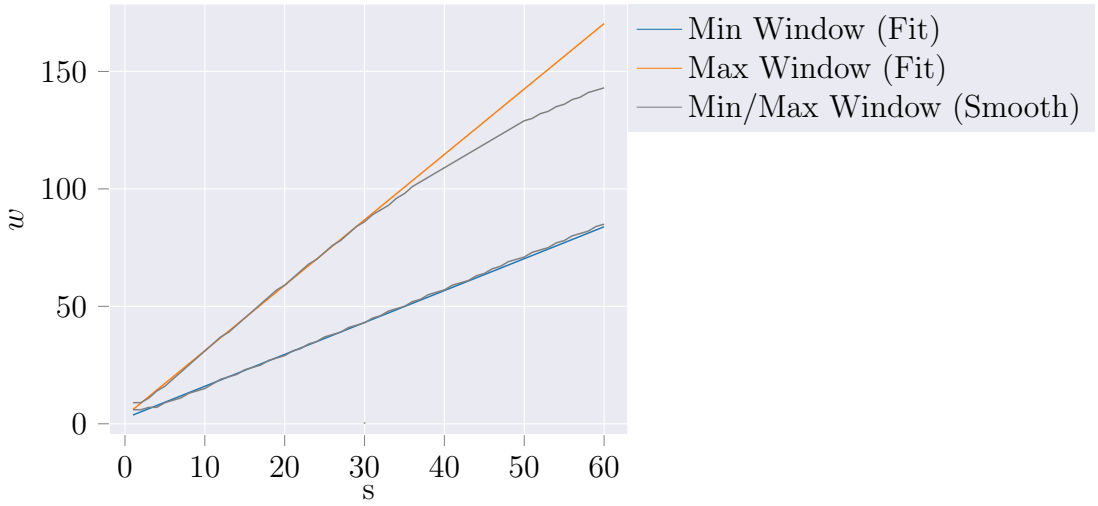


Figure 4.9: Fitted w_{min}, w_{max} for $s \in \{1, 2, \ldots, 60\}$ and their respective data baseline

of $w \in \{2, 3, \ldots, 200\}$. This time, the cross-correlation of two Lorentz signal with same scale and window size was calculated with one signal being shifted by another $l_0 \in \{-1, 0, 1\}$ units. The assumption behind this test was that the cross-correlation would be the highest (i.e., $r = 1$), when both signal had the same center, and less when centers where different. Due to the symmetry of the Lorentz signal, the cross-correlation of the signal will be also symmetric when shifted by either -1 or $+1$ units. Yet, to have a certain familiarity between this and the scale experiment, both shifts were tested and used for performance measure.
First, the definition of cross-correlation r needed to be redefined to fit a concise notation, similar to how it was done for the scale experiment in Eq. 4.24:

$$r_{w,s,l_0} = \sum_{n=1}^{w} L_{w,s,0}(n) L_{w,s,l_0}(n) \tag{4.32}$$

with

$$\begin{aligned} L_{w,s,l_0}(n) &:= \frac{2s}{\pi(s^2 + 4((n - (l + l_0)\Delta x^2)} \\ l &= \lfloor \frac{w+1}{2} \rfloor \ , \\ \Delta x &= 1 \ . \end{aligned} \tag{4.33}$$

The performance measure from Eq. 4.27 was adapted to

$$p_s(w) = |r_{w,s,0} - r_{w,s,-1}| + |r_{w,s,0} - r_{w,s,1}| \ . \tag{4.34}$$

Fig. 4.10 shows the performance score for a Lorentz function with $s = 10$. In con-

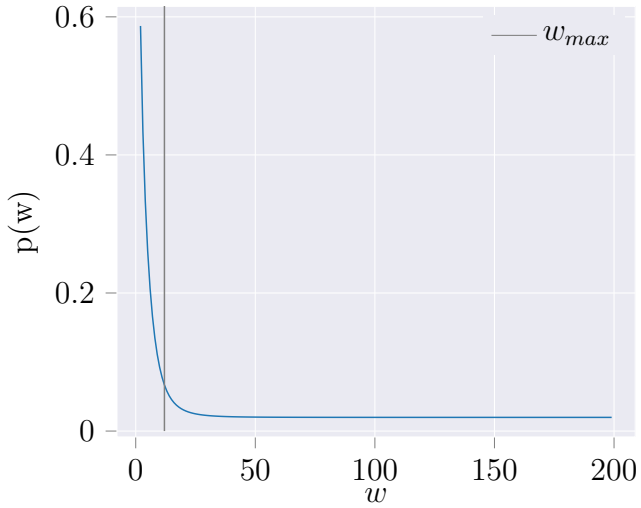


Figure 4.10: $p_{10}(w)$ for $w \in \{2, 3, \ldots, 200\}$ with $w_{max} = 12$

trast to the performance curve from Subsec. 4.3.1.1, this function has no inflection point and just an exponentially decaying performance with increasing window size. Recalling the *Uncertainty Principle* analogon from beginning of this chapter, these results fit the expectations. The larger the window, the higher the uncertainty of the exact position of the peak. Therefore, the best result is achieved when using the minimum possible window (i.e., $w = 2$, because a normalized cross-correlation using $w = 1$ would always result in $r = 1$), and constraining the maximum window at a value which is close to reaching the asymptote. Having "solved" the minimum window definition, a method is needed to define w_{max} as a function of s, so it can be determined for any s (as it was done in Subsec. 4.3.1.1).
The previously used method of allowing 90 % of the performance measure range as a valid window region has proven to yield linear results. Applying this method to this problem, we can calculate w_{max} for every tested Lorentz function. Fig. 4.10 shows an exemplary "cut-off" for a Lorentz function with $s = 10$. After aggregating the results, again, the determined maximum window value was separated by odd and even window sizes to prevent one noise functions. In contrast to the *Scale* experiment, this time both curves are not entangled and have differences of up to 5 width units (cf. Fig. 4.11). Before proceeding with the actual analysis, it needs to be noted that odd windows appear to perform better than their even counterparts, especially

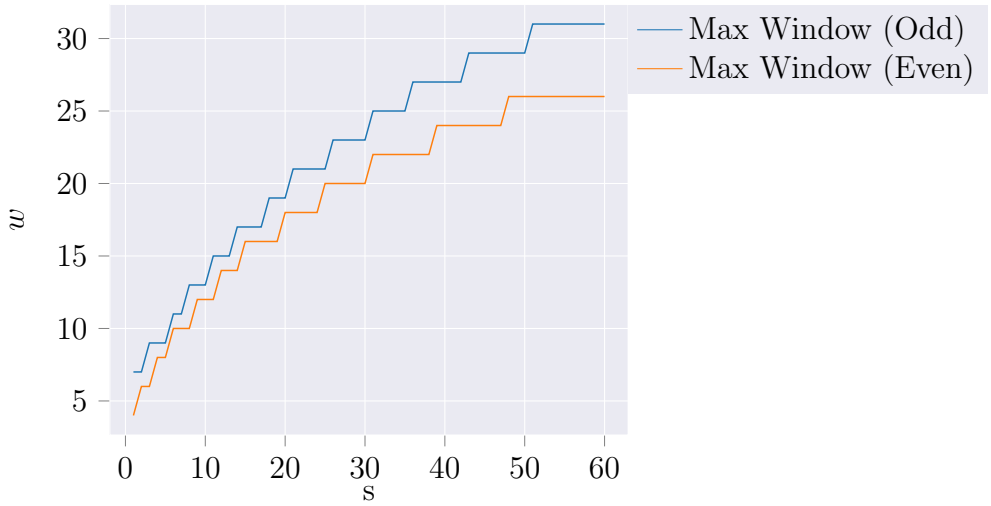


Figure 4.11: Window limits for odd and even windows

when the scale of the function grows larger. Applying the algorithm from Eq. 4.30, simply leaves the values from the odd windows. Unfortunately, in this data experiment no linear function can be fitted, but rather a non-linear function is needed. Using the same least-squares method from Jones et al. [19, scipy.optimize.curve_fit]. After fitting a second-order function $f_1(x) = ax^2 + bx + c$ and a mirrored exponential function $f_2(x) = a(1 - e^{-bx}) + c$ to the data, both performed approximately the same ($RMSE_1 = 0.588$, $RMSE_2 = 0.566$), with the exponential function being slightly better. The resulting functions were

$$\begin{aligned} w_{min}(s) &= 2 \\ w_{max}(s) &= 31.532(1 - e^{-0.029x}) + 5.715 \ . \end{aligned} \tag{4.35}$$

Fig. 4.12 shows the fitted function in comparison to the noisy data from using odd windows, and the fixed minimum window size of constant $w = 2$.

Same as before, these results only provide a window range which reduces uncertainty in determining the exact position of the Lorentz signal.

4.3.1.3 Combined Scale and Position

After analyzing how uncertainty can be reduced for both *Scale* and *Position*, one might have realized that both window ranges have barely any intersections that would make a combination of both results straight forward. First, any possible combination would results from an area between w_{max} of the position experiment and w_{min} of the scale experiment. Due to the fact that the found w_{max} has an exponential decay, there can be only intersections for smaller values of s, which is an undesired property. Second, w_{max} is very constraining and is almost always smaller than the smallest value of w_{min}. When plotting both functions together, one can see that there are barely any intersections, except for values $s \in \{1, 2, \ldots, 6\}$ (cf. Fig. 4.13). One explanation might be that the limiting values were selected too narrow. Yet, imagining allowing windows in the scale experiment far before the inflection point, will only create a different offset for the minimum window function, but not resolve the inevitable divergence from the decaying maximum window exponential

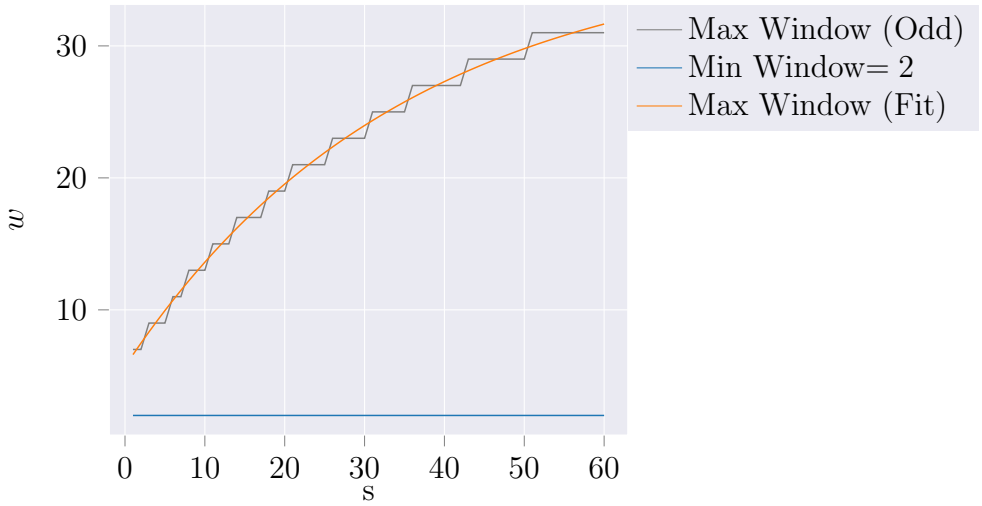


Figure 4.12: Window limits for odd and even windows

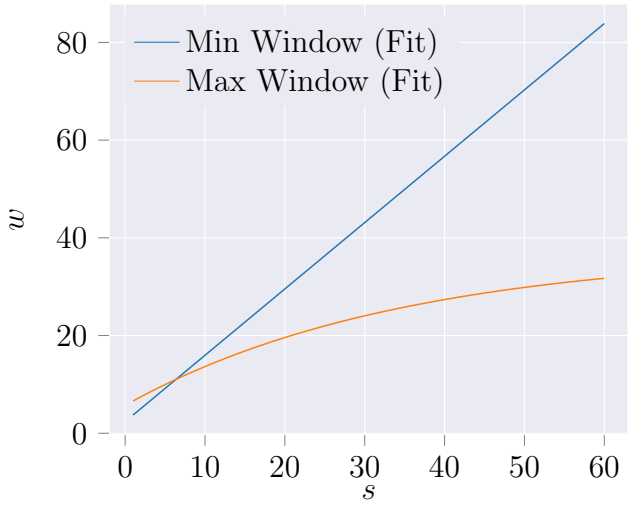


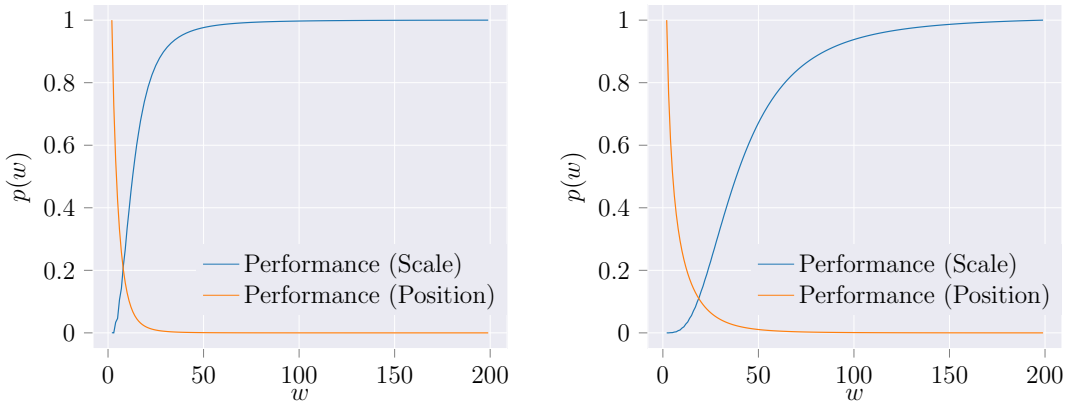
Figure 4.13: Combined window limits from scale and position experiments

function. Another possibility would be to define another (or broader, see Eq. 4.26) performance measure, re-run the experiments and see if other insights can be gained on how to define uncertainty ranges.
Here, a third solution is proposed, which is built on top of the results from the previous experiments and create a range of window sizes as a function of scale parameter s:

Factored Scale and Position Score

Until now, both scale and position uncertainty have been treated as distinct performances. Through the *Uncertainty Principle*, it needs to be accepted, though, that both values are intertwined and can't be treated isolated. Therefore, both performance measures from Subsec. 4.3.1.1 and 4.3.1.2 were plotted in one graph for an initial visual inspection. Unfortunately, using absolute values did not create a useful view for the problem. The scale performance ranged from 0.02 to 0.59, while the

position performance ranged only from 0.000 014 to 0.0025. A solution to this obstacle was to normalize the values from both performances into a comparable range. Using a normalization between 0 and 1 appeared to be the best option: this made all performance measures more comparable between each other. The range can be interpreted as a degree of reaching the best possible score (or the worst possible score, vice versa). Due to the asymptotic behavior of both functions, it can be presumed that the recorded best and worst values from our data set are limiting values. Additionally, having a range from 0 to 1 gives both performances equal weight (if not defined otherwise in an aggregation function). A value of 0.5 is half of the best-possible result. Fig. 4.14 shows both performance functions after normalization for a Lorentz signal with $s = 10$ and $s = 30$. It shows that the scale performance score decays more slowly to its asymptote, while the position performance already reaches its lowest value at a smaller window size. The obvious choice to select a "fair" win-

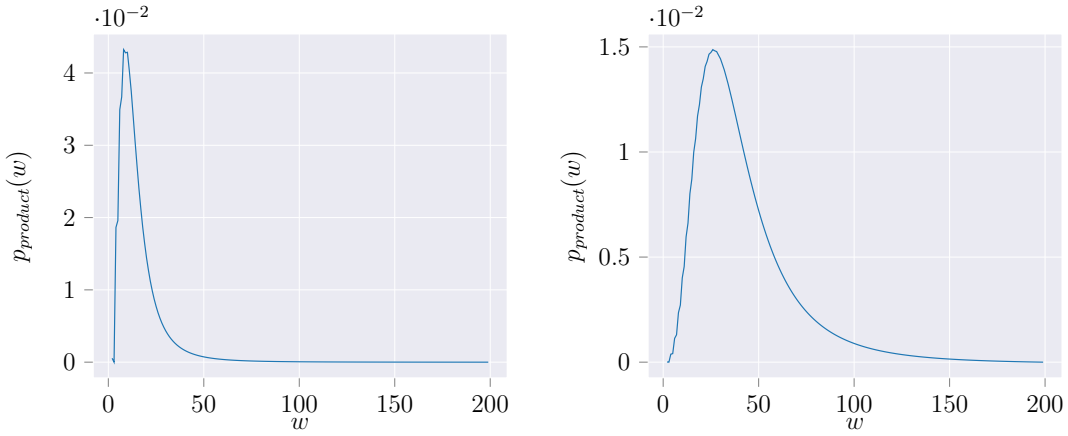


(a) Lorentz function with $s = 10$

(b) Lorentz function with $s = 30$

Figure 4.14: Normalized performance score from scale and position experiment

dow size to satisfy both criteria would be the intersection point, where both have quite low values, but are in balance (in Fig. 4.14 at $p(w = 7.92) = 0.21$)). But due to the non-linear slope for both functions between 0 and 1, other window sizes might also create acceptable performances. Due to the normalization, the product of both performance functions will always be between 0 and 1, too. This appeared to be the best option to combine both performances into one. Fig. 4.15 shows exemplary plots for Lorentz functions with $s = 10$ and $s = 30$ and their product performance. When observing the shape of both performance functions, it showed that when the underlying Lorentz function had a larger scale parameter, the maximum product performance was smaller and but decreased slower due to the broader shape. After having accepted the product performance to be a suitable measure for selecting window sizes, the re-occurring problem of defining a minimum and maximum window size as a function of s needed to be solved. Fig. 4.15 shows that the performance function's shape is similar to a Gauss or Lorentz function, with a certain symmetry around the maximum value. Therefore, the full width at half maximum ($FWHM$) was selected to define the minimum and maximum window size. After calculating those points for every tested Lorentz function with $s \in \{1, 2, \ldots, 60\}$, a picture of linear window function was created. As it can be seen in Fig. 4.16, with increasing



(a) Lorentz function with $s = 10$ (b) Lorentz function with $s = 30$

Figure 4.15: Product performance score from scale and position experiment

s, the maximum window sizes increases much faster than the minimum window size. Furthermore, the window size depending based on the intersection points, follows a similar function as the minimum window size, simply with a larger offset. Seeing the intersection point window size always within the $FWHM$ limits is an affirmation the window range includes optimal window sizes. As a last step, linear functions where

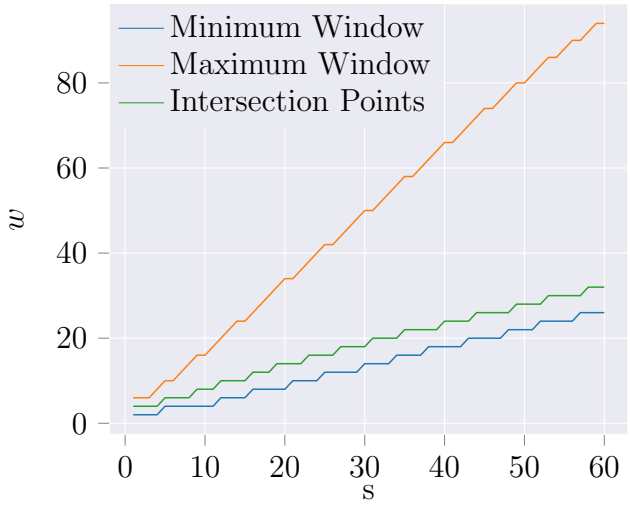


Figure 4.16: Window limits from product performance

fitted into the displayed data to receive general functions that can be implemented in any wavelet algorithm (cf. Fig. 4.17). The resulting functions were

$$\begin{aligned} w_{min}(s) &= 0.433s + 0.492 \\ w_{max}(s) &= 1.574s + 1.768 \end{aligned} \tag{4.36}$$

.

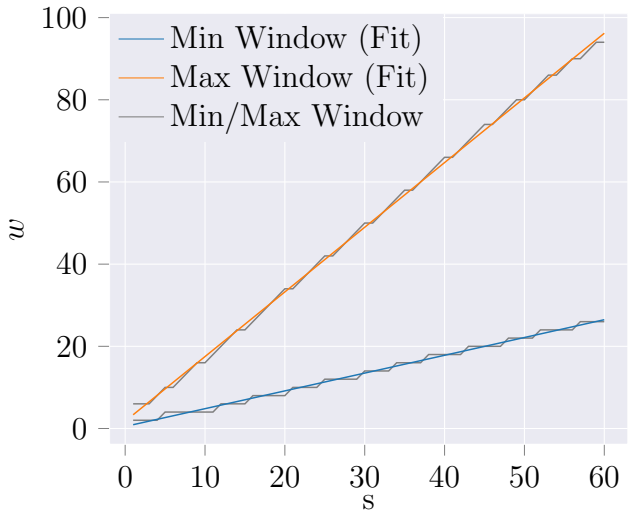


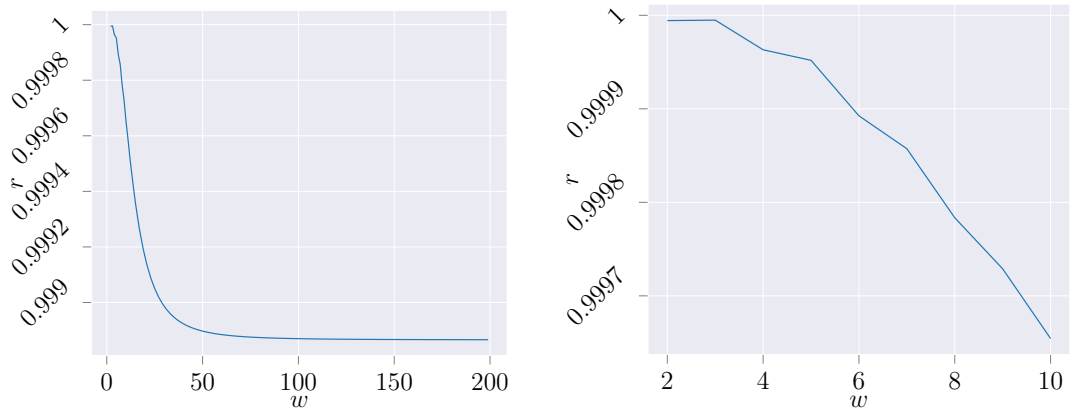
Figure 4.17: Fitted window limits from product performance

4.3.1.4 Result Discussion

After trying multiple approaches to find suitable window sizes that can be generalized into a function of Lorentz functions' scale parameter s, the most promising approach was described last. Normalizing the performance measure between 0 and 1 gives both the scale and position uncertainty equal weight. By creating the product of both at every tested window size, a joint performance could be created. Frankly, there might be better performance measures than used in Subsec. 4.3.1.1 and 4.3.1.2, and through the normalization they can be easily replace the one described in Eq. 4.27. Yet, due to the simplicity of the underlying operations, the window size functions created in this chapter were accepted and used for further wavelet analyses.

4.3.2 Comparison of Odd and Even Window Size

So far, any window size was increased by 1 unit until a maximum window was reached. As a result, the window size will be alternatingly even and odd. In this section it will be analyzed if the usage of an even or odd window has impact on the performance of the wavelet transformation.
The performance was tested by having one Lorentz function as the base signal with a fixed value of s, and one Lorentz function as the wavelet with a fixed value of $s + 1$. These two signals were correlated using the definition from Eq. 4.24 with an incremental window size $w \in \{2, 3, \ldots, 200\}$. For s, the usual range of $s \in \{1, 2, \ldots, 60\}$ was used. An example result for $s = 10$ can be seen in Fig. 4.18. The cross-correlation r decreases until reaching its asymptote value at $w \approx 100$. Yet, when looking closely, there appears to be some *ripple* at the very beginning of the curve. When enlarging the affected region (cf. Fig. 4.18b), the effect becomes clearly visible: the difference between r at an even window size and at a preceding odd window size is always greater than the difference between an odd window preceded by an even one. As a result, the slope of the function is not as smooth as it could be. To make the effect more visible across the full tested window range, for every window size the difference between it and its predecessor was calculated. Fig. 4.19 shows that the ripple effect continues for a larger range of window sizes, yet, admittedly



(a) Window range from 2 to 200 (b) Window range from 2 to 10

Figure 4.18: $r_{w,10,11}$ for $w \in \{2, 3, \ldots, 200\}$

less significant as for smaller windows.

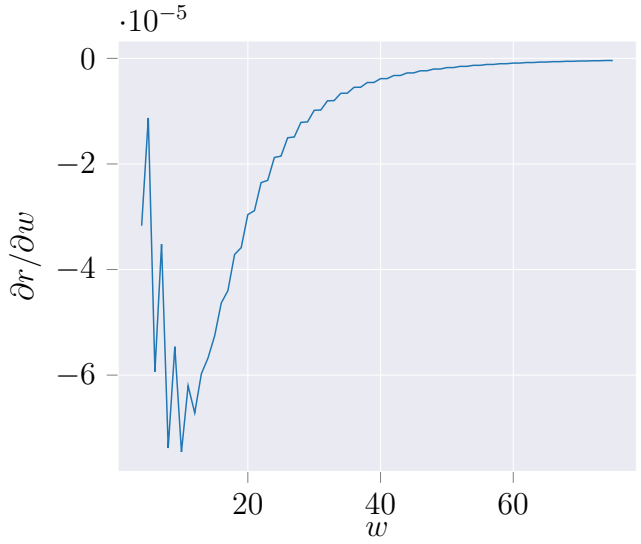


Figure 4.19: Derivative of $r_{w,10,11}$ for $w \in \{2, 3, \ldots, 75\}$

As the previous graphs showed, fortunately, the choice of odd or even window size does not affect the general trend of the cross-correlation, but it has an effect on the smoothness of that trend. The information gain for every increase in window size is not as large as it could be if only odd or even windows were used. When thinking about peak finding algorithms, a smooth curve of cross-correlation values is definitely preferable. Furthermore, if leaving out every second window size (either odd or even), the computation time is halved. Therefore, it is recommended to decide on either odd **or** even window sizes. In this work the decision was made to use only odd window sizes, due to the symmetry of Lorentz functions around the

center and the "balance" of the odd window size.
The results regarding the minimum and maximum window size in Sec. 4.3.1 are still valid. It is recommended to simply round the minimum and maximum window size to the next odd integer, and increment window sizes afterwards by 2.

4.4 Sampling Theorem for Lorentz-shaped Absorptions

Based on the NIR fundamentals described in Sec. 3.1, any recorded NIR spectrum can result in a combination of multiple bell-shaped curves, similar to Lorentz functions. If these functions are at adjacent wavelengths and have a sufficiently large scale, they can overlap and the signals add up at these points. Accumulated signals can lose properties of their original shape. Especially when very close together, the superposed signals will result in one peak, hiding its origin of two separate signals. Considering this effect, any feature detection algorithm could detect an incorrect number of Lorentz functions, an incorrect position or scale. To avoid this, two central questions need to be answered:

1. *Which resolution criterion is suitable for NIR absorption peaks?*
2. *What peak distance $\Delta\tau$ is required to theoretically detect two peaks according to the defined resolution criterion?*

The next section starts with the basic theory regarding resolution criteria and superposition of signals.

4.4.1 Theory

According to Beyerer et al. [3, p. 58], "the physically possible resolution of an optical system, as well as its sharpness, are limited by diffraction. When an image is formed, the diffraction patterns of all the object points are additively superposed on each other on the image plane". Due to the superposition of object points, two objects can be perceived as one, given that their centers are sufficiently close to each other. Fig. 4.20 shows how two objects are perceived through a small aperture on an image plane. The diffraction patterns are accumulated at overlapping segments.
Visualizing the effect from a two-dimensional perspective makes the superposition clearly visible. Fig. 4.21(a) shows that both signals (i.e., *Dirac delta functions*) can be clearly perceived as two separate signals. When moving their centers together, their intensities are accumulated (shown as red, dashed lines). While Fig. 4.21(b) still shows two separated peaks, Fig. 4.21(d) appears as one function with one peak. In fact, the former shows the Rayleigh criterion which states that "the center of the first Airy disk is located at the position of the second diffraction pattern's first minimum" (Beyerer et al. [3, p. 60]). 4.21(c) shows *Sparrow's resolution limit*, which is met when the center between the two peaks is an inflection point. The *Rayleigh criterion* is a well-accepted approach to define the resolution limit of an optical system ("of all the diffraction-related resolution criteria, the classical Rayleigh criterion is certainly the most famous" (den Dekker and van den Bos [6, p. 547])).

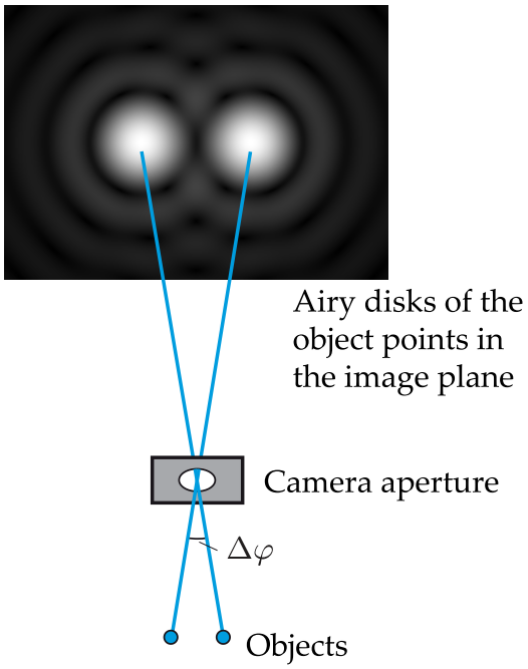


Figure 4.20: "Superposition, in the image plane, of the diffraction patterns of two object points". Adapted from Beyerer et al. [3, p. 59] with permission.

When empirically determining the Rayleigh criterion for different wavelengths λ and aperture sizes D, the observation angle $\Delta\varphi$ will follow the function

$$\Delta\varphi \approx 1.22\frac{D}{\lambda} \tag{4.37}$$

(cf. Beyerer et al. [3, eq. 2.185]).
Since Sparrow's resolution limit uses the inflection point, it allows narrower angles $\Delta\varphi$ than the more conservative Rayleigh criterion. The resolution limit can be calculated using

$$\Delta\varphi \approx 0.95\frac{D}{\lambda} \tag{4.38}$$

(cf. Beyerer et al. [3, eq. 2.187]).

4.4.2 Resolution Criterion for Lorentz Functions

The previously introduced criteria were designed to define limits when two diffraction patterns are not distinguishable anymore. The underlying signal is a Dirac delta function. Unfortunately, the shape of absorption spectra in NIR spectroscopy follow another function shape, i.e., a Lorentz function. Due to this fact, it is unclear whether the known criteria hold. Yet, it would be an excellent tool to discard detected peaks as *not possible*, if they didn't meet an adequate resolution criterion. Additionally, they would give an indication which sensor resolution is required to detect peaks of specific scale and position. This is important to precisely identify intensities of *known* absorptions, but also to detect any *unknown* absorptions, i.e., anomalies.
Therefore, a resolution criterion for Lorentz function is evaluated. A transfer of the

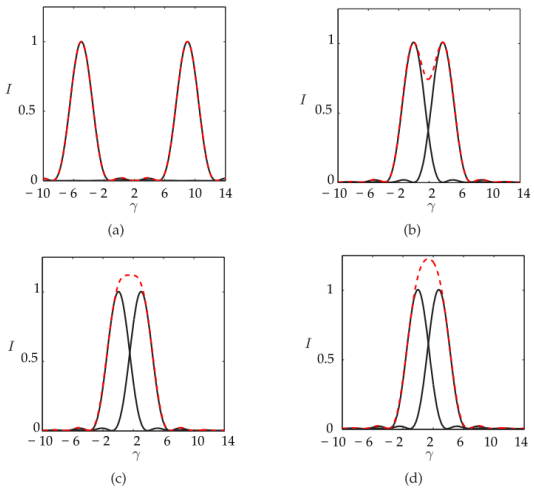
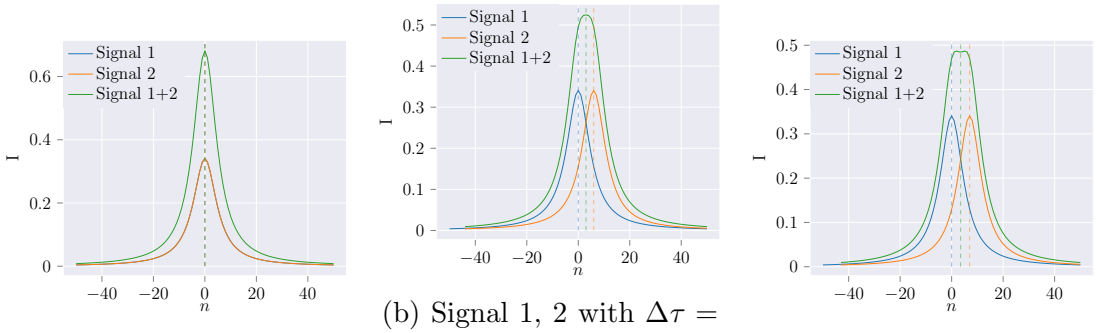


Figure 4.21: "Resolution of imaging systems for two superposed Airy diffraction patterns: (a) Clearly resolvable diffraction patterns; (b) Rayleigh criterion; (c) Sparrow's resolution limit; (d) Diffraction patterns which can no longer be resolved". Adapted from Beyerer et al. [3, p. 59] with permission.

Rayleigh criterion is not possible, since it requires both functions to have a (first) local minimum. The Lorentz function has one global maximum, but no minimum. It decreases asymptotically towards zero on either side of the maximum. Furthermore, the Rayleigh criterion was defined "based on presumed resolving capabilities of the human visual system" (den Dekker and van den Bos [6, p. 548]). In contrast to the human visual system, an algorithm can theoretically detect the smallest changes in superposition of two functions, given a sufficient resolution of the sensor.
Sparrow's resolution limit could be used with Lorentz functions, but didn't have the best possible definition of "resolvable". Having an inflection point in the center is a very specific criterion, which can be easily evaluated when using continuous functions. Finding an inflection point in shifting discrete functions is a problem, though. It might be possible using a very high resolution, but is not guaranteed. Additionally, such a resolution is rarely the case in NIR spectroscopy[1]. Yet, an inflection point in the center is only a small increase in centers shift away from becoming a local minimum. Therefore, the resolution criterion for Lorentz functions was defined to be just when there is still a local minimum between two functions' peaks. Fig. 4.22 visualizes the definition of the criterion: From (a) to (c), signal 2 is shifted outwards, and the superposition's peak becomes more round. After reaching $\Delta\tau = 7$, the first local minimum is visible and noted down as the criterion for two Lorentz functions of same intensity and scale $s = 11$.
One major difference in this criterion is the resulting value: instead of measuring an angular resolution of $\Delta\varphi$, the distance between both function's centers $\Delta\tau$ in discrete units is determined. By using a factor $\Delta\tau_{rel}$, the absolute value should be

[1] As an example, the *OceanOptics NIRQuest 512* used in Ch. 6 has a resolution of 1.6 nm with a range from 900 to 1700nm



(a) Signal 1, 2 with $\Delta\tau = 0$ (b) Signal 1, 2 with $\Delta\tau = 6$ (c) Signal 1, 2 with $\Delta\tau = 7$

Figure 4.22: Signal 1, Signal 2: Lorentz function with $s = 11$, $res = 1$, $l = 101$ Signal 2 center shifted along x-axis

computable for every scale s in combination with every wavelength resolution. To create a generic equation, the function's unit scale $s_n = \frac{s}{\Delta\lambda}$ is used:

$$\Delta\tau = \Delta\tau_{rel} s_n = \Delta\tau_{rel} \frac{s}{\Delta\lambda} \ . \tag{4.39}$$

This represents an adaptation of the *classic* resolution criterion to computationally resolvable Lorentz functions, i.e., comparing two functions of the same intensity. In addition, the criterion should also be defined for functions of different intensity. The relative intensity between two functions will be defined as *amplification* a. As an example, if two peaks have intensity I_1 and I_2, respectively, the amplification is $a = \frac{I_2}{I_1}$. Due to the symmetry of Lorentz functions, an amplification $a = \frac{I_2}{I_1} < 1.0$ can be treated as $a = \frac{I_1}{I_2} > 1.0$. Therefore, the resulting definition of amplification is

$$a = \frac{max(I_1, I_2)}{min(I_1, I_2)} \ . \tag{4.40}$$

4.4.3 Resolution for Continuous Functions

To determine the exact criterion, continuous functions are used in combination with differential calculus. The function of a Lorentzian is recalled from Eq. 4.41. Since the criterion should define the resolution of two superposed functions, the sum L_3 of two functions needs to be calculated. For design reasons one function (L_1) is held with its center at $x = 0$ by setting $\tau = 0$, and the second function (L_2) is shifted by τ. Additionally, the second function is scaled by the amplification a (cf. Eq. 4.42).

$$L(x, \tau) = \frac{2s}{\pi(s^2 + 4(\tau - x)^2)} \tag{4.41}$$

$$\begin{aligned} L_1(x) &= \frac{2s}{\pi(s^2+4x^2)} \\ L_2(x,\tau) &= \frac{2s}{\pi(s^2+4(\tau-x)^2)} \\ L_3(x,\tau,a) = L_1(x) + aL_2(x,\tau) &= \frac{2s(a(s^2+4x^2)+s^2+4(t-x)^2)}{\pi(s^2+4x^2)(s^2+4(t-x)^2)} \end{aligned} \tag{4.42}$$

Being able to calculate the exact superposition as a function of all variable parameters, it is to detect the first local minimum x_{min} between x and $x+\tau$ by applying basic differential calculus and solving for roots of the first derivative of L_3. To validate whether the found extremum is indeed a local minimum, the second derivate of L_3 is calculated at the found extremum position. τ is accepted as the minimum shift if $\frac{\partial^2 L_3(x_{min},\tau,a)}{\partial x^2} > 0$.

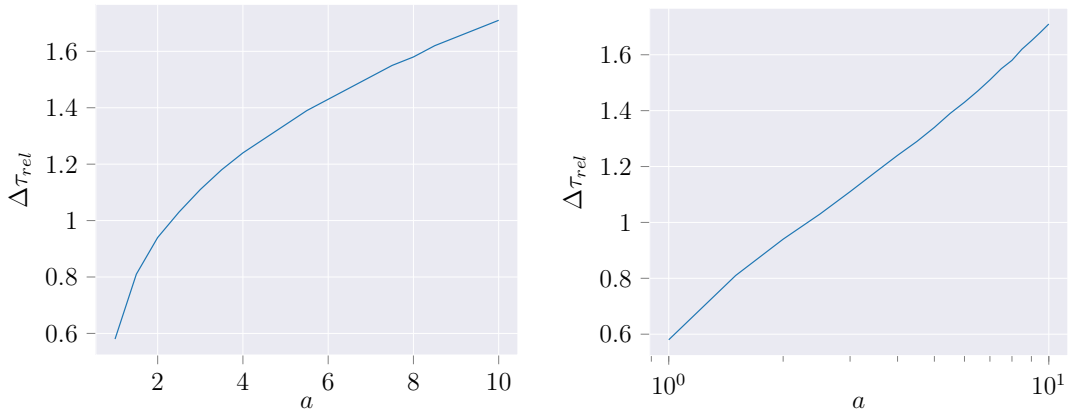
$$\frac{\partial L_3(x,\tau,a)}{\partial x} = \frac{16s(a(s^2+4x^2)^2(t-x) - x(s^2+4(t-x)^2)^2)}{\pi(s^2+4x^2)^2(s^2+4(t-x)^2)^2} \tag{4.43}$$

$$\begin{aligned} \frac{\partial^2 L_3(x,\tau,a)}{\partial x^2} =& \frac{-16as^3}{\pi(s^2+4(t-x)^2)^3} + \frac{192as(t-x)^2}{\pi(s^2+4(t-x)^2)^3} \\ & - \frac{16s^3}{\pi(s^2+4x^2)^3} + \frac{192sx^2}{\pi(s^2+4x^2)^3} \end{aligned} \tag{4.44}$$

For cases of no amplification, i.e., $a = 1$, the determination of the minimum $\Delta\tau$ required to meet the criterion could have been solved by additional partial derivation, since the position of the minimum would be exactly at $x_{min} = \frac{(x+\tau)-x}{2}$. For $a > 1.0$, the position is not clear, therefore an iterative approach was chosen: incrementally test values of $\Delta\tau > 0$ until a value of x_{min} meets the conditions described above. The increment was chosen as 0.01 units.
Fig. 4.23a shows the results of $\Delta\tau_{rel}$ for $a \in \{1.0, 1.5, 2.0, \ldots, 10.0\}$. With increasing a, $\Delta\tau_{rel}$ appeared to decay exponentially. Drawing a conclusion about generic behavior for all possible values of a from an exponential function is not optimal. After plotting the function on a logarithmic x-axis, the shape followed a linear function, as can be seen in Fig. 4.23b. The small deviations from perfect linear behavior is attributed to the imperfect increment in tested values for $\Delta\tau$. As usual, the experimental data points are used to fit a function of form $f(x) = b\log_{10}(x) + c$ which can be used for generic description of minimum shift as a function of a. The resulting function was

$$\Delta\tau_{rel}(a) = 1.098\log_{10}(a) + 0.590 \tag{4.45}$$

and can be compared to the experimental data function in Fig. 4.24. In contrast to the classical criteria described in Sec. 4.4.1, this criterion can be used to define the resolution for peaks with different intensities.
This function serves as the asymptotical limit for any following determination using



(a) Normal x-axis (b) log_{10} x-axis

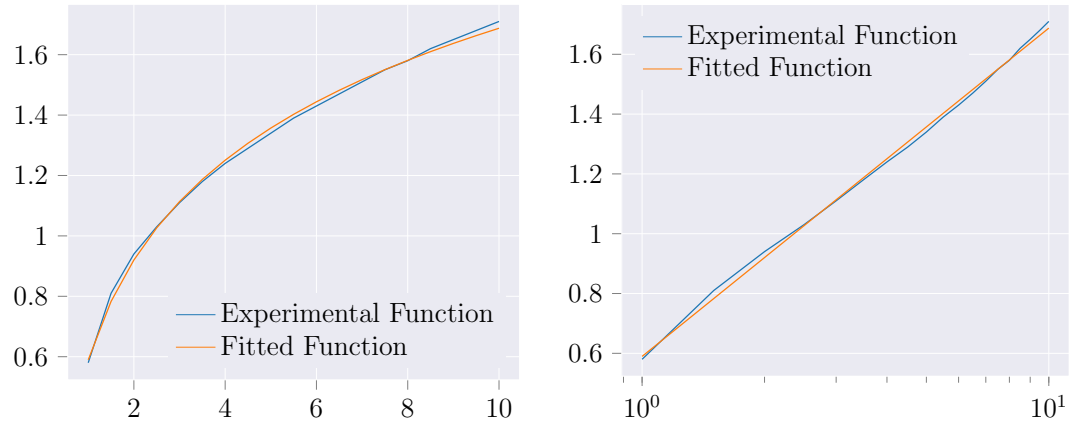
Figure 4.23: Resolution limits for continuous functions and $a \in \{1.0, 1.5, 2.0, \ldots, 10.0\}$

discrete functions.
Some exemplary values for $a \in \{1.0, 2.0, \ldots, 10.0\}$ are

$$\begin{aligned}
\Delta\tau_{rel_{1.0}} &= 0.58 \\
\Delta\tau_{rel_{2.0}} &= 0.92 \\
\Delta\tau_{rel_{3.0}} &= 1.11 \\
\Delta\tau_{rel_{4.0}} &= 1.25 \\
\Delta\tau_{rel_{5.0}} &= 1.36 \\
\Delta\tau_{rel_{6.0}} &= 1.44 \\
\Delta\tau_{rel_{7.0}} &= 1.52 \\
\Delta\tau_{rel_{8.0}} &= 1.58 \\
\Delta\tau_{rel_{9.0}} &= 1.64 \\
\Delta\tau_{rel_{10.0}} &= 1.69\ .
\end{aligned} \tag{4.46}$$

4.4.4 Discrete Functions

NIR spectroscopy returns a set of data points that describe the intensity of an absorption as a function of discrete wavelengths. Applying a resolution criterion that is based on continuous functions will not yield satisfactory results if (a) two peaks are very close together and/or (b) the resolution of the sensed wavelengths is not sufficiently large. Therefore, this subsection describes experiments that will try to find a function that returns the minimum resolvable distance between two peaks as a function of their amplitude relation a, scale s and resolution $\Delta\lambda$.
For the following experiment, the dimensionless unit scale $s_n = \frac{s}{\Delta\lambda}$ was used to produce universal results. Using discrete Lorentz functions with unit scale of $s_n \in \{1, 2, \ldots, 200\}$ were superposed, analyzed for a local minimum, shifted, superposed again and so on. When a local minimum was found, the current value of $\Delta\tau$ was noted and the next function pair tested. A local minimum at x_{min} is accepted when



(a) Normal x-axis (b) log_{10} x-axis

Figure 4.24: Resolution limits with fitted function

$I(x_{min} - 1) > I(x_{min})$ and $I(x_{min} + 1) > I(x_{min})$. Fig. 4.25 shows the steady slope of fitted functions with increasing s_n.

$$\begin{aligned}
\Delta\tau_{1.0} &= 0.58s_n + 1.09 \\
\Delta\tau_{2.0} &= 0.93s_n + 0.61 \\
\Delta\tau_{3.0} &= 1.10s_n + 0.56 \\
\Delta\tau_{4.0} &= 1.23s_n + 0.55 \\
\Delta\tau_{5.0} &= 1.34s_n + 0.57 \\
\Delta\tau_{6.0} &= 1.43s_n + 0.65 \\
\Delta\tau_{7.0} &= 1.51s_n + 0.59 \\
\Delta\tau_{8.0} &= 1.58s_n + 0.62 \\
\Delta\tau_{9.0} &= 1.64s_n + 0.61 \\
\Delta\tau_{10.0} &= 1.70s_n + 0.62
\end{aligned} \tag{4.47}$$

Eq. 4.47 indicate by the slope factors how close the experimental results using discrete functions are to the "ground truth" produced using continuous functions in Subsec. 4.4.3, specifically displayed in Eq. 4.46. The slope and the offset also reveal why a small unit scale fails: when applying $s_n = 1.0$ to any of those functions, the offset has a relatively high impact on the result of $\Delta\tau$. When using $s_n = 200$, the offset is negligibly small. Plotting the relative center distance $\Delta\tau_{rel} = \frac{\Delta\tau}{s_n}$ in Fig. 4.26 shows clearly how large s_n has to be to take advantage of the best possible resolution. Until s_n is about 25 - independent of a - $\Delta\tau_{rel}$ decreases fast, for $s_n > 25$ it converges towards its asymptotical limit from Eq. 4.45.

4.4.5 Result Discussion

After defining a new resolution criterion that is not designed with the human visual system in mind, both continuous functions and discrete functions were tested. While the continuous functions provided the general limit of resolution, the discrete

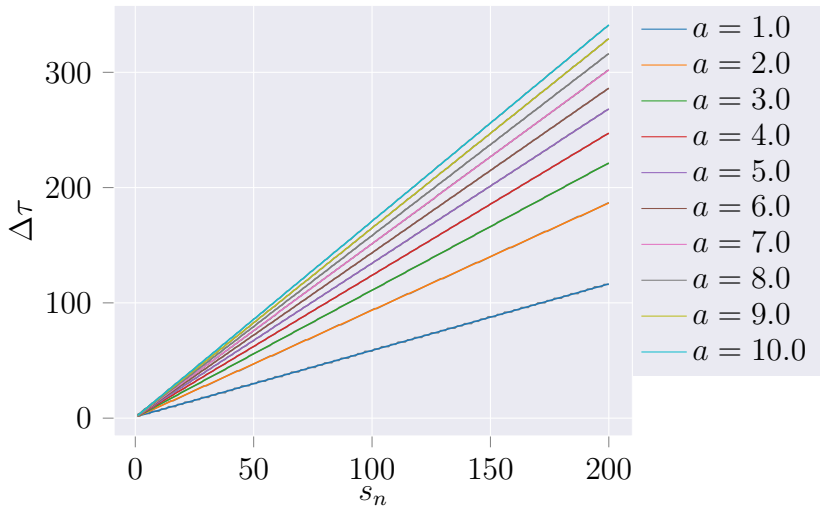


Figure 4.25: Resolution limits for functions with $a \in \{1, 2, \ldots, 10\}$. Fitted functions are in color with its experimental data in grey underneath

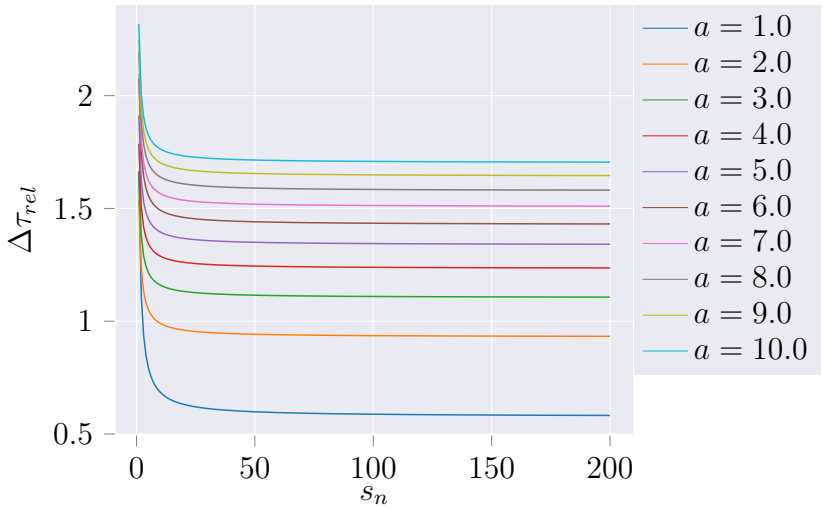


Figure 4.26: Relative resolution limits for functions with $a \in \{1, 2, \ldots, 10\}$

functions provided a tool box that can answer questions when designing NIR spectroscopy experiments or in-field machines:
Is it possible to resolve two peaks that are $\Delta\lambda$ away from each other?
What resolution does a sensor need to resolve those peaks?
An algorithm detected two peaks with scale s and distance $\Delta\lambda$. Is it generally a valid detection?
Especially the last question can be posed when applying any peak detection algorithm. Instead of only using the energy between wavelet and tested data, a second feasibility check can be implemented that discards peaks that are theoretically not possible to resolve, according to Eq. 4.46 and 4.47. The first question can give an indication if a detection problem can be solved in a certain domain, e.g., if two close functional groups (or its overtones) in an NIR spectrum can be theoretically detected by high resolving equipment or not. Subsequently, this question can be also used in anomaly detection: When knowing the expected anomaly and its functional group, validating the resolution by this sampling theorem can give indication if a detection is possible or if another method is required.

5. Implementation of a Peak Detection Algorithm using Wavelet Transformation

Based on the findings from Ch. 3 and 4, an algorithm named *wavelet analysis* is implemented. It is used for the analysis of vegetable oils in Ch. 6.
The programming language was Python 3.6, with the support of the scientific frameworks *numpy* (Oliphant [28]), *scipy* (Jones et al. [19]), *pandas* (McKinney [25]), *scikit-learn* (Pedregosa et al. [29]), *scikit-image* (van der Walt et al. [37]) and *matplotlib* (Hunter [18]).

5.1 Database Extraction

All hyperspectral data are stored in an in-house database named "HyperspectralDB". It provides an HTTP REST-based interface, which is wrapped in a Python SDK for simple access. All data sets can be queried by name or material (e.g., "oil"), and selected data sets are returned. The SDK is built as a dedicated python package because of separation of concerns.

5.2 Preprocessing

Two of the major preprocessing techniques for hyperspectral data are implemented:

Standard Normal Variate

Standard Normal Variate (SNV) is a combined normalization and scaling technique. It subtracts the average signal from the original signal and divides it by the standard deviation. Rinnan et al. [31] provide the formal definition

$$\boldsymbol{x}_{corr} = \frac{\boldsymbol{x}_{org} - a_0}{a_1} \tag{5.1}$$

with

$$a_0 \coloneqq \frac{\sum_{i=1}^{N} x_{org,i}}{N} \ , \quad N = |\boldsymbol{x}_{org}|$$
$$a_1 \coloneqq \sqrt{\frac{\sum_{i=1}^{N} (x_{org,i} - a_0)^2}{N}} \ . \tag{5.2}$$

While many works use SNV as a normalization technique, the resulting data x_{corr} is susceptible for distortions in the signal. Assuming that a spectrum includes some unexpectedly intense peaks due to incorrect handling of equipment or any other source of noise, the complete spectrum will be normalized using that peak implicitly in a_0 and a_1. Additionally, when comparing two spectra with different λ_{min} and λ_{max}, peaks from other bands will be included in the SVN in one data set but not in the other. Instead of developing a new SNV implementation in Python, the *scale* function from Pedregosa et al. [29, sklearn.preprocessing.scale] is used. It provides an implementation of the exact functionality required.
For the applied anomaly detection in Ch. 6 no SVN is used to test the algorithm on unaltered spectral data. For reference, Fig. 3.5b shows SNV normalized data.

Derivation

According to Rinnan et al. [31, p. 1213], "derivatives have the capability to remove both additive and multiplicative effects in the spectra and have been used in analytical spectroscopy for decades". The additive effect, i.e., the baseline or offset, can be removed by using the first derivative of the data. To remove any multiplicative effect, i.e., trend or slope, the second derivative needs to be used. Fig. 5.1 visualizes how signals are de-trended by applying derivation. The exemplary function is a Gauss function which is similar to the Lorentz function. The second derivative is also known as the (inverse) *Mexican hat*, due to its shape. As an example, Fig. 5.2a shows the absorbance spectra for crystal and powdered sugar. The inducted light has a smaller chance to be reflected by powdered sugar due to its smaller particle size. Therefore, the full absorbance spectrum has a lower baseline than the one for crystal sugar. When deriving the data twice, both signals have their baseline at 0 and overlapping absorption peaks (cf. Fig. 5.2b).

The derivative is implemented using the *gradient* function from Oliphant [28, numpy.gradient]. It has the advantage of using forward/backwards differences at the boundaries, so the derived signal has the same length as the original signal. Additionally, it handles different spacing between values which is a characteristic of sensor data as described in Sec. 6.2. Of course, when using the first or second derivative of the data, also a derived version of the wavelet function needs to be used.

5.3 Wavelet Analysis

The adapted wavelet transformation is described and formally defined in Sec. 4.2. Using the recommended window sizes from Sec. 4.3, a complete wavelet analysis algorithm is implemented. The algorithm takes a matrix of spectra as input, which are then tested for a fit with the user-specified wavelet function at all positions. Normally, a single spectrum or array of spectra would be sufficient, yet, the implementation is designed to also handle hyperspectral images in form of a 3D matrix.

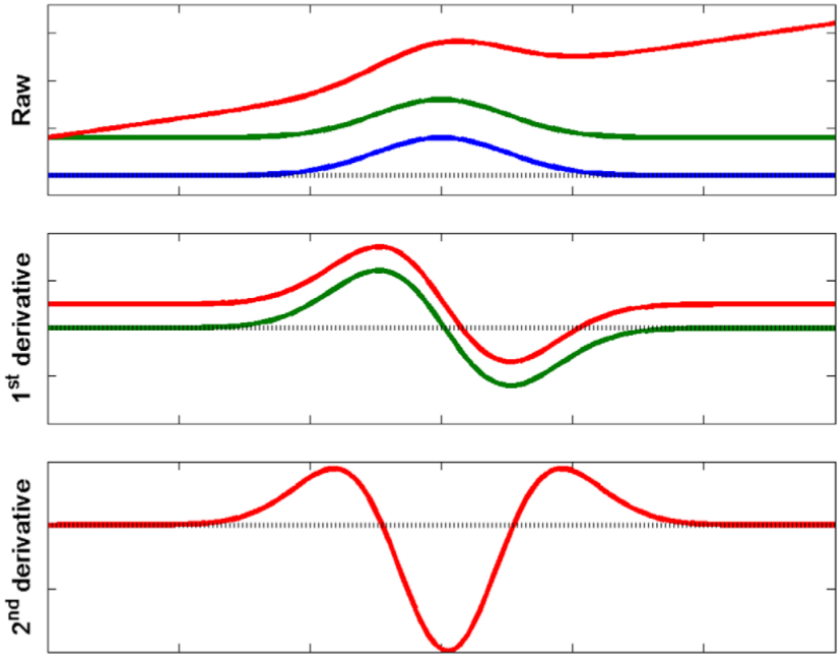


Figure 5.1: Blue: No Baseline, no trend; Green: Baseline, no trend; Red: Baseline, Trend. Adapted from Rinnan et al. [31, p. 1213] with permission.

The to be tested wavelet and scale parameters are also provided as a function parameter in nm units. The algorithm calculates the recommended window size from the provided scale parameters and executes the wavelet transformation for every combination. The results are stored in a 2D matrix as a surface that can be used by any peak detection algorithm.
To improve the runtime of the algorithm, it is implemented as a multi-core application. Firstly, all of the input spectra can be processed independently, secondly, the convolution of wavelets with different scales can be parallelized, as long as the results are stored in the correct order.
The pseudo code in Alg. 1 is a non-parallel implementation of the wavelet analysis, accepting a single spectrum. The handling of multiple spectra would have made the pseudo code unnecessarily more complicated without adding value to the core functionality. Additionally, in *line 16*, the wavelet function takes the scale parameter and the window size, while originally defined in Eq. 4.9 as also receiving the shift parameter l. The practical implementation changed to a symmetric wavelet around zero, with $|w|$ data points. This reduces the amount of data processed in every scalar product. In Subsec. 4.2.1, the usage of a window range is compared against a single window. In Alg. 1 in combination with Alg. 2 a window range is used, which is reduced by simply taking the maximum value of all tested window sizes for each wavelength. To allow an easy interchange of these selection methods, the code section (from *line* 21 to 29) is more inefficient than necessary.

5.4 Peak Detection

After returning the wavelet analysis results as a 2D matrix, a peak detection algorithm is needed to decide where absorption peaks are present and which scale they

Algorithm 1 Wavelet Analysis

procedure WAVELETANALYSIS(spectrum, wavelet, scales)
  $N \leftarrow$ length of *spectrum*
  $result_{wavelet} \leftarrow$ matrix with size $|spectra|$ x N
  for s in *scales* **do**
    $w_{min} \leftarrow$ OPTIMALMINWIN(s)
    $w_{max} \leftarrow$ OPTIMALMAXWIN(s)
    $\mathcal{W} \leftarrow \{w_{min}, w_{min} + 1, \ldots, w_{max}\}$
    $results_w \leftarrow$ array of length $|\mathcal{W}|$
    for each w in $\mathcal{W}$ **do**
      $result_{w,temp} \leftarrow$ array of length N
      $w_{half} \leftarrow \lfloor \frac{w-1}{2} \rfloor$
      $spectrum_{pad} \leftarrow pad\_with\_zeros(spectrum, w_{half})$
      for $i \leftarrow 0; i < N; i \leftarrow i + 1$ **do**
        $spectrum_i \leftarrow spectrum_{pad}[i - w_{half} : i + w_{half}]$
        $spectrum_i \leftarrow \frac{spectrum_i}{\|spectrum_i\|_2}$
        $\psi_i \leftarrow wavelet(s, w)$
        $result_{w,temp}[i] \leftarrow \langle spectrum_i, \psi_i \rangle$
      end for
      $results_w \leftarrow results_w + result_{w,temp}$
    end for
    $result_{s,max} \leftarrow$ array of length N
    for $i \leftarrow 0; i < N; i \leftarrow i + 1$ **do**
      $max_i \leftarrow 0$
      for each w in $\mathcal{W}$ **do**
        $max_i \leftarrow \max(max_i, results[w, i])$
      end for
      $result_{s,max}[i] \leftarrow max_i$
    end for
    $result_{wavelet}[s] \leftarrow result_{s,max}$
  end for
  return $result_{wavelet}$
end procedure

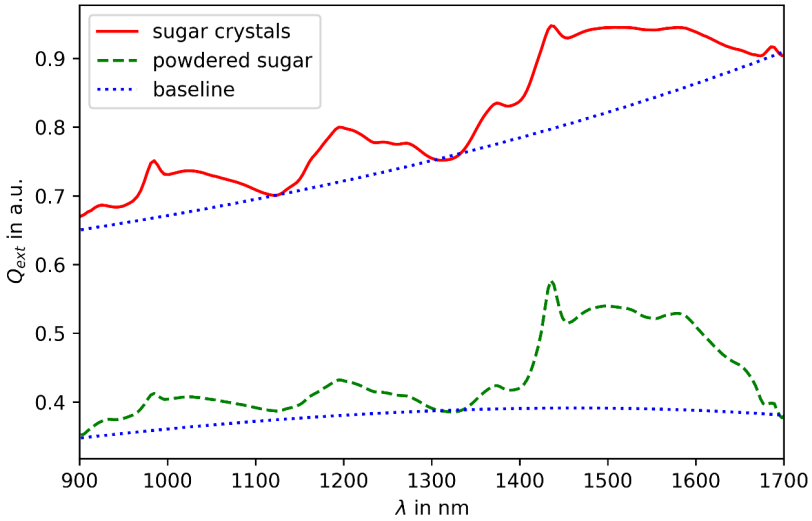
Algorithm 2 Optimal Window Size

procedure OPTIMALMINWIN(s)
  $window_{min} \leftarrow 0.433 * s + 0.492$ // cf. Subsec. 4.3.1.3
  return $window_{min}$
end procedure
procedure OPTIMALMAXWIN(s)
  $window_{max} \leftarrow 1.574 * s + 1.768$ // cf. Subsec. 4.3.1.3
  return $window_{max}$
end procedure

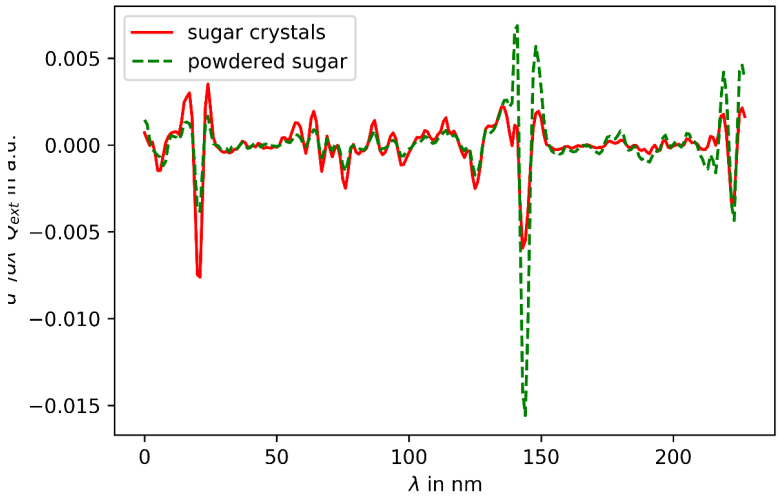
have. During the research it became more and more visible that a well-performing peak detection algorithm needs to be created. This detection algorithm could incorporate research results from Sec. 4.3 and 4.4. Within the limits of this work, no dedicated peak detection algorithm could be created. As an alternative, the *peak_local_max* algorithm by van der Walt et al. [37, skimage.feature.peak_local_max] was used. It was created to detect local maxima within 2D images, as for example the results of the wavelet analysis, based on absolute or relative thresholds. The most sensitive parameter is the number of pixels around a potential peak, that need to be less intense than the target pixel. When using finer scale parameter resolution, the number of pixels in the wavelet analysis result becomes larger, so the filter parameter needs to be refined. As a result of this local peak detection, one wavelengths can be assigned multiple peaks at different scale parameters. Since there can be only one peak per wavelength, a simple elimination procedure was chosen, i.e., accepting the peak with the highest wavelet transformation result at its respective scale. The result of Alg. 3 is an array of scales for every wavelength. If the value is zero, no peak was detected at that wavelength, otherwise the detected peak had the value as scale.

Algorithm 3 Peak Detection

procedure DETECTPEAKS($result_{wavelet}$)
  $N \leftarrow$ length of $columns(result_{wavelet})$
  $\boldsymbol{s} \leftarrow$ title of $rows(result_{wavelet})$
  $S \leftarrow$ length of $\boldsymbol{s}$
  $peaks_{map} \leftarrow$ PEAK_LOCAL_MAX($result_{wavelet}$)
  $peaks_{reduced} \leftarrow$ array of length N
  for $i \leftarrow 0$; $i < N$; $i \leftarrow i + 1$ **do**
    $s_{max} \leftarrow 0$
    $g_{max} \leftarrow 0$
    for $j \leftarrow 0$; $j < S$; $j \leftarrow j + 1$ **do**
      $g_{curr} \leftarrow peaks_{map}[i, j]$
      if $g_{curr} > 0$ and $g_{curr} > g_{max}$ **then**
        $s_{max} \leftarrow \boldsymbol{s}[j]$
      end if
    end for
    $peaks_{reduced}[i] \leftarrow s_{max}$
  end for
  return $peaks_{reduced}$
end procedure



(a) Underived spectra



(b) 2x derived spectra

Figure 5.2: Absorbance spectrum for differently shaped sugar. Adapted from Krause [23] with permission.

6. Applied Anomaly Detection in Oil NIR Spectra

Olive oil is amongst the products which have been repeatedly involved in food fraud cases. Most consumers prefer it due to its high amount of monounsaturated fatty acids and sensory features as smell and taste. (Mono)Unsaturated fatty acids are considered to be healthy and were part of what is labeled today as the *Mediterranean diet*. Willett et al. [39] define this diet as "food patterns typical of Crete [...], and southern Italy in the early 1960s". They found that "adult life expectancy for populations in these areas was among the highest in the world, and rates of coronary heart disease, certain cancers, and some other diet-related chronic diseases were among the lowest in the world in the early 1960s, despite limitations of existing medical services"(Willett et al. [39][p. 1402S]). While, undoubtedly, the Mediterranean diet consisted of several components attributing to its positive health impact, olive oil is defined as "the principal fat". Willett et al. [39][p. 1404S] list multiple reasons why olive oil is preferable to other oils, including that "diets high in monounsaturated fat are less likely than those high in polyunsaturated fat to be involved in the oxidation of low-density lipoproteins (LDLs), a process thought to increase the risk of atherogenesis and coronary heart disease". Due to these health benefits, the demand for olive oil is persistent and prices for olive oil are higher than for most other vegetable oils used in the daily kitchen. Unfortunately, high demand and high prices create a fruitful base for "adulteration of food products, involving the replacement of high-cost ingredients with lower grade and cheaper substitutes [...] for a food manufacturer or raw material supplier. Adulteration of food products is not only a major economic fraud but can also have major health implications for consumers"(Tay et al. [34][p. 99]). One of the larger-scale "health implications" were found and described by Kilbourne et al. [20] as the "Spanish toxic-oil syndrome": In 1981, Spain had an unexpected outbreak of pneumonitis, which could be later attributed to illegally marketed cooking oil. This oil is thought to have been marketed as olive oil and sold in unlabeled 5-liter plastic containers. While the exact content of these containers could not be determined, it was recognized that it had a high proportion of rapeseed oil. Rapeseed oil is generally not creating any health problems, yet, for industrial purposes it is denaturated using Aniline. Due to its

cheap price, Kilbourne et al. [20] suspect that a following de-denaturation process, i.e., the removal of Aniline, might have created substances that are responsible for the epidemic.
Admittedly, the *Spanish toxic-oil syndrome* is an extreme example, but it shows that anomalies can be defined by the most unexpected substances or adulterations.
In addition to the real-life olive oil food fraud problem, the decision to use oils as a trial food for anomaly detection was mainly based on the following properties: Oils are of homogeneous matter, which makes a NIR transmittance measurement reliable. As opposed to the recurring apple example, there are no offsets in measure spectra by the angle the light hits the surface, and no hidden anomalies (like bad apple flesh under the skin) that could be mistaken for a normal state. Additionally, when mixing different oils, a new state of homogenization can be reached without addition of any solvents or using a disproportional amount of time for the stirring process. Also, vegetable oils are available in different varieties and quantities in any supermarket.

6.1 Analyzed Oil Types

For this work a selection of olive oils and other vegetable oils was done. Using the analysis from Stiftung Warentest [33], one olive oil with result *good* in chemical and sensory quality (**OL2**) and one olive oil with *poor* sensory quality (therefore, raising suspicion of an adulterated state) (**OL1**). Since their test was executed using another season's olive oil, it can only be assumed that the quality levels were comparable at time of buying the products for this experiment.
The choice for other vegetables was doing creating a broad variety of saturated, monounsaturated and polyunsaturated fatty acids: rapeseed oil (**RP**) has one of the lowest saturated fatty acid levels amongst vegetable oils. Sunflower (**SO**) and linseed oil (**LS**) both have high levels of polyunsaturated fatty acids, but have a transparent or yellow color, respectively. The last selected oil is a Styrian[1] pumpkinseed oil (**KB**). It was not selected due to its nutrition contents, but due to its very dark green color. Fig. 6.1 shows a picture of all selected oils with their original bottles and samples on partially white background for visual color and opaqueness comparison. Table 6.1 shows all analyzed oils by their abbreviation and their contents of fatty acids, sorted in ascending order by saturated fatty acids (more details can be found in Table A.2. The displayed values are taken as far as possible from the oil container's labels. If no values were provided by the manufacturer, they were estimated using Heseker and Heseker [17], of course while maintaining ratios between fatty acid types.
Due to limitations, no chemical analysis of the samples were executed, but the provided information from manufacturers and nutrition tables had to be used. Unfortunately, the provided nutrition facts are a gross average over an oils' season and can derive from product to product.

6.2 NIR Spectroscopy Sensor and Equipment

For measuring the transmittance spectrum of the oil samples, a setup was designed that is outlined in Fig. 6.2. The basic structure was: a source emits light into a

[1] **Styria** is a state in south-east Austria, sharing a border with Slovenia.

Figure 6.1: Overview of selected oils

Table 6.1: Fatty Acids (FA) comparison of selected oils per 100 g (values marked with * were not provided by manufacturer and have been estimated using Heseker and Heseker [17])

Abb.	Saturated FA [g]	Monounsaturated FA [g]	Polyunsaturated FA [g]
RP	7.6	62.0	30.4
SO	9.7	21.9*	68.4*
LS	9.8	18.5	71.7
OL1	15.0	78.9	6.1
OL2	15.4	74.9*	9.8*
KB	17.6*	29.7*	52.7*

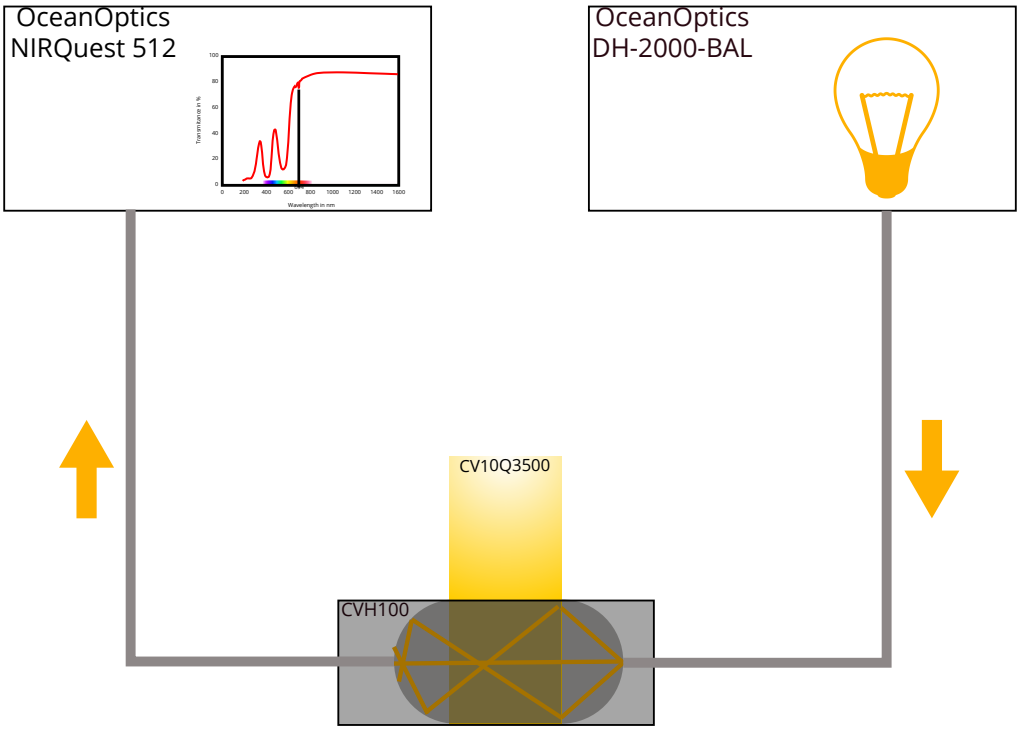


Figure 6.2: Setup of devices for oil experiment

conducting medium, which emits the light again through the sample. Afterwards the transmitted light is inducted into another conducting medium, and then sensed by a NIR spectroscope.

All parts are hereinafter described in greater detail:

Light Source: OceanOptics DH-2000-BAL

The *DH-2000-BAL* has two light sources, of which the Tungsten-Halogen lamp was used for NIR spectroscopy. It has a wavelength range from 300 to 1500nm with focused radiation. According to the specifications sheet by the manufacturer (OceanOptics [26, p. 6]), the stability and drift are $\leq 0.01\%/h$ @ 700 nm. The light source is actively cooled by a built-in cooling fan.

Light Conductor: THORLABS BF20LSMA

For in- and outgoing at the cuvette holder, the *BF20LSMA* multi-mode fiber bundle was used. The total diameter is 2.0 mm, created by 7 individual fibers in round shape. The supported wavelength range is from 400 to 2200 nm. Both of the fibers had a length of 1 m, each side equipped with a *SMA 905* connector, which the default connector for all used equipment in this experiment.

Cuvette Holder: THORLABS CVH100

For measuring transmittance, the *CVH100* was selected. Besides allowing the addition of optical filters in-line, which were not used, it supported one main optical axis

and one secondary axis. The beam height was set at 8.5 mm above the base of the inserted cuvette. The supported outer cuvette dimensions were 12.5 mm x 12.5 mm x 45 mm, according to THORLABS [35]. Recalling the recommendation by Workman and Weyer [41] described in Subsec. 3.1, the chosen path length $l = 12.5$mm through the cuvette appears to be optimal.

Cuvette: THORLABS CV10Q3500

Given the selected cuvette holder, cuvettes with the fitting outer dimensions were required. Thorlabs' *CV10Q3500* met this criterion. The cuvettes were made from UV fused quartz and had two polished and two frosted sides. Having two opposite frosted sides made the handling easier and was more robust towards any remainders from fingerprints. The capacity was 3500 µl. Since the cuvette itself has a reduced transmission spectrum (cf. THORLABS [36, Fig. "UV Fused Quartz Cuvettes Transmission"]), every calibration was executed with an empty cuvette inside the holder for compensation.

Spectrometer: OceanOptics NIRQuest512-1.7

The spectrometer was the *NIRQuest-512* with a detection range from 900 to 1700 nm using a *Hamamatsu G9204-512* sensor with 512 pixel resolution. Therefore, the effective resolution was at $\approx (1700\text{nm} - 900\text{nm})/512 = 1.56\text{nm}$. [2] The device's product sheet (cf. OceanOptics [27]) sets the signal-to-noise ratio at 15000 : 1 at 100 ms integration time. For the experiment an integration time of 50 ms was used, so the given signal-to-noise-ratio could be assumed to be similar in the recorded spectra.

Software

For capturing data from the sensor, an in-house developed software from *Fraunhofer IOSB SPR* department was used. It was configured to record a spectrum with an integration time of 50 ms.

All samples were transferred to glass cuvettes and measured between 900 nm and 1700 nm with an average resolution of 1.6 nm, accumulating 15 scans per spectrum. The accumulation (and later averaging) of one spectrum can be used to filter noise in the data and increase the signal-to-noise ratio. Assuming noise in the signal to follow a random process with a certain distribution, it will eventually average out to zero, given a sufficiently large number of recordings over time. The time to record 15 spectra was on average 20 s. Recordings were obtained at room temperature, using no hardware to control the exact temperature of the samples.
For oil mixtures, oils were sequentially added to the cuvette and its relative concentration determined by weight, assuming the same density for all oils. For transferring oil from their source container to the cuvette, regular plastic *Pasteur pipettes* were used. An exemplary weigh-in process can be seen in Fig. 6.3a, applying the correct share in SB oil with the help of a special accuracy weighing machine. After filling the cuvette, the content was fully homogenized by stirring and visual inspection. Fig. 6.3b shows how mixture samples were not sufficient for measurement and needed an extended manual homogenization process.

[2] The measurements showed that the true distance between recorded wavelengths was not always equidistant, but increased slightly with the wavelength. On average, $\Delta\lambda$ was 1.5996 nm, with a standard deviation of 0.0243 nm.

(a) Weigh-in process (b) Non-homogenized KB-OL2 mixture

Figure 6.3: Creation of mixed samples

6.3 Qualitative Analysis of Fatty Acids

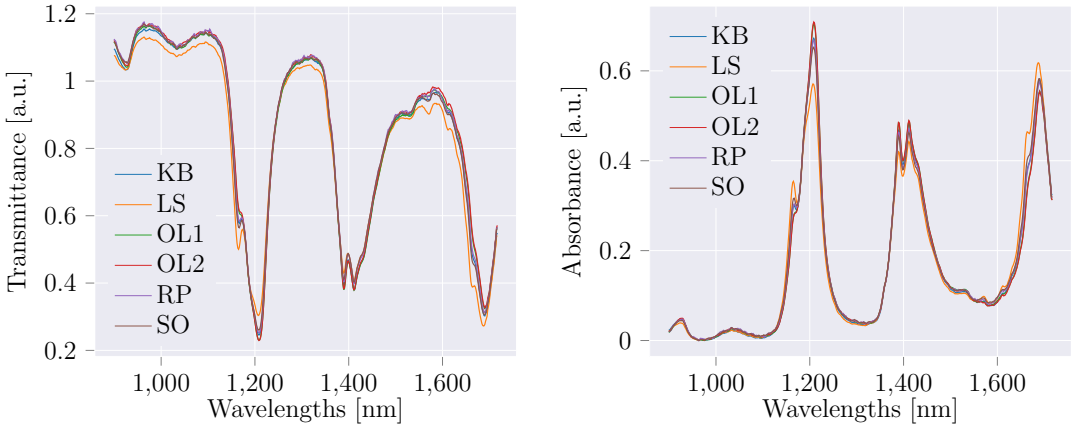
All oils were recorded as pure 100% mixtures and as all possible combination with ratio 25/75, 50/50 and 75/25. For **OL2** and **LS** additional mixtures with finer resolution (90/10, 80/20,...,20/80, 10/90) were measured. The exact ratios and weighed samples can be found in Tables A.3, A.4, A.5 and A.6.

The first qualitative analysis was executed on pure oil samples. Fig. 6.4a shows the recorded spectra for all six oils. The relation between transmittance T and absorbance A has been described by Workman and Weyer [41, eq. 1.3] as

$$T = \frac{I}{I_0} = 10^{-\epsilon c l} \implies A = -\log_{10}(\frac{I}{I_0}) = -\log_{10}(T) = \epsilon c l \quad . \tag{6.1}$$

Using the common logarithm requires $T \in [0, 1]$, so a normalization of the data is required by dividing through the maximum transmittance value in the spectrum, assuming that $min(T) \geq 0$. In Fig. 6.4 both the recorded transmittance and the converted absorbance spectrum are shown. Admittedly, any algorithm could easily work on either of those spectra, but the absorbance spectrum has a more intuitive perspective. The higher the peak, the higher the absorbance, the higher the content of a functional group in the sample.

When looking on the absorbance spectrum in Fig. 6.4b, all of the spectra appeared to follow a similar function, except with different intensities at absorption peaks. This behavior was a confirmation that the measurement of oils was successful, and that there are some chemical differences which are represented by different absorption at different wavelengths. Recalling the functional group assignment from Table 2.1, based on Woodcock et al. [40], the visible peaks at 1168 nm and 1211 nm can be attributed to the second overtone of C-H stretch vibrations in CH3— and —CH2— functional groups, respectively. The —CH2— functional group is associated with the content of unsaturated fatty acids and can be used to predict the content inside



(a) Transmittance (b) Absorbance

Figure 6.4: Transmittance and Absorbance spectra for pure oils

the sample. Additionally, the peaks from 1390 to 1410nm are also due to —CH2— first/second overtone vibrations in combination with ArOH, which is an indicator for Vitamin E and polyphenol (cf. Gómez-Rico et al. [13]) content. Fig. 6.5 shows magnified absorbance spectra. Mapping the sequence of oils by absorption intensity to the nutrition facts in Table 6.1, indicate that the absorption at 1168 nm is related to polyunsaturated fatty acids and at 1211 nm to monounsaturated fatty acids.

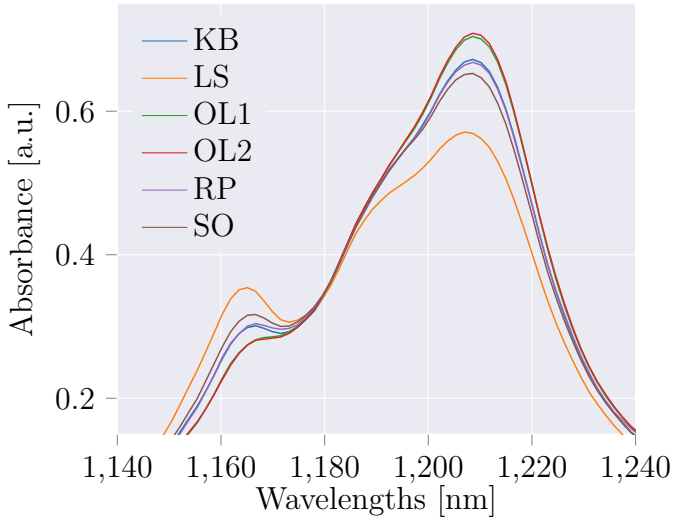


Figure 6.5: Absorbance spectra from 1140 to 1240nm

6.4 Prediction of Fatty Acids Content

6.4.1 Partial Least Squares

Partial Least Squares regression (PLS) is a common tool for training regression models on high-dimensional data, e.g., hyperspectral data. Also for olive oil analysis PLS

is one of the favorite tools. In Armenta et al. [2, Table 2] is an overview of NIR based methods used in different papers to determine quality parameters of olive oil (i.e., amongst others, fatty acids and peroxide values). Almost all of those listed publications use PLS regression or its close neighbor *Principal Component Regression* (PCR). Both methods reduce the number of dimensions to a smaller amount. This is an important property when analyzing high-dimensional data, due to the *curve of dimensionality*. Hastie et al. [16, p. 23] draw a simple analogy that explains the problem as follows: "consider the nearest-neighbor procedure for inputs uniformly distributed in a p-dimensional unit hypercube [...]. Suppose we send out a hypercubical neighborhood about a target point to capture a fraction r of the observations. Since this corresponds to a fraction r of the unit volume, the expected edge length will be $e_p(r) = r^{1/p}$. In ten dimensions $e_{10}(0.01) = 0.63$ and $e_{10}(0.1) = 0.80$, while the entire range for each input is only 1.0. So to capture 1% or 10% of the data to form a local average, we must cover 63% or 80% of the range of each input variable. Such neighborhoods are no longer "local"". One can easily imagine that having 512 dimensions (as in this experiment) drastically reduces the ability to train a regression model with "classical" methods, e.g., *Multiple Linear Regression* (MLR), which embeds a hyper-dimensional linear function into the training data. One property that differentiates PLS from PCR, according to Hastie et al. [16, p. 81], is that it does not only use the input variables but also the response variables for creating the solution path: "it can be shown [...] that partial least squares seeks directions that have high variance *and* have high correlation with the response, in contrast to principle component regression which keys only on high variance [...]". Recalling the previous qualitative analysis of the oil spectra, the property of using the correlation between response and explanatory variables is a favorable property. Just by visual inspection there appeared to be a correlation between higher unsaturated fatty acids (response) and higher absorption at $\approx 1200\,\text{nm}$ (explanatory). Using that information while building a regression model will definitely improve the model or at least accelerate the training.

To have a baseline for comparing performance of methods, three PLS regression models were trained and tested for saturated fatty acids, monounsaturated fatty acids and polyunsaturated acids. All 6 pure and 54 mixed oil samples (cf. Table A.3-A.6) were used. The data were split into training (60%) and test (40%) data. The training phase was executed using 10-fold cross validation, of which the best performing model was selected. This was repeated for maximum number of 20 principal components, of which again the best performing model was selected. Finally, the selected model was tested using the test data, and the RMSE and R2 calculated. With respect to explained variance, the worst performing response were saturated fatty acids with $R2 = 0.766$. The $RMSE = 1.104\text{g}$ is the smallest amongst all response variables, but comparing it to the average content of saturated fatty acids in oils (cf. Table 6.1), it is relatively high. Predicting the content of mono- and polyunsaturated fatty acids had $R2 > 0.9$ and an RMSE of 4.688g and 4.047g, respectively. Comparing this again to the qualitative analysis from above, these results appear to be explainable. While there were clear indications that absorptions were due to unsaturated fatty acids, there was no direct hint about saturated fatty acids. More information regarding the number of principal components, factors for PC1 and the test prediction can be found in Fig. A.1, A.2 and A.3.

As a result it can be concluded, that PLS is a convenient tool to train regression models for fatty acids based on a sufficiently large training data base. Based on

these regression models, *expected* anomalies in variables can be detected, e.g., too high content of saturated fatty acids. The criterion for labeling a datum as an anomaly is another task, but one could imagine to create density-based clusters for *normal* olive oil's saturated, monounsaturated and polyunsaturated fatty acids values, and use the size of the border sphere as a fixed threshold.

6.4.2 Multiple Linear Regression using Wavelet Analysis

The wavelet analysis described in Ch. 4 and implemented in Ch. 5 was tested if it could serve as a robust dimensionality reduction for regressing any of the fatty acids, as it was done with the PLS. Due to the reduced dimensions, classical learning methods can be used. In this case, a simple *Multiple Linear Regression* (MLR) was trained.
First, all oils were searched for absorptions using a Lorentz function as wavelet and $s = \{30, 31, \ldots, 100\text{nm}\}$. No derivation or smoothing filter was applied, to have "unaltered" results for comparison. The results from the wavelet transformation were detected for peaks by using *peak_local_max* (van der Walt et al. [37, skimage.feature.peak_local_max]) function with a custom filter mask of size 10$\dot{5}$0. Border regions were excluded from the filter search. The resulting peaks were searched for directly neighboring wavelengths, and eventually merged into the peak with the larger detected scale. Using a more complex validation approach based on the sampling theorem described in Sec. 4.4 was not necessary, since all peaks were sufficiently distant to each other and could be easily validated by visual inspection. Fig. 6.6 shows an exemplary result from wavelet transformation and peak detection. Peaks were found at 1030.55 nm, 1166.71 nm, 1206.96 nm, 1389.13 nm and 1411.36 nm. These include the relevant absorption wavelengths for fatty acids, as described at the beginning of this chapter.

After using the same parameters for wavelet transformation and peak detection on all 60 samples, the absolute frequency of peaks at their respective wavelengths was counted. Fig. 6.7 shows how the 4 most frequent peaks are always at the same wavelengths without any deviations. For peaks from 1030 to 1040nm the frequency is still clearly visible, but also not at a clear position. Using this information, the decision was made to use the top 4 peak wavelengths for the MLR. These were located at 1166.71 nm, 1206.96 nm, 1389.13 nm and 1411.36 nm.

With respect to prediction of prediction unsaturated fatty acids, the MLR performed worse than the PLS. For monounsaturated fatty acids the RMSE was 11.076g and $R2 = 0.578$, for polyunsaturated fatty acids they were 10.363g and 0.664, respectively. The model for saturated fatty acids had the best performance results with a RMSE of 1.024g and $R2 = 0.798$. Admittedly, when predicting any of the fatty acids, the PLS using all available information is the better choice. Yet, the wavelet analysis reduced the features dramatically from 512 to 4 and still had very acceptable results. Additionally, the influence of certain wavelengths is better interpretable in the MLR. The factors of the wavelengths into the linear model can be compared in Fig. A.4, A.5 and A.6.
Since the models were solely trained on the intensity, the wavelet transformation only served for detecting frequent peaks as a feature reduction. Yet, while finding the peaks, each of them is assigned a position and a scale. So far, the information regarding the scale had not been using in the regression model. Therefore, the detected scale at all top 4 wavelengths were extracted and used as explanatory

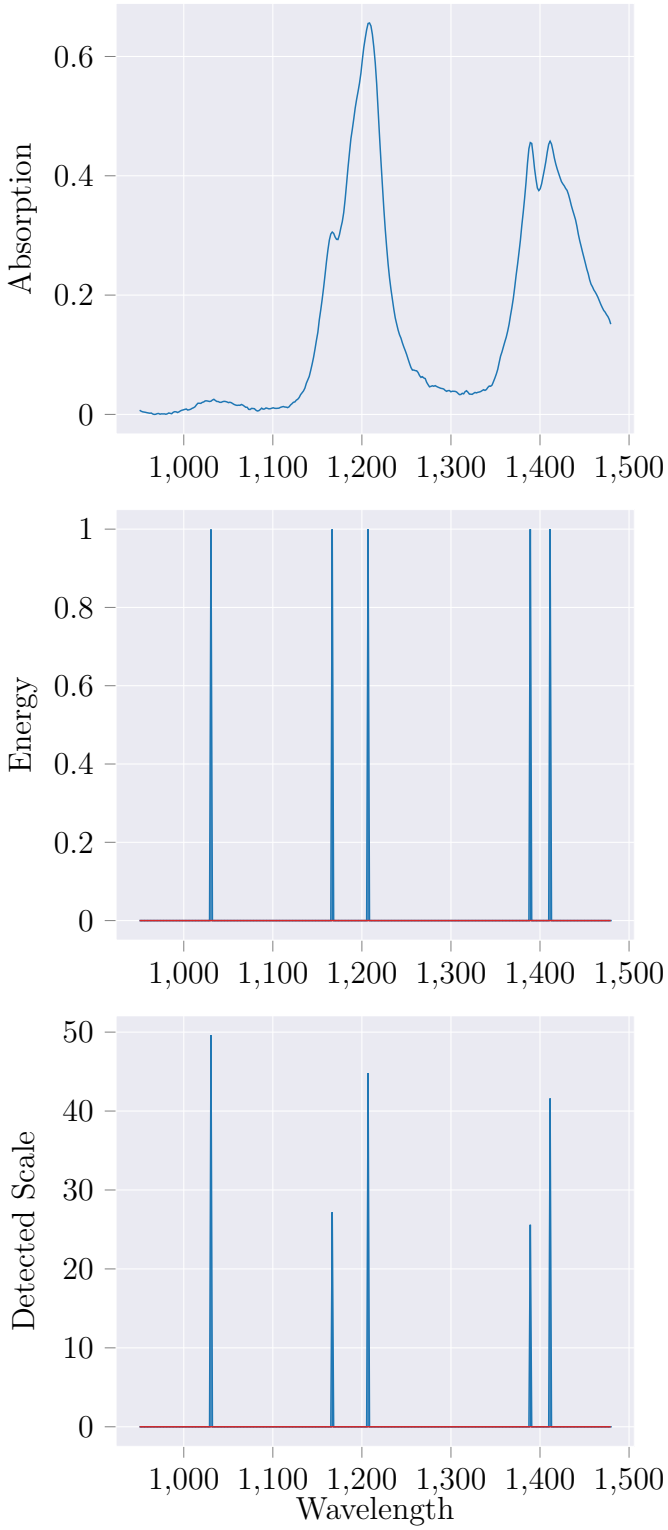


Figure 6.6: Spectrum, wavelet energy and detected scale for a SO-RP (25:75) sample

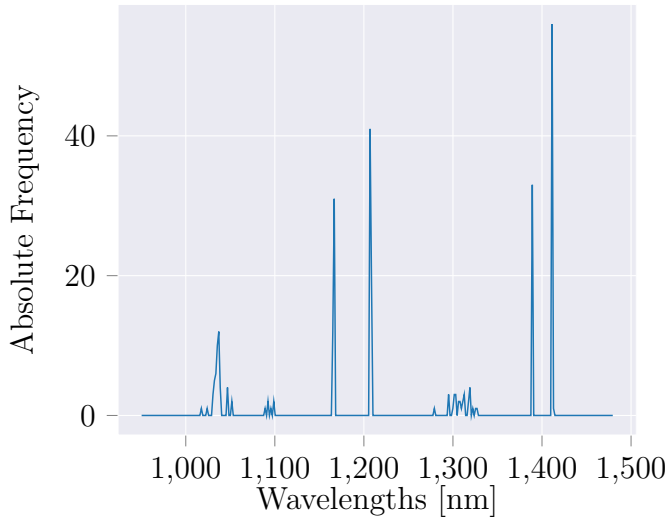


Figure 6.7: Absolute frequency of detected peaks across all 60 samples

variables for a linear regression model. The results were disappointing, for saturated fatty acids the R2 was 0.086, and for both unsaturated fatty acids between 0.3 and 0.4. So the detected scale of peaks can not solely be used for regressing fatty acids. A combination of both the intensity and the detected scale as explanatory variables yielded slightly worse results, than with just using the absorbance itself. Therefore, this more complicated model was discarded. For polyunsaturated fatty acids, both models can be compared in Fig. A.7 and A.8.

6.5 Detection of Unknown Anomalies

The previous section predicted the concentration of molecules in the matter using multiple training samples to either search for principle components (PLS) or the most frequent detected peaks (wavelet analysis). By defining a range of concentration values that are considered to be normal, anomalous behavior can be detected. Recalling the three types of anomaly from Sec. 3.3 which can occur, only **Changed Absorption** and **Removed Absorption** can be detected.
The wavelet analysis is designed to also detect anomalies that are caused by **New Absorption**. If an olive oil samples is contaminated with substances that have their absorptions at different regions than saturated and unsaturated fatty acids, the wavelet analysis will detect a new peak. By applying the L^2 norm, the wavelet transformation returns $0 \leq \Gamma_s \leq 1$, which can be interpreted as a confidence that a peak with s occurred at a position. If a new peak receives a high confidence, it can be considered to be an anomaly. Methods as MLR or PLS will consider the new peaks due to a low coefficient, but some distance-based methods could.
One proposal for a distance-based methods is the creation of hashes, of which distances (or vice versa: similarities) can be calculated. A fixed number of peaks with highest confidence are selected, ordered by wavelength occurrence, and stored in that hash with their absorbance intensity and detected scale s. When comparing two normal samples, the distance will be small since the wavelet transformation will detect similar peaks at similar position with small variance. When comparing one

normal and one anomalous (an additional absorption peak occurred in the spectrum), the anomalous peak will be added to the hash and create a larger distance to a normal sample's hash. A sophisticated implementation that uses hashes to *fingerprint* audio files and detect similarity only based on frequency occurrences at time location, is described by Wang [38]. The hash is created based on a 2D map (time vs frequency), which has multiple anchor points. For every point in the map, a relative distance to an anchor point is stored together with the position of the anchor point in the hash. As a result, the comparison is translation-invariant, so longer audio segment can be compared with shorter audio segments.
An implementation of such an algorithm in combination with the wavelet analysis in NIR spectra would be a promising endeavor. Food samples could be stored as hashed values using one sensor and compared to hashes from another sensor, without requiring large sets of training data or scaling operations.

7. Conclusion

In this thesis it was analyzed which information about a near-infrared spectrum are required to potentially detect anomalies of three different types (changed absorption, removed absorption and new absorption). Afterwards, the wavelet transformation was adapted in a way, that such information can be detected in a NIR spectrum. By analyzing which window size for discrete data yields optimal results with respect to absorption location and scale, the wavelet transformation can be used as a tool to extract absorption peaks from most spectra for any peak scale. Furthermore, the sampling theorem defined a resolution criterion for Lorentz functions. Absorption peaks have a similar shape to that of a Lorentz function, which makes the resolution criterion applicable for peaks in NIR spectra. The result were functions which provide the minimum distance between two Lorentz peaks as a function of their scale and amplification, called the sampling theorem. The sampling theorem enables researchers and users of NIR sensors, to plan required sensor resolution according to their target detection. If close absorption bands need to be resolved, a finer device resolution is required, which can be estimated by using the sampling theorem. Vice versa, the theorem can also be used after data was recorded to discard results from peak detection algorithms. Given that two peaks were detected in a proximity, which is not resolvable, one or both of them can be rejected. This enables other algorithms to validate their results in an efficient way.
The definitions from the adapted wavelet transformation and the results for an optimal window size were implemented in a ready-to-use algorithm, named wavelet analysis. It was used to apply wavelet transformation different vegetable oil samples and search for Lorentz functions of different scales at all positions. The results were analyzed by a primitive peak detection algorithm, which extracted peak information, according to the requirements for a successful anomaly detection. These peak information were used to predict the concentration of different fatty acids in vegetable oils, and compared against Partial Least Squares regression (PLS) using the full spectrum, a method used by a majority of researchers in the fields of NIR food analysis. The results using PLS yielded better results than fitting a linear regression on four extracted peaks, but the latter method's results were still very promising. Especially, considering the dimensionality reduction from 512 to 4 dimensions.
The most difficult and least researched component in this work was the peak detec-

tion / extraction algorithm. In this work's implementation, a simple local minima finding algorithm was used that works on two-dimensional surfaces. The results from the optimal window experiment showed, though, that the occurrence of a peak can be better detected when using a custom filter, instead of a simple rectangle grid. Moreover, the selection of the most significant peaks was done measuring the absolute frequency across all oil samples. This process can be improved by selecting peaks based on their wavelet transformation result. By defining a normalization of signal window and wavelet during the adaptation of the wavelet transformation, the result is now a correlation coefficient between zero and one. This correlation coefficient can be used to cut off peaks below a certain threshold or by using a statistical procedure that selects peaks based on occurrence density.
The detection of unknown peaks was found to be possible using the wavelet analysis algorithm, yet, a concrete implementation of an algorithm still needs to be evaluated. Possible approaches based on hashes or peak constellation maps were proposed in this work and can be considered promising. Due to the way how the wavelet analysis extracts correlation between a scaled wavelet and a location in the spectrum, the foundation for such algorithms has been created.
A first regression based on extracted peak information using the wavelet analysis did not perform as well as PLS. As a conclusion, a PLS using the full spectral information can be considered more suitable to detect anomalies in NIR regions, which are known and expected possibly change. Improving the wavelet analysis algorithm by a more sophisticated peak detection algorithm could make it compete with the PLS. The results regarding determination and mean error are still a good baseline. With respect to unknown anomalies, wavelet transformation has the potential to thrive and out-perform the existing major algorithms. This makes it a robust algorithm which can be implemented in many scenarios where unknown anomalies are amongst the more dangerous ones, for example in food production.

A. Appendix

A.1 Oil Experiment Details

Table A.1: Brand and product names for all tested oils.

Abb.	Brand	Product
SO	Thomy	Reines Sonnenblumen Öl
OL2	Bertolli	Olio Extra Vergine Di Oliva
OL1	ja!	Natives Olivenöl Extra
RP	Mazola	100% reines Rapsöl
KB	Familie Reitzer-Ferl	100% reines Kürbiskernöl
LS	SchneeKoppe	Lein Öl Klassik

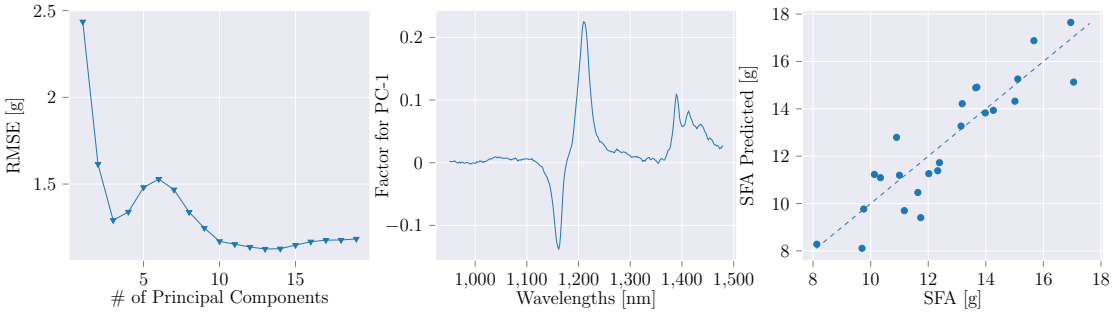


Figure A.1: PLS for Saturated Fatty Acids (SFA) with 13 PCs, $RMSE = 1.104$g, $R2 = 0.766$

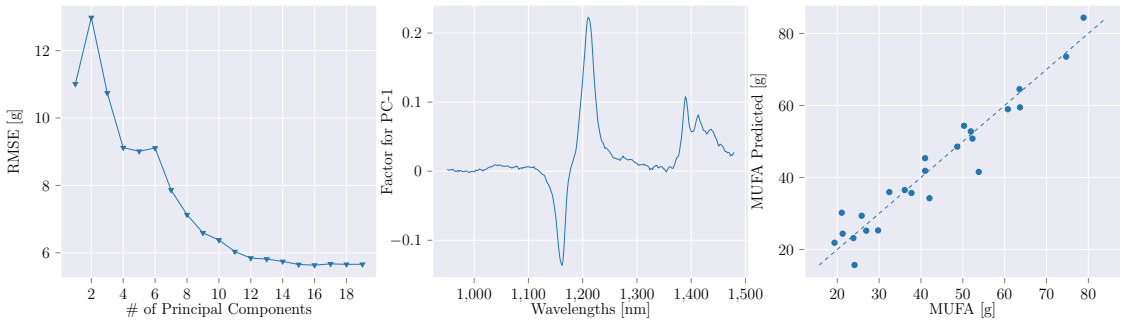


Figure A.2: PLS for Monounsaturated Fatty Acids (MUFA) with 16 PCs, $RMSE = 4.688$g, $R2 = 0.924$

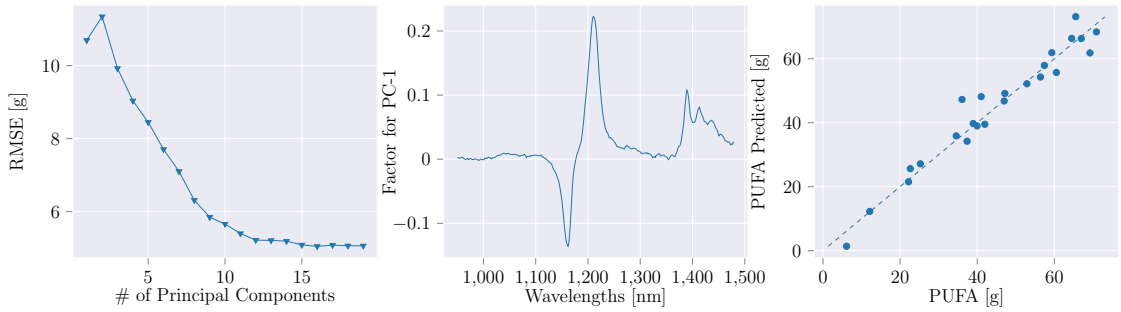


Figure A.3: PLS for Polyunsaturated Fatty Acids (PUFA) with 16 PCs, $RMSE = 4.047$g, $R2 = 0.949$

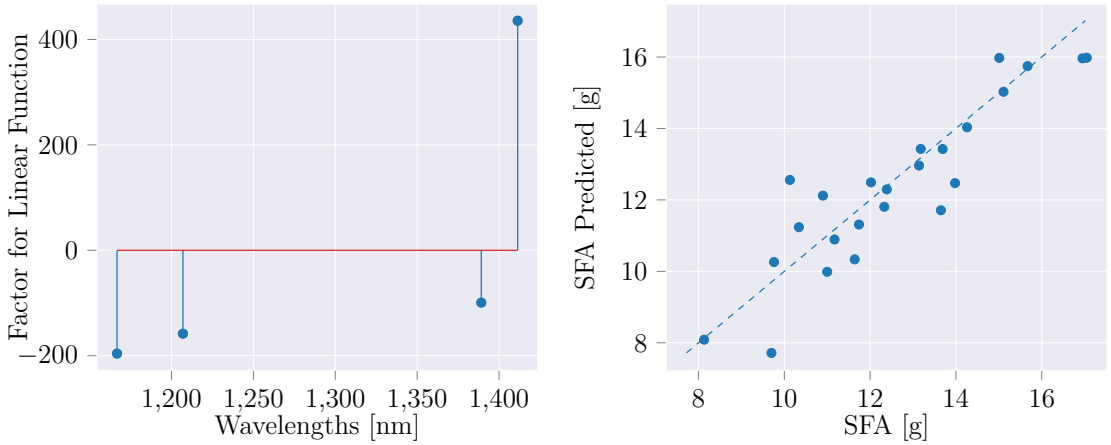


Figure A.4: MLR for Saturated Fatty Acids (SFA) with $RMSE = 1.024$g, $R2 = 0.798$

Table A.2: Nutrition of selected oils per 100 g (values marked with * were not provided by manufacturer and have been estimated using Heseker and Heseker [17]). SFA: saturated fatty acids; MUFA: monounsaturated fatty acids; PUFA: polyunsaturated fatty acids.

Abb.	Energy [kJ]	Total Fat [g]	SFA [g]	MUFA [g]	PUFA [g]	Carbohydrate [g]	Protein [g]	Sodium [g]	Vit. E [mg]
SO	3700.0	100.0	9.7	21.9*	68.4*	0.0	0.0	0.0	57.9
OL2	3712.1	100.0	15.4	74.9*	9.8*	0.0	0.0	0.0	11.9*
OL1	3699.9	100.0	15.0	78.9	6.1	0.0	0.0	0.0	11.9*
RP	3700.0	100.0	7.6	62.0	30.4	0.0*	0.0*	0.1*	32.6
KB	3762.0*	100.0*	17.6*	29.7*	52.7*	0.0*	0.0*	0.0*	3.5*
LS	3700.0	100.0	9.8	18.5	71.7	0.0	0.0	0.0	54.3

Table A.3: Oil mixtures with targeted 50 : 50 ratio

ID	Oil 1 (50)		Oil 2 (50)		Ratio
	Name	Weight [g]	Name	Weight [g]	
1	OL1	1.51	OL2	1.52	50 : 50
2	OL1	1.44	KB	1.47	49 : 51
3	OL1	1.48	LS	1.48	50 : 50
4	OL1	1.47	SO	1.48	50 : 50
5	OL1	1.48	RP	1.5	50 : 50
6	OL2	1.48	KB	1.48	50 : 50
7	OL2	1.48	LS	1.49	50 : 50
8	OL2	1.47	SO	1.51	49 : 51
9	OL2	1.49	RP	1.48	50 : 50
10	KB	1.47	LS	1.48	50 : 50
11	KB	1.47	SO	1.46	50 : 50
12	KB	1.48	RP	1.48	50 : 50
13	LS	1.48	SO	1.47	50 : 50
14	LS	1.5	RP	1.51	50 : 50
15	SO	1.5	RP	1.51	50 : 50

Table A.4: Oil mixtures with targeted 25 : 75 ratio

ID	Oil 1 (50)		Oil 2 (50)		Ratio
	Name	Weight [g]	Name	Weight [g]	
31	OL1	0.76	OL2	2.26	25 : 75
32	OL1	0.75	KB	2.24	25 : 75
33	OL1	0.75	LS	2.24	25 : 75
34	OL1	0.75	SO	2.26	25 : 75
35	OL1	0.75	RP	2.24	25 : 75
36	OL2	0.76	KB	2.28	25 : 75
37	OL2	0.74	LS	2.25	25 : 75
38	OL2	0.74	SO	2.24	25 : 75
39	OL2	0.74	RP	2.24	25 : 75
40	KB	0.75	LS	2.25	25 : 75
41	KB	0.75	SO	2.27	25 : 75
42	KB	0.76	RP	2.25	25 : 75
43	LS	0.75	SO	2.24	25 : 75
44	LS	0.74	RP	2.25	25 : 75
45	SO	0.76	RP	2.27	25 : 75

Table A.5: Oil mixtures with targeted 75 : 25 ratio

ID	Oil 1 (50)		Oil 2 (50)		Ratio
	Name	Weight [g]	Name	Weight [g]	
16	OL1	2.21	OL2	0.73	75 : 25
17	OL1	2.25	KB	0.77	75 : 25
18	OL1	2.23	LS	0.75	75 : 25
19	OL1	2.26	SO	0.75	75 : 25
20	OL1	2.26	RP	0.74	75 : 25
21	OL2	2.24	KB	0.76	75 : 25
22	OL2	2.25	LS	0.75	75 : 25
23	OL2	2.24	SO	0.76	75 : 25
24	OL2	2.25	RP	0.75	75 : 25
25	KB	2.25	LS	0.74	75 : 25
26	KB	2.26	SO	0.75	75 : 25
27	KB	2.26	RP	0.75	75 : 25
28	LS	2.27	SO	0.75	75 : 25
29	LS	2.24	RP	0.76	75 : 25
30	SO	2.26	RP	0.76	75 : 25

Table A.6: **OL2-LS** mixtures with finer ratio resolution

ID	Oil 1 (50)		Oil 2 (50)		Ratio
	Name	Weight [g]	Name	Weight [g]	
46	OL2	2.71	LS	0.31	90 : 10
47	OL2	2.39	LS	0.6	80 : 20
48	OL2	2.1	LS	0.9	70 : 30
49	OL2	1.8	LS	1.2	60 : 40
50	OL2	1.5	LS	1.49	50 : 50
51	OL2	1.19	LS	1.79	40 : 60
52	OL2	0.9	LS	2.11	30 : 70
53	OL2	0.6	LS	2.4	20 : 80
54	OL2	0.3	LS	2.7	10 : 90

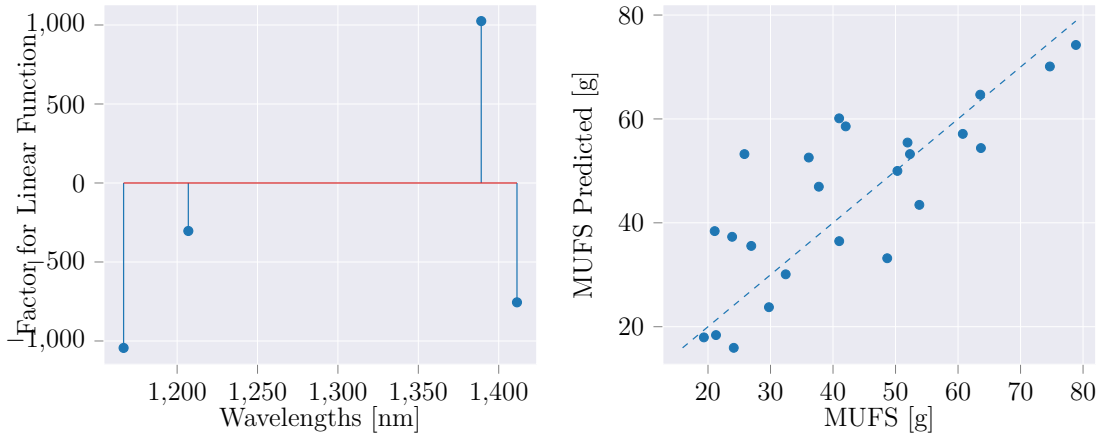


Figure A.5: MLR for Monounsaturated Fatty Acids (MUFA) with $RMSE$ = 11.076g, $R2 = 0.578$

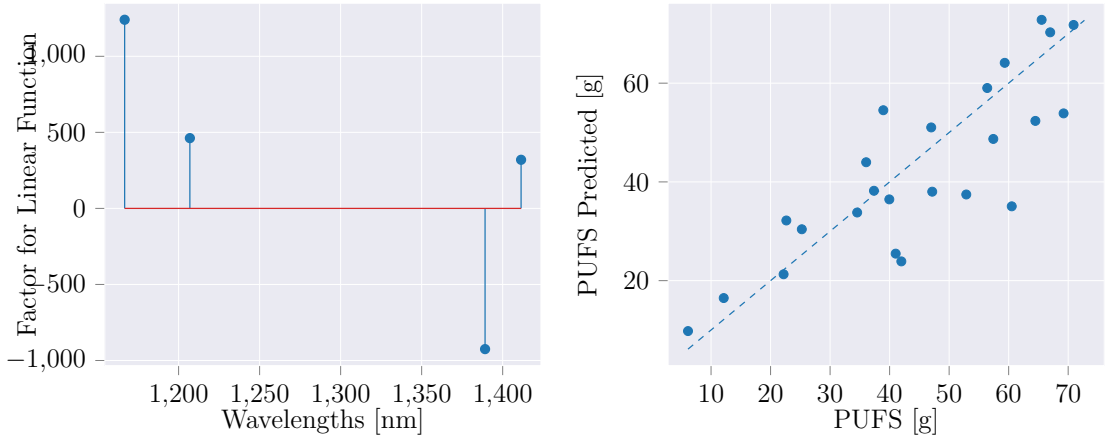


Figure A.6: MLR for Polyunsaturated Fatty Acids (PUFA) with $RMSE = 10.363$g, $R2 = 0.664$

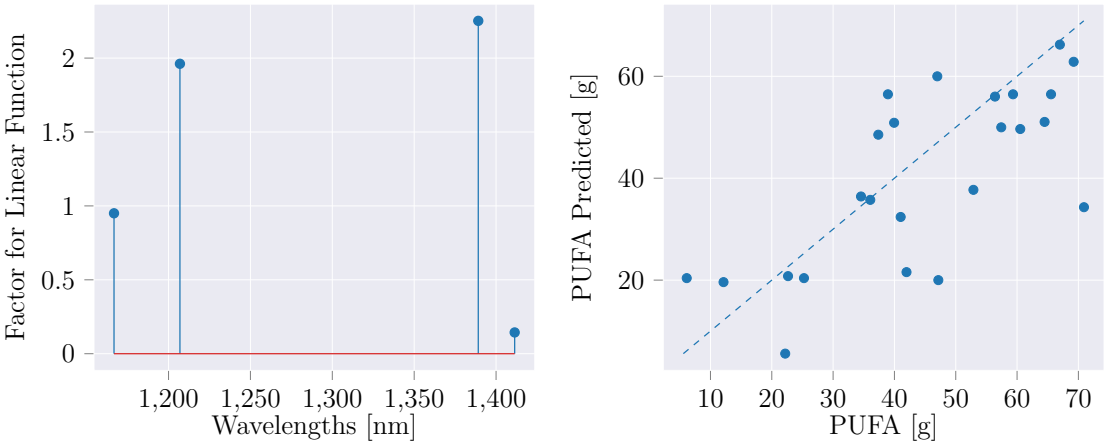


Figure A.7: MLR for Polyunsaturated Fatty Acids (PUFA) using detected scale s with $RMSE = 13.800$g, $R2 = 0.404$

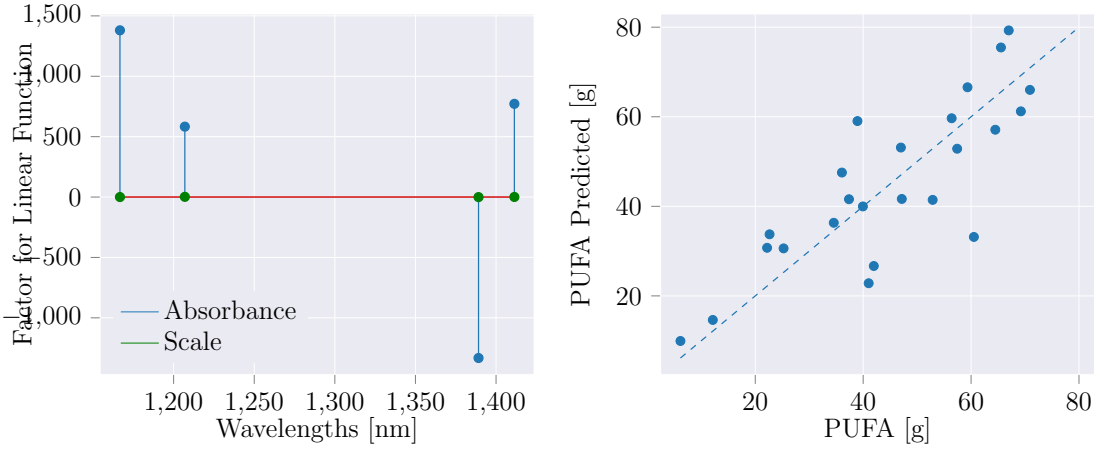


Figure A.8: MLR for Polyunsaturated Fatty Acids (PUFA) using detected scale s and absorbance with $RMSE = 10.785$g, $R2 = 0.636$

Bibliography

[1] Armenta, S., S. Garrigues and M. de la Guardia: *Determination of edible oil parameters by near infrared spectrometry.* Analytica Chimica Acta, 596(2):330–337, 2007.

[2] Armenta, S., J. Moros, S. Garrigues and M. de la Guardia: *The Use of Near-Infrared Spectrometry in the Olive Oil Industry.* Critical Reviews in Food Science and Nutrition, 50(6):567–582, 2010.

[3] Beyerer, Jürgen, Fernando Puente León and Christian Frese: *Machine Vision - Automated Visual Inspection: Theory, Practice and Applications.* 2016.

[4] Chandola, Varun, Arindam Banerjee and Vipin Kumar: *Anomaly detection.* ACM Computing Surveys, 41(3):1–58, jul 2009.

[5] Dekker, A. J. den and A. van den Bos: *Resolution: a survey.* Journal of the Optical Society of America A, 14(3):547, mar 1997.

[6] Demtröder, Wolfgang: *Experimentalphysik 3.* 2016.

[7] ElMasry, Gamal M. and Shigeki Nakauchi: *Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality - A comprehensive review.* Biosystems Engineering, 142:53–82, 2016.

[8] European Commission: *Commission Regulation (EEC) No. 2568/91 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis.* Official Journal L 248, 34(5 September):1–83, 1991.

[9] Fischer, Gerd: *Lineare Algebra - Eine Einführung für Studienanfänger.* 17 edition, 2010.

[10] Food and Agriculture Organization of the United Nations: *FAOSTAT DB*, 2017 (accessed May 5, 2018).

[11] Gabor, Dennis: *Theory of communication. Part 1: The analysis of information.* Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of, 93(26):429–441, 1946.

[12] Golub, G H and C F V Loan: *Matrix Computations.* The Johns Hopkins University Press. Baltimore and London., 3 edition, 1996.

[13] Gómez-Rico, Aurora, M. Desamparados Salvador, Marta La Greca and Giuseppe Fregapane: *Phenolic and Volatile Compounds of Extra Virgin Olive Oil (Olea europaea L. Cv. Cornicabra) with Regard to Fruit Ripening and Irrigation Management.* Journal of Agricultural and Food Chemistry, 54(19):7130–7136, sep 2006.

[14] Han, Jiawei, Micheline Kamber and Jian Pei: *Data Mining - Concepts & Techniques.* In Han, Jiawei, Micheline Kamber and Jian Pei (editors): *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Boston, Third Edit edition, 2012.

[15] Harris, F.J.: *On the use of windows for harmonic analysis with the discrete Fourier transform.* Proceedings of the IEEE, 66(1):51–83, 1978.

[16] Hastie, Trevor, Robert Tibshirani and Jerome Friedman: *The Elements of Statistical Learning*, volume 1. 2009.

[17] Heseker, Helmut and Beate Heseker: *Die Nährwerttabelle 2018/2019.* Neuer Umschau Buchverlag, 5 edition, 2017.

[18] Hunter, John D.: *Matplotlib: A 2D Graphics Environment.* Computing in Science & Engineering, 9(3):90–95, 2007.

[19] Jones, Eric, Travis Oliphant, Pearu Peterson et al.: *SciPy: Open source scientific tools for Python*, 2001–. [Online; accessed <today>].

[20] Kilbourne, Edwin M, G Rigau-Perez Josf, Clark W Heath, Matthew M Zack, Henry Falk, Manuel Martin-Marcos and Ana de Carlos: *Clinical Epidemiology of Toxic-Oil Syndrome.* New England Journal of Medicine, 309(23):1408–1414, 1983.

[21] Kim, M S, A M Lefcourt, K Chao, Y R Chen, I Kim and D E Chan: *Multispectral Detection of Fecal Contamination on Apples based on Hyperspectral Imagery: Part I. Application of Visible and Near-Infrared Refrectance Imaging.* Transactions of the ASAE, 45(6):2027–2037, 2002.

[22] Kokaly, Raymond F, Roger N Clark, Gregg A Swayze, K Eric Livo, Todd M Hoefen, Neil C Pearson, Richard A Wise, William M Benzel, Heather A Lowers, Rhonda L Driscoll and Anna J Klein: *USGS Spectral Library Version 7.* Technical Report, Reston, VA, 2017.

[23] Krause, Julius: *Identification of one or multiple spectral features in a spectrum of a sample for ingredient analysis*, 2017.

[24] Manolakis, Dmitrios, R. Lockwood, T. Cooley and J. Jacobson: *Is there a best Hyperspectral detection algorithm?* SPIE, Proceedings of, pages 733402–16, 2009.

[25] McKinney, Wes: *Data Structures for Statistical Computing in Python.* In Walt, Stéfan van der and Jarrod Millman (editors): *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[26] OceanOptics: *DH - 2000 Deuterium - Halogen Light Source Installation and Operation Manual*, 2009.

[27] OceanOptics: *NIRQuest 512-1.7 Product Sheet*, 2017.

[28] Oliphant, Travis E.: *Guide to NumPy*. Methods, 1:378, 2010.

[29] Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot and E Duchesnay: *Scikit-learn: Machine Learning in {P}ython*. Journal of Machine Learning Research, 12:2825–2830, 2011.

[30] Reed, Irving S. and Xiaoli Yu: *Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(10):1760–1770, 1990.

[31] Rinnan, Åsmund, Frans van den Berg and Søren Balling Engelsen: *Review of the most common pre-processing techniques for near-infrared spectra*, 2009.

[32] Stiftung Warentest: *Von wegen Güteklasse*. Stiftung Warentest, (2):18–28, 2016.

[33] Stiftung Warentest: *Der Extra-Bluff*. Stiftung Warentest, (2):10–19, 2017.

[34] Tay, A., R. K. Singh, S. S. Krishnan and J. P. Gore: *Authentication of olive oil adulterated with vegetable oils using Fourier transform infrared spectroscopy*. LWT - Food Science and Technology, 35(1):99–103, 2002.

[35] THORLABS: *CVH100(/M) Operation Manual 2014*, 2018.

[36] THORLABS: *UV Fused Quartz Cuvettes*, 2018.

[37] Walt, Stéfan van der, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart and Tony Yu: *scikit-image: image processing in Python*. PeerJ, 2:e453, 2014.

[38] Wang, Avery Li-Chun: *An Industrial Strength Audio Search Algorithm*. Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR 203), Baltimore, Maryland (USA), 26-30 October 2003, pages 7–13, 2003.

[39] Willett, W C, F Sacks, A Trichopoulou, G Drescher, A Ferro-Luzzi, E Helsing and D Trichopoulos: *Mediterranean diet pyramid: a cultural model for healthy eating*. The American Journal of Clinical Nutrition, 61(6):1402S–1406S, 1995.

[40] Woodcock, Tony, Gerard Downey and Colm P. O'Donnell: *Confirmation of declared provenance of European extra virgin olive oil samples by NIR spectroscopy*. Journal of Agricultural and Food Chemistry, 56(23):11520–11525, 2008.

[41] Workman, Jerry Jr. and Lois Weyer: *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy*, volume 13. 2012.