Wavelet Based Feature Extraction in Near Infrared Spectra for Compositional Analysis of Food

Julius Krause

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
julius.krause@kit.edu

Technical Report IES-2017-02

Abstract: Near infrared spectroscopy is a common method for analysis of food, soil and pharmaceutical products. New developments in sensor technology, like hyperspectral camera systems and mobile spectrometers, allow broad applications of spectroscopy with devices out of specialized laboratories. Therefore, it is necessary to develop robust algorithms for classification and regression, regardless of the device. The key to robust analysis lies in data preparation to get standardized spectral information from each device. Wavelet based feature extraction could be a possible method to compress spectral data to its material specific absorption information. A method for wavelet based feature extraction, which also reduces the influence from elastic scattering effects is proposed in this report.

1 Introduction

In order to ensure the high standards of food quality, monitoring measurements are required throughout the entire production process right up to the customer. Optical spectroscopy in the visible and near-infrared spectrum can be used as a non-destructive and non-contact measuring method on foods for quality determination. Compared to laboratory tests, the result of an optical measurement is immediately available [LGGFR17].

In the future, the development of compact and cost-effective sensor technology will facilitate the dissemination of spectroscopy. Due to advancing developments

in microsystems technology, it has been possible to integrate different measurement methods like tunable Fabry-Perot filter, fourier-transform or scanning grating systems into miniaturized sensors. A series production at prices of a few US dollars has already been announced. A ”food scanner” is just one possible application. The integration of these sensors in the Internet of Things (IoT) or a smartphone is also possible and opens up a variety of other applications in the field of quality and process control [RDC17, DWKR16].

The comparability of spectroscopic data across different devices is one of the great challenges in spectroscopy. It is gaining in importance as networking of spectral sensors grows. In the history of spectroscopy, many approaches to spectral preprocessing [RvdBE09] and transfer of models [FWT+02] have been developed. Wavelet transformation has also been used for pre-processing [MNHG96].

The approach presented below attempts to extract physical features from spectral data, which only represent the absorption by molecule vibrations or electron excitations. Therefore, a wavelet transformation of the spectral data with an approximation function like Gaussian or Lorentzian shape is used to get these features in connection with derivative pre-processing.

2 Physical model of the interaction between light and matter

The method presented in this article is based on the idea of describing the measurement signal by a physically motivated model. The analysis, based on physical model parameters, provides a level of abstraction in which specific disturbing influences can be specifically suppressed. At the same time, sensor-independent chemometric modeling is possible. The following section summarizes the physical factors that the author considers relevant to the theoretical signal model.

Interaction with matter leads to an extinction Q of light intensity. The extinction process depends on the wavelength and contains an overlay of different signals from different origin. In the following, the name spectral signature is used as an umbrella term for the raw signal, which is an overlay signal of chemical and

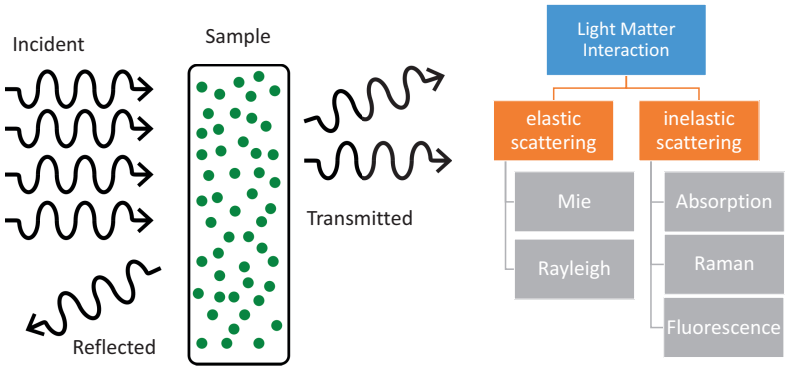


Figure 2.1: The incident light interacts with a sample. By elastic scattering within the sample, the photons are deflected by an angle θ. The most common interaction processes of the photons with the sample are shown in a block diagram.

physical properties as well as the environment. In the following, influencing variables are exemplified and summarized in three categories.

- **Chemical properties** of the sample, which are determined by absorption and fluorescence effects, which are related to specific excitations of electron states and molecular vibrations and thus produce a material-specific spectral signature.
- **Physical properties**, which depend, for example, on the shape of the sample, a surface condition of the sample or, in the case of a sample in powder form, on a degree of grinding of the powder and other properties. There is also an influence of the geometry between the measurement object and the sensor or the light sources and the sensor. This influence changes the measured spectral signature and thus also makes the comparability of different measuring devices more difficult.
- **Sensor properties**, which are caused for example by different sensitivity or different spectral measuring ranges, complicate the comparability of measurement results of different measuring devices.

It is reasonable to assume that the three categories are independent. The scattering theory will be used, to describe these processes in detail.

The occurrence of fluorescence effects should not initially be considered in the following model. However, the analysis model shown below can be extended at any time by fluorescence effects. Inelastic scattering signals from Stokes and Antistokes processes, also known as the Raman effect, are neglected due to the signal strength from 10^{-6} to 10^{-9} compared to the output signal.

The following model summarizes the influencing variables of the measurement signal: An optical sensor detects the light emitted by a sample. The measured reflection or transmission signal is considered with respect to the light emitted from the light source. Only a part of the light emitted by the sample can be detected in the solid angle $\Delta\Omega_{\text{sensor}}$ of the detector. This extinction Q_{ext} is composed of the scattering of the incident light Q_{sca} into the solid angle Ω not detected by the sensor (Fig. 2.1). The absorption process by the electrons and the molecular states is considered independent of the scatter and added as an additional term Q_{abs}:

$$Q_{\text{ext}}(\lambda, \theta, r) \coloneqq Q_{\text{sca}}\left(\frac{r}{\lambda}, \theta\right) + Q_{\text{abs}}(\lambda)\,.$$

The model includes the particle size or micro structure with a radius r, the angle θ between the light source and sensor, and the wavelength λ dependency. The two terms of the equation are described in detail below.

2.1 Elastic Scattering Theory

Many optical systems can be well described by geometric optics. In cases where the object radius r is in or below wavelength ranges, the phenomena occurring can be well described by the scattering theory of Rayleigh and Mie [CDL02].

The Rayleigh theory can be applied to describe the light scattering by particles with radius $r < 1/10\lambda$. In this regime, the particle act as an oscillating dipole driven by the electromagnetic field. A microscopic dipole absorbs a photon in a virutal state, and the subsequent emission has the characteristic of a dipole antenna. The intensity distribution

$$I(\theta, \lambda) \propto 1/\lambda^4(1 + \cos^2\theta)$$

of the scattered light results from the probability of the individual scattering angles.

In particles and structures whose dimension correspond to the wavelength λ, plasmon resonances can be excited. A complete analytic solution of the Maxwell equations exists only for spherical objects and is described in detail in the Mie theory [Mie08]. The Mie theory can also be used for particle size determination by laser diffraction and, in particular for small objects, provides a better result than Fraunhofer diffraction. For the following analysis, however, it is sufficient to know that the solution of the Mie theory is given by the so-called Bessel functions, which are smooth and differentiable.

In summary, the elastic scattering gives a smooth and differentiable low-frequency signal contribution in the detected spectral signature.

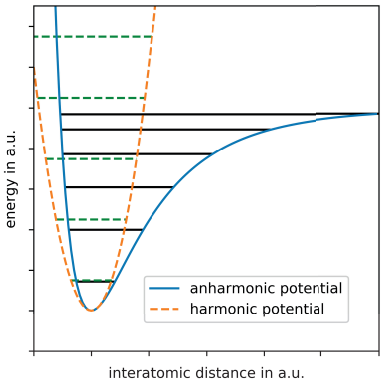
2.2 Absorption

The model of a harmonic oscillator (Lorentz oscillator) can approximately describe the absorption by molecular vibrations. Wherein the solution for determining refractive indices and absorption coefficients is reproduced substantially correctly. The dependence of wavelength of the absorption is given by the so-called dielectric function which be approximated by a Lorentz profile. In addition, the Lorentz profile corresponds to the so-called natural linewidth, which e.g. can be derived from the Fourier transform of a damped harmonic oscillator.

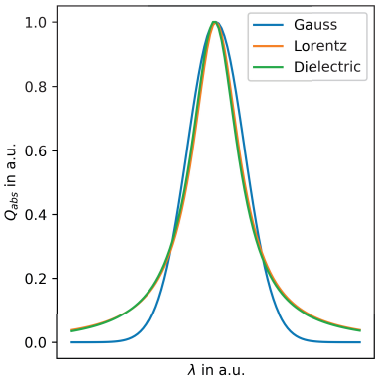
Regardless of which model is used, the exact absorption spectrum can not be calculated as long as the individual coefficients of the electric field distribution in the solid state, the anharmonicity of the molecular vibration, and the interaction with neighboring molecules are unknown. However, the course of the dependence of wavelength of a single absorbance (or emission by fluorescence) is represented approximately correctly by a bell shaped curve such as Lorentz profile or Gaussian function [Dem10].

Following the preceding qualitative analysis of the dependence of wavelength of the absorption, the relationship between an amount of substance and its absorption is now to be determined in a simple model. From the exponential attenuation of a light beam after entering a medium, the Beer-Lambert law can be derived:

$$\ln\left(\frac{I_1}{I_0}\right) = -c\eta\delta\,.$$



(a) The Lennard-Jones potential describes the interaction potential of a diatomic molecule and can be described in quadratic approximation by a harmonic oscillator. With increasing energy above the ground state, anharmonic corrections must be considered.



(b) Clearly recognizable is the good approximation of the dielectric function ϵ by a Lorentz profile. A Gaussian profile differs more, but this is justifiable because absorption lines are usually extended by disturbing effects.

Figure 2.2: Absorption on the model of diatomic molecular vibrations.

The extinction $Q_{\text{abs}} = I_1/I_0$ causes the scaling of the bell shaped absorption curve. A quantitative content determination appears possible due to the connection to the substance-dependent attenuation factor η, the concentration c, and the optical path length δ.

In summary, a single absorption can be described approximately by three parameters of a bell curve like a Gaussian.

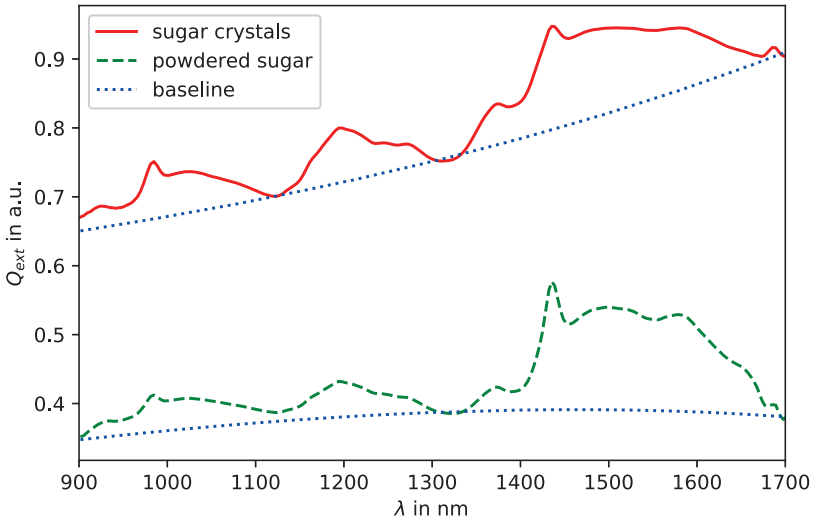


Figure 2.3: The extinction Q_{ext} of the chemically identical substances differs in scale and in the baseline due to different scattering properties and due to different particle size.

3 Wavelet Based Feature Extraction

The extinction Q_{ext} of the measurement signal in Figure 2.3 is compared to a white reflection standard. Both spectra describe the identical substance which is sugar and only the particle size differs. It can clearly be seen that the figure shows differences in the baseline of the two spectra and the signal strength. Therefore, pre-processing is needed for chemical component analysis.

The most common methods for spectral pre-treatment like scatter correction, normalization, and dimensional reduction are based on the spectral signature of a single measurement system as a whole, e.g. by inclusion of the mean signal. In addition, linear operators are destroying the connection to Beer-Lambert law. Therefore, analysis models based on these methods can not readily be used in another measurement environment.

Another established possibility for the correction of multiplicative influences is the gradient formation over the spectrum. However, the noise is amplified and human interpretation is difficult. The wavelet analysis based on the derived spectrum is intended to counteract these two disadvantages.

The wavelet analysis includes the neighbourhood information in the spectrum which counteracts the noise. The easily interpretable parameters of the approximated bell curve can be taken from the wavelet scalogram afterwards. In addition, the mean value of the wavelet transformation corrects another term of the baseline. Moreover, a later normalization based only on the absorption bands used in the model is more robust to changes in the spectral signature.

The algorithm is based on two assumptions that were explained previously:

- The baseline of the spectrum is due to the anisotropy of the elastic scattering and can be approximated by a smooth polynomial function.
- The absorption can be approximated by a bell-shaped absorption function with three parameters.

For the mathematical description, the spectrum is referred to as a function $g(\lambda)$ and the continuously differentiable bell-shaped approximation function is referred to as $\psi_{\lambda_0,s}(\lambda)$. Where λ_0 is the center and s is the width of the approximation function.

Step 1: Baseline Correction and Peak Deconvolution

The n-fold derivative of the spectrum reduces the polynomial order of the baseline, at the same time superposed peaks are unfolded [NW84]. Noise is greatly amplified by the derivative, which is why smoothing according to Savitzky-Golay is used in many cases [SG64].

In the following the fact is used that the derivative operator can also be applied to the function ψ. In the case of the second derivative, one obtains the *Mexican Hat* function, which is widely used in signal processing.

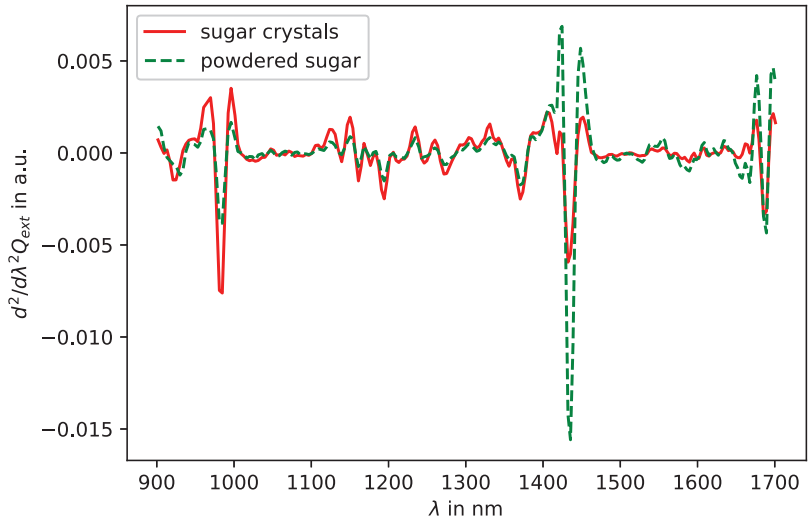


Figure 3.1: By applying the second derivative, the baseline of the spectrum was eliminated. Characteristic features are clearly shown in the shape of an inverted Mexican Hat with identical width and position in both spectra.

Step 2: Wavelet Transformation

The wavelet transformation has the character of a correlation analysis [Mal89, Mor83]. The description of the wavelet transformation as

$$\Gamma_\psi(\lambda_0, s) \coloneqq \langle \psi_{\lambda_0,s}(\lambda), g(\lambda) \rangle$$

shows this fact.

The scalar product is performed for different values of λ_0 and s, the resulting wavelet coefficients from the example of sugar is shown in a scalogram (Fig. 3.2).

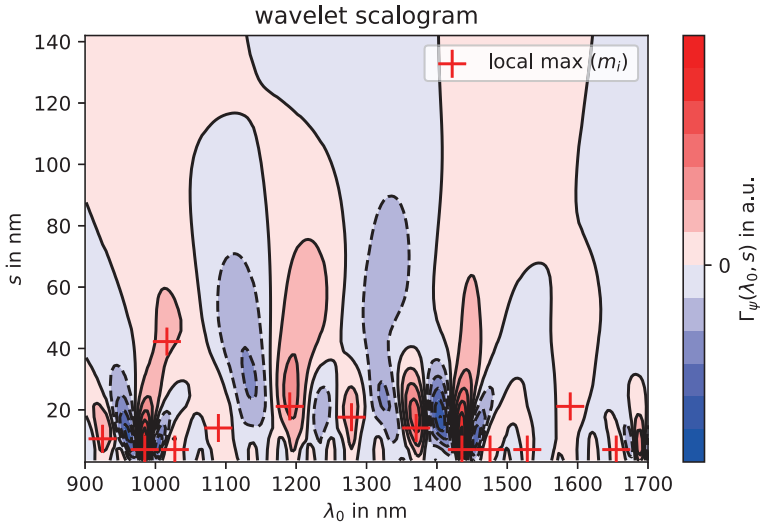


Figure 3.2: For the example of sugar, the wavelet coefficient is plotted for a selected region of s_i and λ_i. Where of the red (blue) color indicates positive (negative) energy of the wavelet coefficient. Local maxima are highlighted by red crosses.

Step 3: Feature Extraction

Local maxima of the wavelet coefficient show the location of the best match between the correlation function and the spectrum. From the coordinates of the local maxima, the parameters of each feature

$$m_i = (\lambda_i, s_i, \Gamma_\psi(\lambda_i, s_i))$$

can be found. Wherein the wavelet coefficient also indicates the height or the strength of the peak, which is linked via the Beer-Lambert law with the amount of existing ingredients. Although the amount $i \in \mathbb{N}^+$ of found features is initially not limited.

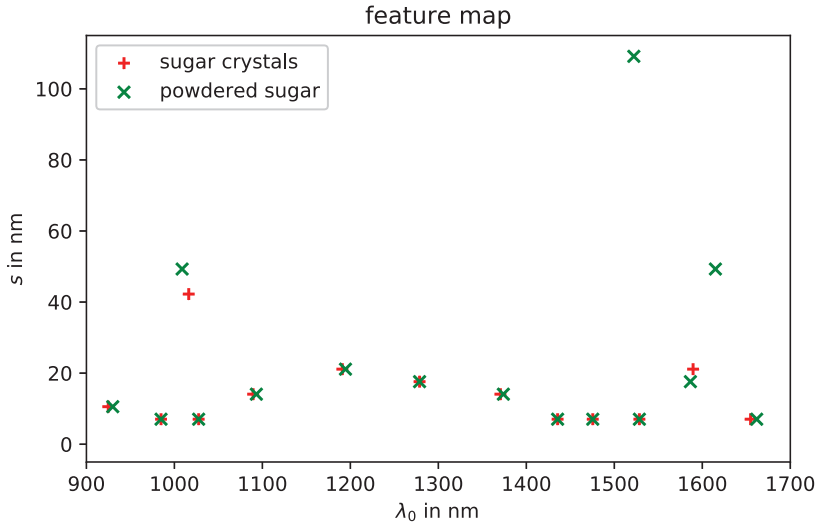


Figure 4.1: In many cases, the algorithm has extracted the identical values even for superimposed peaks.

4 Experimental Results

Using different sugar granules as an example, it could be shown that, despite different scattering properties, nearly identical features are found for the individual features (Fig. 4.1). This confirms the invariance of the found features against changes in the scattering properties.

Another example is to show that the features contain additional information of individual ingredients. For this purpose, the data set of a competition for the determination of protein in cereals by NIR spectroscopy was selected. The training dataset comprises 1488 spectra from 248 different samples, measured with 6 spectrometers, three spectrometers of the same model from two different manufacturers.

Figure 4.2 shows an example of two spectra from different Instruments. For comparison, a virtual absorption spectrum was formed from the previously extracted

features m_i. The relevant feature for the protein concentration also determined in the competition literature is the small peak at $1016\ nm$. The information about the protein content is thus in a small hidden peak [IAB+17].

The determined features m_i of all 6 spectrometers for all 248 samples are shown in a three-dimensional feature map in a section around the relevant protein peak (Fig. 4.3). The wavelet coefficient is normalised by referencing to a protein independent peak. The presented method found the relevant peak in all 1488 spectra. The figure also clearly shows, that the width s_i and wavelet coefficient $\Gamma_\psi(\lambda_i, s_i)$ of the protein peak increases with the protein content.

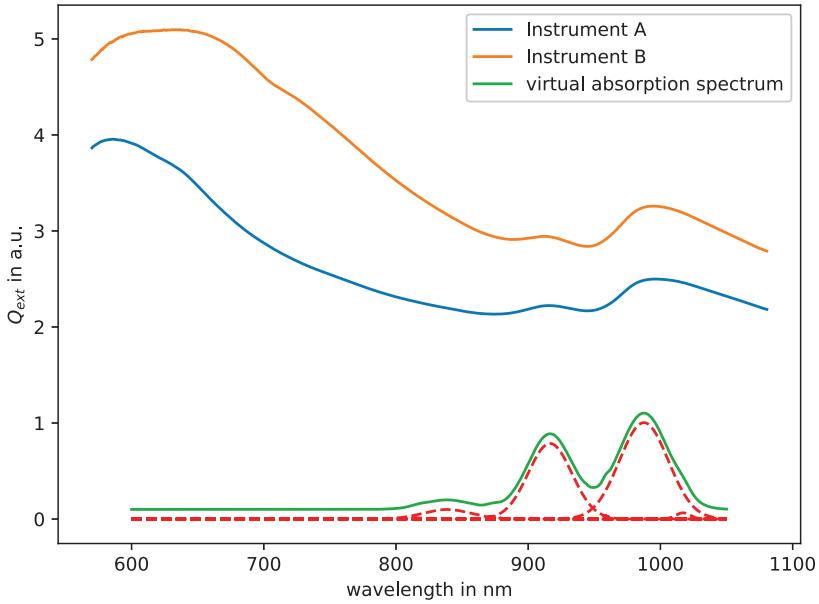


Figure 4.2: The spectra of an identical sample recorded by two spectrometers models from different manufacturers A and B. In the lower part is shown as an example a virtual absorption spectrum, which was determined from the previously extracted features m_i

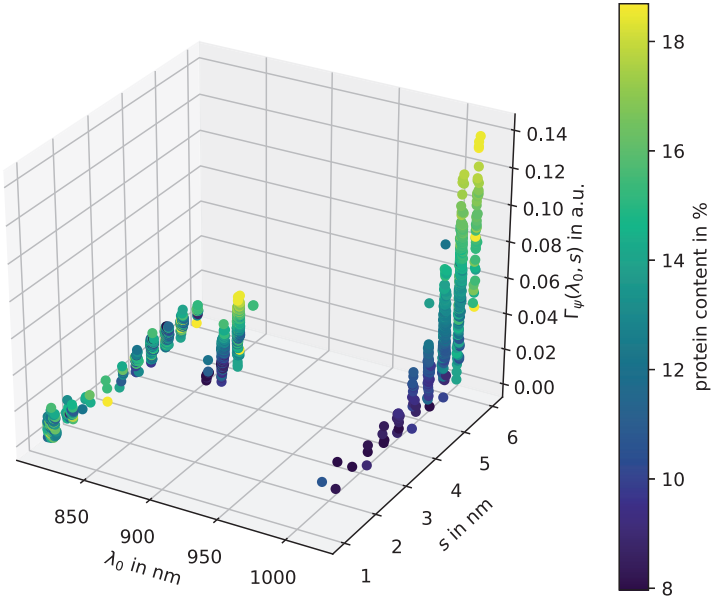


Figure 4.3: The three-dimensional feature map shows the parameters of the features extracted from the training set. The protein content of the sample is shown in color. For better visibility of the protein content, the presentation is limited to three small features.

5 Summary

The presented method is suitable for the feature extraction of superimposed absorption signals. The position and width of a peak are independent of the scaling of the signal strength and are therefore suitable for a robust material identification, regardless of the sample geometry and the measuring device used. The information of the signal intensity is included in the wavelet coefficient and offers the possibility to quantify an ingredient. The representation of the spectral features m_i as a list of triplets $(\lambda_i, s_i, \Gamma_\psi(\lambda_i, s_i)$ can be created for measurements of sensors of different types. Classification and regression models based on evaluation of the triplets are thus invariant with respect to the sensor used.

The knowledge of the position and width of the absorption features also allows further evaluations. Optical filters can be selected based on the individual features. In addition, components of the elastic scattering parameters can be determined from the residuum of the spectral signature after deduction of the absorption properties. In hyperspectral imaging, feature extraction can be used for compression.

Bibliography

[CDL02] A. J. Cox, Alan J. DeWeerd, and Jennifer Linden. An experiment to measure Mie and Rayleigh total scattering cross sections. *Am. J. Phys.*, 70(6):620–625, 2002.

[Dem10] Wolfgang Demtröder. *Expirmentalphysik 3, Atome, Moleküle und Festkörper*. Springer-Lehrbuch. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[DWKR16] Anshuman J. Das, Akshat Wahi, Ishan Kothari, and Ramesh Raskar. Ultra-portable, wireless smartphone spectrometer for rapid, non-destructive testing of fruit ripeness. *Sci. Rep.*, 6(September):1–8, 2016.

[FWT+02] Robert N. Feudale, Nathaniel A. Woody, Huwei Tan, Anthony J. Myles, Steven D. Brown, Joan Ferre, and Joan Ferré. Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst.*, 64(2):181–12, 2002.

[IAB+17] Benoit Igne, Md Anik Alam, Dongsheng Bu, Pierre Dardenne, Hanzhou Feng, Ali Gahkani, David W Hopkins, Shikhar Mohan, Charles R Hurburgh, and Cathleen Brenner. Summary of the 2016 IDRC software shoot-out. *NIR news*, 28(4):16–22, 2017.

[LGGFR17] Mercedes G. López, Ana Sarahí García-González, and Elena Franco-Robles. Carbohydrate analysis by NIRS-chemometrics. In *Dev. Near-Infrared Spectrosc.* InTech, mar 2017.

[Mal89] Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[Mie08] Gustav Mie. Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen. *Ann. Phys.*, 330(3):377–445, jan 1908.

[MNHG96] C. R. Mittermayr, S. G. Nikolov, H. Hutter, and M. Grasserbauer. Wavelet denoising of Gaussian peaks: A comparative study. *Chemom. Intell. Lab. Syst.*, 34(2):187–202, 1996.

[Mor83] J. Morlet. Sampling theory and wave propagation. In C. H. Chen, editor, *Issues in Acoustic Signal — Image Processing and Recognition*, pages 233–261, Berlin, Heidelberg, 1983. Springer Berlin Heidelberg.

[NW84] KH Norris and PC Williams. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size., 1984.

[RDC17] Giovanni Rateni, Paolo Dario, and Filippo Cavallo. Smartphone-based food diagnostic technologies: A review. *Sensors (Switzerland)*, 17(6), 2017.

[RvdBE09] Asmund Rinnan, Frans van den Berg, and Soren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201 – 1222, 2009.

[SG64] Abraham Savitzky and Marcel J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.