Simulation-Based Evaluation of Wavelet Coefficients for Robust Analysis of Near Infrared Spectra

Julius Krause

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

Technical Report IES-2018-08

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. Wavelet coefficients are a promising approach for the detection and estimation of spectral absorption bands. The robustness of wavelet based features against typical measuring influences and calibration errors will be analyzed in the following by using simulations.

1 Introduction

In near infrared spectra, information from the chemical and physical properties of a sample are superimposed. Therefore, methods of multivariate data analysis and machine learning are needed to link the optically measured information with the properties of the sample. In this area, computer science overlaps with chemometrics, a sub-discipline of chemistry for the evaluation of statistical data. For this reason, there is already a large number of methods for processing spectral data. However, it turns that some classical data evaluation methods are difficult to apply to current tasks, such as networks of spectral sensors. Models for the evaluation of near infrared spectra can often only be used for single sensors calibrated by complex reference analyses [Wor18, FWT$^+$02].

One of the most serious calibration errors of spectrometers is the wavelength mapping of the individual bands, e.g. due to aging processes, temperature fluctuations or tolerances in the production process. This error has a massive influence on the model accuracy. The transferability of models between sensors and the long-term stability of the models is therefore limited.

Therefore, the wavelet transformation for the detection of absorption bands has already been presented as a promising method [BPT18]. This method explicitly determines the absorption bands, more precisely the chemical parameters of the sample. This results in feature vectors which are partly invariant to multiplicative influences such as the measuring distance and particle size.

In the following, the influence of wavelength calibration and noise errors will be investigated in more detail. A classical data preprocessing method by smoothing with Savitzky-Golay filters, derivatives and normalization with *Standard Normal Variate* (SNV) is compared with the wavelet coefficient approach [RvdBE09a].

2 Signal model of the near infrared reflection

The determination of chemical parameters, such as the concentration of individual components, is based on the properties of certain absorption bands. In many examples, bandpass filters can be used for evaluation [LBF+15a, Lu04a]. However, the absorption can only be determined indirectly, because only the light reflected from the sample reaches the sensor. Scattering effects also applies to transmission measurements, which are therefore not considered separately.

First, the signal input at the spectrometer is to be described. The number of photons hitting the sensor is determined by

$$g(\lambda) := N_{\text{sensor}}(\lambda) = (1 - P_{\text{abs}}(\lambda)) \cdot P_{\text{sca}}(\lambda) \cdot N_{\text{source}}(\lambda),$$

whereby the scattering probability is determined by integrating the radiance in the solid angle of the sensor.

The so-called *Bidirectional Reflectance Distribution Function* (BRDF) [BFP12] provides a formal description of the reflection behavior. In a spectral measurement, however, the necessary angles and the BRDF of the object are usually unknown.

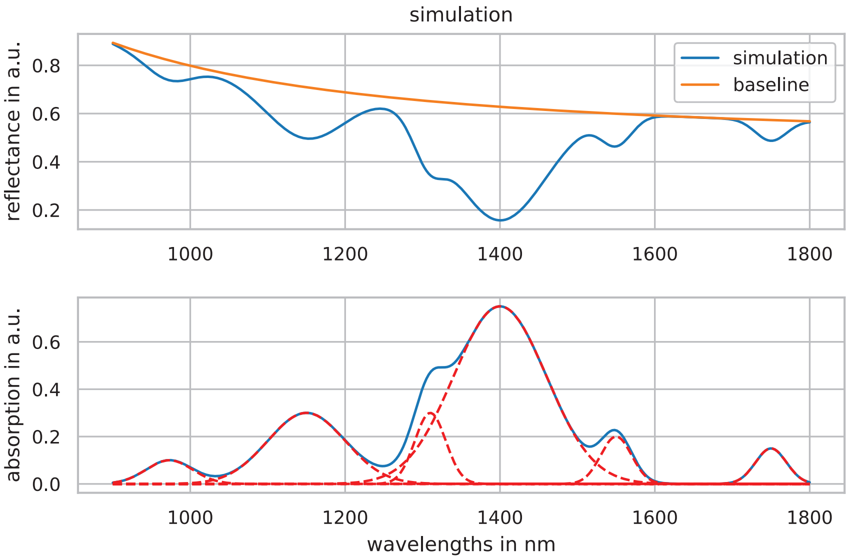

Figure 2.1: Near infrared spectra are formed from a sum of different absorptions by polar hydrogen bonds. The absorptions can be determined indirectly from the reflectance or transmission of a sample. But this signal additionally contains a wavelength dependent baseline by physical effects like Mie and Rayleigh scattering.

A theoretical description of the reflection behavior allows in parts the physical scattering theory according to Mie and Rayleigh. The result is a wavelength-dependent scattering probability that a photon is scattered into the detection range of the sensor. In order to simplify the complex scattering process, it is assumed in the following that the scattering probability is a smooth and continuously differentiable function. For this reason, the derivation in spectral direction is an established method for reducing the influence of particle size in a spectral measurement [NW84a].

The absorption probability results from

$$P_{\text{abs}}(\lambda) = \sum_{j}^{N} c_j \cdot \exp\left\{ -\left(\frac{\lambda - \lambda_j}{\sqrt{2}\sigma_j} \right)^2 \right\},$$

whereby the relationship between the expression c_j of the corresponding absorption bands and the substance concentration is well approximated by Beer-Lambert's law [Bro06a]. In many cases, constituents also act on several absorption bands, because the absorption bands are formed by excitation of molecular groups of different polar hydrogen bonds. Therefore, spectral analysis is an ill-posed non-linear inverse problem. A further complicating factor is that the absorption bands of the different molecule groups also overlap.

A spectrometer or a spectral camera forms a discrete measurement signal

$$k_i := (g * h_{i,\text{sensor}})(\lambda_i) + N_{i,\text{dark}}$$

from the incident photons. For a large number of spectral bands, depending on the system several hundred or more than thousand bands are recorded, the term *hyperspectral* is used.

In the ideal case there is no crosstalk of the bandpasses between single spectral channels

$$h_{i,\text{sensor}}(\lambda) \approx \delta(\lambda, \lambda_i + \Delta\lambda).$$

The wavelength assignment is done via the filter position λ_i, which can be shifted by $\Delta\lambda$ due to calibration errors.

The reflectance **R** of a sample is determined by comparison with an standard with ideal reflectivity

$$\begin{aligned} R_i &= \frac{k_i - k_{i,\text{dark}}}{k_{i,\text{reference}} - k_{i,\text{dark}}} \\ &= (1 - P_{\text{abs}}(\lambda_i + \Delta\lambda)) \cdot P_{\text{sca}}(\lambda_i + \Delta\lambda). \end{aligned}$$

This removes the dark current and the spectral curve of the light source. The wavelength-dependent sensitivity of the sensor, which has not been described in detail so far, is also compensated, because the detection probability of a photon is comparable to a changed number of photons from the source.

3 Wavelet feature extraction

The purpose of the feature extraction method presented here is to estimate the absorption bands contained in the spectrum. Wavelet transformation is a well-established method in image and signal processing, especially for edge detection [Mal89a] [Mor83]. Due to the similarity to P_{abs}, the analysis with the Mexican-Hat wavelet

$$\psi_{\lambda_0,s}(\lambda) = s^2 \frac{d^2}{d\lambda^2} \frac{1}{\sqrt{2\pi s}} \exp\left\{ - \left(\frac{\lambda - \lambda_0}{\sqrt{2}s} \right)^2 \right\}$$

allows a direct estimation of the absorption bands.

The Wavelet transformation

$$\Gamma_\psi^R(\lambda_0, s) := \langle \psi_{\lambda_0,s}(\lambda), R(\lambda) \rangle$$

is performed for a variety of possible positions λ_0 and scaling factors s. The absorption bands can then be determined from the positions (λ_i, s_i) of the local maxima within the wavelet scalogram. The evaluation of the wavelet coefficient $\Gamma_\psi^R(\lambda_i, s_i)$ also makes it possible to estimate the amount of absorption.

Assuming the scattering probability can be approximated as a smooth, continuously differentiable function, an approximation as a Taylor series is possible at the position of the absorption bands. This approximation allows a simplified description of scattering effects because the Mexican-Hat wavelet has two vanishing moments. Thus the influence of the scattering probability to the determination of the molecule concentration c_i is reduced to a localy static multiplicative factor

$$\begin{aligned} \Gamma_\psi^R(\lambda_i, s_i) &= \langle \psi_{\lambda_0,s}(\lambda), (1 - P_{\text{sca}}(\lambda)) P_{\text{sca}}(\lambda) \rangle \\ &\approx c_i P_{\text{sca}}(\lambda_i), \end{aligned}$$

which can be removed e.g. by quotient formation with neighbouring absorption bands. However, the position and width of the estimated absorptions are not influenced by this multiplicative factor.

4 Impact of noise and wavelength shift on prediction accuracy

Based on the signal model presented at the beginning (see fig 2.1), spectral data with different concentrations of a fictive ingredient were generated. The fictive ingredient forces linear changes of the absorption band at 1310 nm. In addition this absorption band is superimposed by a larger absorption at 1400 nm. This problem is exemplary e.g. for the determination of acid concentration by OH absorption near the water band. In addition, static absorption bands of different characteristics were added at 973 nm, 1150 nm, 1550 nm and 1700 nm (see fig 2.1). In addition, the reflectance, more precisely the baseline, was varied by a random value. For comparability all models were trained with 41 spectra, each spectrum is sampled by 300 bands between 900 nm and 1800 nm. The validation was done by newly generated spectra, which differ from the original training data by a random baseline. *Partial least squares regression* (PLSR) models with 6 components were used to compare feature extraction by wavelets with a classical spectral pre-processing.

The classical pre-processing steps are:

- a smoothing of the spectrum using Savitzky-Golay filters (11,2)
- the generation of a spectral derivative to correct the baseline
- normalizing the intensities by using *Standard Normal Variate* (SNV)

The wavelet based feature extraction was performed in the following steps:

- by use of wavelet transformation the absorption bands were determined from local maxima in Γ_{ψ}^{R}
- feature vector **m** was created with $m_i = s_i \cdot \Gamma_{\psi}^{R}(\lambda_i, s_i)$
- the feature vector **m** was normalized to the variance due to the different intensities of the spectra

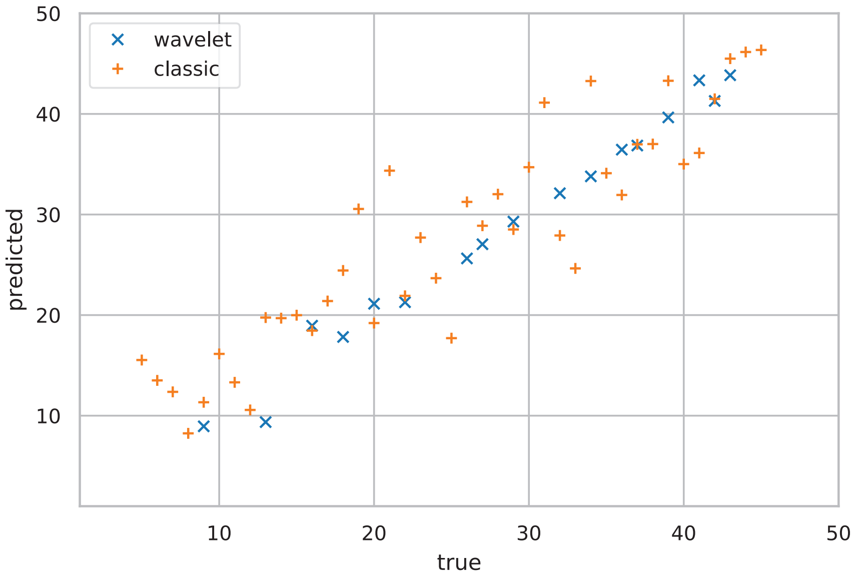

Figure 4.1: For a noise level of 0.015 of the spectral intensity, the type of prediction of both methods differs significantly. While the classical method scatters more strongly, the wavelet method, in individual cases especially at low concentrations, does not make a prediction.

4.1 Simulation of different noise levels

In real spectroscopic sensors, the measurement signal is superimposed by noise due to various processes such as discretization and amplification in signal acquisition. Therefore, the noise level of spectral bands is an important quality parameter of the sensor. A comprehensive description can be found in the EMVA1288 standard. For further evaluation, the spectral signals

$$R_i' = R_i \cdot (\text{rand [-1,1]} \cdot \text{noise level} + 1)$$

are superimposed with random noise.

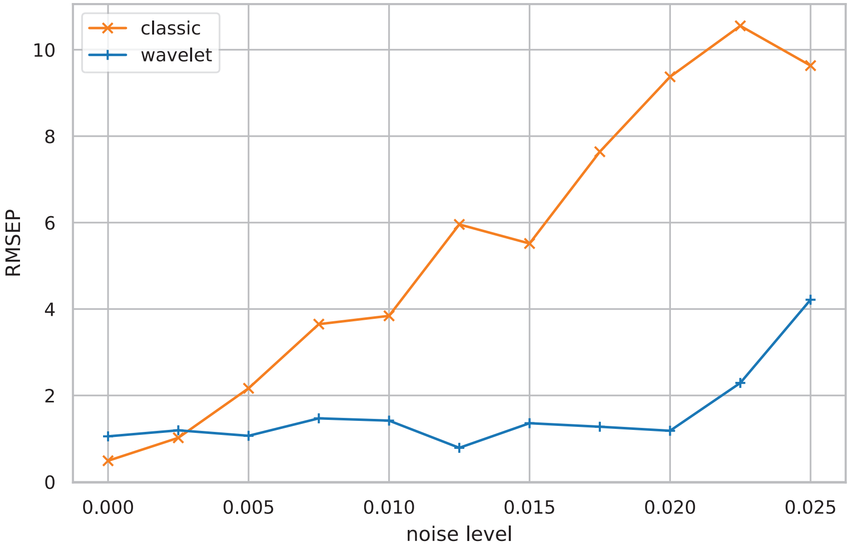

Figure 4.2: The *Root Mean Square Error of Prediction* (RMSEP) was determined for different noise levels using 41 random spectra. The wavelet method shows nearly no changes over a wide range, but this is also based on the implicit rejection of faulty spectra.

For different noise levels the *Root Mean Square Error of Prediction* (RMSEP) was determined by 41 random spectra of different concentrations 4.2. It was found that the RMSEP increased continuously with increasing noise levels when classical pre-processing was used. Using the wavelet method, the required absorption band, especially at low concentrations, was not detected with increasing noise level in some cases. Because an incorrect detection of absorption bands already leads to an error in the creation of the feature vector **m**. Therefore, the wavlet method implicitly rejects faulty data and remaining predictions lead to a better result (see fig 4.1).

4.2 Simulation of systematic wavelength shifts

As in the previous section, the two pre-processing methods were compared using systematically modified test data. The spectra were shifted step by step compared to the trained data sets.

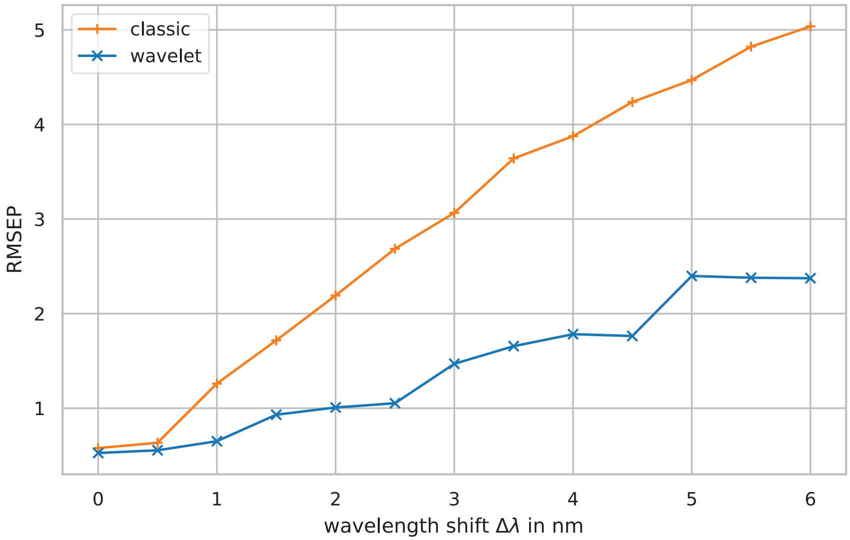

Figure 4.3: The spectra of the test dataset were shiftet systematically in wavelength compared to the training dataset. This creates a bias and therefore a continuously increasing error in prediction by the classical method. The prediction by the wavelet method changes step by step, depending on detection of the peaks within the allowed tolerance range.

Using classical preprocessing, a bias in the prediction was generated, resulting in a continuous increase of the RMSEP with increasing shift. This was to be expected in this way.

The features determined by wavelet transformation also showed a shift dependent error. However, this error is lower compared to the classical method and changes approximately stepwise. The simulation therefore evaluates the effects on the prediction error already for shifts below the spectral resolution.

5 Summary

The results of the simulations have already been expected a-priori due to the mathematical structure. The presented classical method is based on the scalar multiplication of the spectrum with a vector of coefficients, which are exactly adapted to the resolution and measurement range of the sensor. The SNV preprocessing method also uses the variance of the entire spectrum for normalization. This is the case in almost all established methods. Due to the strong adaptation to the sensor used, problems arise in the transferability of the learned models to different sensors.

The feature extraction by detection of absorption bands with a wavelet transformation uses prior knowledge about the shape of the absorption bands. This makes it possible to consider the neighborhood relationship of spectral bands. This results in a reliable detection of superimposed absorption bands and noise influences are minimized. The application of the Mexican-Hat wavelet also enables the minimization of scattering effects due to the included vanishing moments.

Especially at low concentrations, less absorption bands were detected with an increasing noise level. This correlation could be used in the future as a possibility to characterize spectral sensors. Because the signal-to-noise ratio given in the data sheets refers only to an intensity-noise-ratio of the spectral bands. Wavelets take into account the neighborhood relationship of spectral bands. Due to the limit and the uncertainty in the detection of the wavelets, a combination of resolution and noise is possible. Together with an already existing model based on wavelet features, a sensor-specific *Limt of Detection* (LoD) for certain ingredients is thinkable.

Bibliography

[BFP12] Jürgen Beyerer, Christian Frese, and Fernando Puente Leon. *Automatische Sichtprüfung : Grundlagen, Methoden und Praxis der Bildgewinnung und Bildauswertung*. Springer Vieweg, 2012.

[BPT18] Jürgen Beyerer, Alexey Pak, and Miro Taphanel, editors. *Proceedings of the 2017 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*. Karlsruher Schriften zur Anthropomatik. KIT Scientific Publishing, Karlsruhe, [2018].

[Bro06a] Adrian Jon Brown. Spectral curve fitting for automatic hyperspectral data analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6):1601–1607, 2006.

[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. *Chemometrics and Intelligent Laboratory System*, 64(2):181–12, 2002.

[LBF+15a] Magali Lafontaine, Zrinka Bockaj, Maximilian Freund, Kai Uwe Vieth, Christian Negara, and Thomas Langle. Non-destructive determination of grape berry sugar concentration using visible/near infrared imaging and possible impact on wine quality. *Technisches Messen*, 2015.

[Lu04a] Renfu Lu. Multispectral imaging for predicting firmness and soluble solids content of apple fruit. *Postharvest Biology and Technology*, 31(2):147–157, 2004.

[Mal89a] S.G. Stephane G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, jul 1989.

[Mor83] J. Morlet. Sampling theory and wave propagation. In *Issues in acoustic signal/image processing and recognition*, volume 1, pages 233–261. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[NW84a] 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.

[RvdBE09a] Åsmund Rinnan, Frans van den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222, nov 2009.

[Wor18] Jerome J. Workman. A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy, mar 2018.