Automatic signal enhancements for spectroscopic measurements

Dissertation
zur Erlangung des Grades eines

Doktors der Ingenieurwissenschaften

der Technischen Universität Dortmund
an der Fakultät für Informatik

von
Jan Schlenke

Dortmund
2013

Tag der mündlichen Prüfung: 30.10.2014
Dekan: Prof. Dr.-Ing. Gernot A. Fink
Gutachter: Prof. Dr. Heinrich Müller, Prof. Dr. Günter Rudolph, Dr. Lars Hildebrand.

Contents

Chapter 1

Aim and motivation

Material analysis is an important and diverse scientific field often requiring a high degree of expert knowledge to achieve satisfactory results. Therefore automation of process steps is essential to allow untrained users access to such methods and reduce manual workload for experts. This thesis proposes robust corrective measures that can be applied to measured data containing severe distortions with the goal to enable automatic identification of compounds in so called real world spectra, meaning spectra that were recorded under unknown and sometimes adverse conditions outside a laboratory environment. To offer an intuitive understanding of why corrective measures to real measurements are essential in highly automatic analysis systems figure 1.1 shows a number of normalized Raman spectra acquired at standoff distance. All six spectra show the same substance in different mixtures and on different background materials. The task of automatic signal enhancements is to reduce the influence of distortions to a level at which unknown measurements taken in the field under adverse conditions can be compared to reference measurements acquired under laboratory conditions in order to extract information of their contents.

Extracting information from chemical systems in order to classify or even identify their components is a critical task for almost every application that is aimed to detect and/or identify unknown or hidden compounds. Methods to achieve that goal employ spectroscopic techniques that vary in instrumentation and underlying physical principals. The analysis of the collected data is often non-trivial and interpretation requires experts with years of training to extract reliable information from measurements. The reasons for these difficulties are multiple but can be roughly categorized in two categories: instrument dependent and environmental dependent. Instrumental dependent challenges are related to discrepancies between the characteristics of the measuring instruments and the requirements necessary to observe certain chemical or physical phenomena with sufficient accuracy. Typical examples of critical constraints are the resolution available to record measurements, the sensitivity and selectivity of sensors

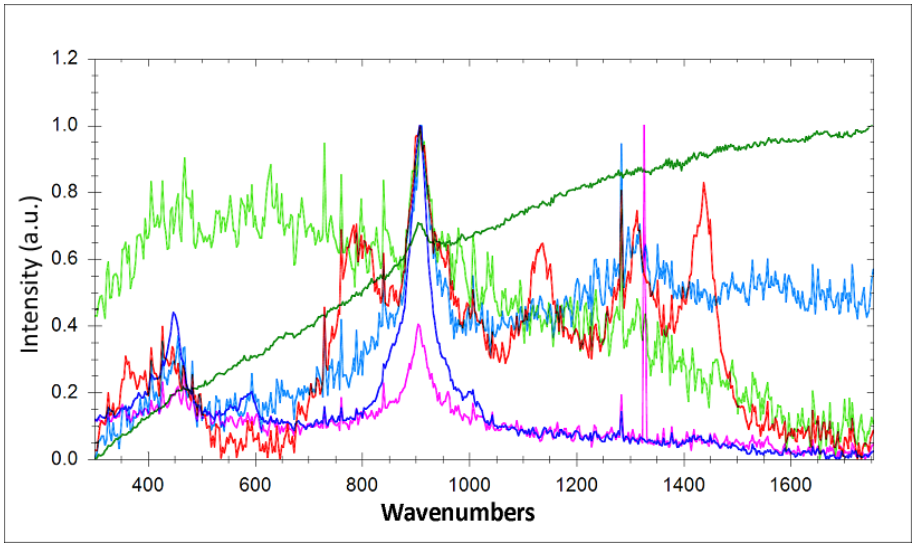

Figure 1.1: Exemplary standoff Raman spectra of an explosive substance in mixtures with different materials and backgrounds.

and timing constrains.

Environmental factors often manifest in the form of distortions that are introduced by sources which influence the measurements from the outside like e.g. humidity, air pressure, exposure to sunlight, wind or movement. As well as sources that lie within the measured compound itself like impurities or simply chemical combinations that produce a signal which is more complex than the simple sum of its parts. Instrumental limitations can sometimes be migrated by repeated measurements and the use of multivariate statistics but in general data analysis cannot create information that is not contained in a measurement and instrumental limitations have to be solved by adjusting the measuring technique, not the acquired data. Therefore it is sensible to focus on the estimation and reduction of distortions contained in real world measurement in order to make techniques more accessible and reliable in the hands of non-expert users. Reduction of distortions also plays an important role in automatic identification scenarios as identification often involves the comparison of known structures to an unknown measurement, and mathematics and computer algorithms are not yet able to match the pattern recognition power of the human brain.

The main focus of this thesis is the development of techniques that reliably reduce distortions in measurements without the need for fixed models, training sets or repeated measuring. The techniques are developed with the goal of being used for measurements that are acquired outside of controlled laboratory environments and therefore must be robust against a wide array of possible distortions without complicated control mechanisms that need to be adjusted to every measurement.

The idea for such techniques was motivated by discussions with practical physicists

and chemists who expressed their discontent with currently available methods. Available correction methods often work very well in simple cases, but return poor results in more challenging scenarios making robust behavior a primary concern. Distortions can be specific to a particular spectroscopic technique or based on known physical processes, that can be detected and identified. These types of distortions of best treated with specialized methods, however other types of distortions are common to almost all spectroscopic techniques and while causes may differ from technique to technique the displayed effects are similar and can be treated by similar correction methods.

The most prominent examples of distortions that appear in multiple techniques are noise, a random signal introducing a measure of uncertainty to all recorded signals, and baseline distortions, a non-constant background signal of unpredictable intensity that changes the values of recorded data values significantly making quantitative comparisons between measurements impossible. This thesis presents new techniques that allow the robust and reliable estimation and reduction of noise and baseline distortions in order to improve the comparability and therefore identifiably of measurements made outside of controlled laboratory environments.

This work introduces new advancements in noise suppression via wavelet transform that focus on the automatic reduction of noise of varying intensity within a single measurement and a new baseline estimation technique based on adaptive regression that is highly robust against distortions and requires only minimal information about expected signal characteristics.

Chapter 2 gives an introduction to spectroscopy and introduces several different spectroscopic techniques, each possessing different characteristics and capabilities, to give the reader a glimpse at the wide array of specialized fields the proposed methods can be applied to and which usually require field specific experts. Chapter 3 discusses noise, different types and possible sources and reviews common noise reduction techniques and their effect on sampled data. The basics of baseline distortions are introduced in Chapter 4 existing baseline estimators are presented in detail and discussed. Chapter 5 introduces wavelet transform which is essential to the new noise reduction technique presented in chapter 6 that can be used to accurately estimate and remove noise even if its intensity varies within a single measurement. Combining hard thresholding, section 6.1, and shift invariant wavelet transform, section 6.2, with a new approach of using a threshold function instead of a static threshold value, the new noise reduction method is able adapt to changing noise intensities, section 6.6, and suppress noise distortions with minimal degradation of the underlying signal, section 6.7. Chapter 7 presents a new baseline estimation technique based on local regression that is more robust to distortions than other state of the art techniques and is able to perform well in a wide array of scenarios by automatically adapting to baseline and measurement characteristics. The core regression algorithm is presented in section 7.1

while challenges inherent to the task of baseline estimation are presented in section 7.2. Section 7.3 presents the new enhancements and solutions that change and optimize the original core algorithm for the task of baseline correction and draws first conclusions by comparison with other state of the art baseline estimation techniques. Strengths and weaknesses of the new technique are analyzed in detail by discussing synthetic and real examples in chapter 8.

1.1 Results

Changing noise conditions within a single measurement were observed in real several spectra which let to poor results using conventional noise estimation and suppression methods. Usage of a threshold function in place a the global, static threshold can effectively suppress or remove noise that exhibits varying characteristics within a single measurement improving both noise suppression and signal reservation by adapting to the actual, local noise level instead of a value that minimizes the global error.

The advanced noise estimation can work independently for a variety of different data and also is a key factor in the presented new baseline estimation technique. Used in combination with Adaptive local regression (ALR) the noise threshold function removes the need for iterative noise estimation from the baseline estimation process allowing better discriminability between noise and signal features. ALR is able to produce high quality baseline estimations in the presence of variable noise, multidirectional distortions and a wide variety of baseline shapes. ALR does not use underlying baseline models and is not dependent on specific signal shapes in order to separate signal and baseline features. It also removes the assumption of models that are limited one directional errors made by other state of the art correction methods thus making baseline correction more robust and more reliable in real world scenarios.

Advanced noise reduction and adaptive local regression are powerful and easy to control tools that can significantly enhance the performance of any signal recognition technique by reliably reducing two of the most common and detrimental distortions found in almost all spectroscopic measurements. Both methods have been tested on real world spectra acquired at standoff distances displaying challenging, unpredictable distortion characteristics.

Both methods are applied in the chemometrics module of the Optix prototype to significantly enhance signal quality characteristics of standoff measurements enabling the automatic detection of explosive materials by comparison with stored reference spectra.

1.2 Non-disclosure information

Real data discussed in this work was largely acquired during the Optix project, see 2.4 and as such are classified as security sensitive information. Information about explosive substances which were tested during the project are not to be made available to the general public as information about detection limits and priorities can easily be derived. Information about the types of substances used for the initial testing is classified as it reveals which explosive substances are considered typical or important in the context of improvised explosive devices (IED). The names of non-explosive materials tested during the project are security sensitive because a list of tested non explosive materials gives information about what materials could be used to hamper detection methods by either masking possible explosive materials or by purposefully triggering false alarms. For these reasons clear names of substances in real measurements will not be given at any point.

Chapter 2

Introduction to spectroscopy

> *Spectroscopy* : Study of the absorption and emission of light and other radiation by matter, as related to the dependence of these processes on the wavelength of the radiation.
> *Encyclopædia Britannica*

The chapter presents a general introduction to spectroscopy and its applications, as well as a brief overview over several different spectroscopic techniques, to convey a sense of similarities and differences that can be found in the different specialized forms of spectroscopy. As large parts of the analyzed data were acquired during the development of the Optix project the goals and challenges of Optix are presented with special regard to the implications for data analysis methods.

2.1 Overview

In 1666 Isaac Newton showed that the sun's natural light can be dispersed into a continuous series of colors he called spectrum. The relatively simple apparatus he used, consisting of a screen a prism and a lens can be thought of as the first spectroscope. Spectroscopy was the term used for the study of objects based on their interactions with radiation of different wavelengths and can be traced back to Thomas Melville who in 1752 observed changes in the colors emitted by a flame burning pure alcohol or a mixture of alcohol and sea salt [MIT13]. Today spectroscopy is the term used for a group of methods which analyze the spectrum of probes according to the emitted energy when subjected to a specific excitation with the goal to assess its qualitative consistency. Closely related to spectroscopy is the field of spectrometry which is defined by quantitative measurements of physical properties like concentrations of specific compounds, or parameters of electrical or magnetic fields. The key idea discovered by Thomas Melville however remains that substances, given certain conditions and external stimuli, can be identified by characteristic emission lines in spectra that are obtained from them.

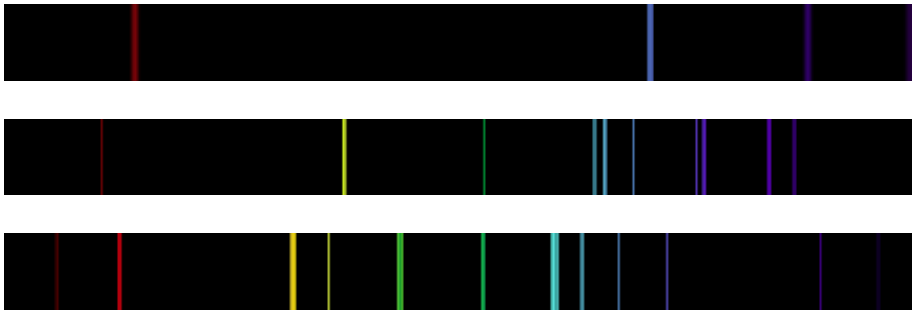

Figure 2.1: Emission lines of hydrogen (top), helium (middle) and carbon (bottom) recorded by NASA

Various incarnations of spectroscopy and spectrometry are used in the fields of physics and analytical chemistry to identify unknown substances through characteristics found in the spectra that are emitted or absorbed. The scope of techniques ranges from the detection of smallest traces of compounds in a microscopic environment to telescopes equipped with spectrometers in order to determine the composition or velocity of astronomical objects based on properties found in their characteristic spectra. Other prominent examples are space explorations missions. The mars rover *Curiosity* which is part of the Mars Science Laboratory that landed on mars in 2012 is equipped with several spectrometers including an Alpha particle X-ray spectrometer and a Laser-Induced Breakdown Spectrometer (LIBS) to determine the composition of examined samples underlining the importance of spectroscopic techniques for the discovery of new knowledge [JPL12a], [JPL12b].

2.2 Spectral data

Spectroscopic data is in most cases given as an intensity distribution over an axis, some techniques also produce higher dimensional data but these are usually based on a one dimensional technique that has been extended to include additional spatial or time information. In the general case the x-axis defines the spectral bandwidth and the sampling rate with which data was acquired. Typical units of measurements are wavelengths, frequencies or corresponding energies, derived units like e.g. wave numbers or relative units that compensate for bias introduced by different excitation sources. The unit of the intensity or y-axis is often given as counts - for CCD sensors - or simply intensity/absorption in arbitrary units when no sensible physical unit is appropriate. Data is usually presented as a graph that interpolates linearly between data points; an example is given in figure 2.2. The spectral line representation shown in figure 2.1 can also be used but is generally reserved for non-scientific subjects or very simple examples. The wavelength information is encoded by position and color while

the intensity is represented as luminescence. In this format it is usually much harder for a human observer to make quantitative statements since luminescence is difficult to quantify and certain colors are perceived as lighter or darker even if their luminescence values are identical.

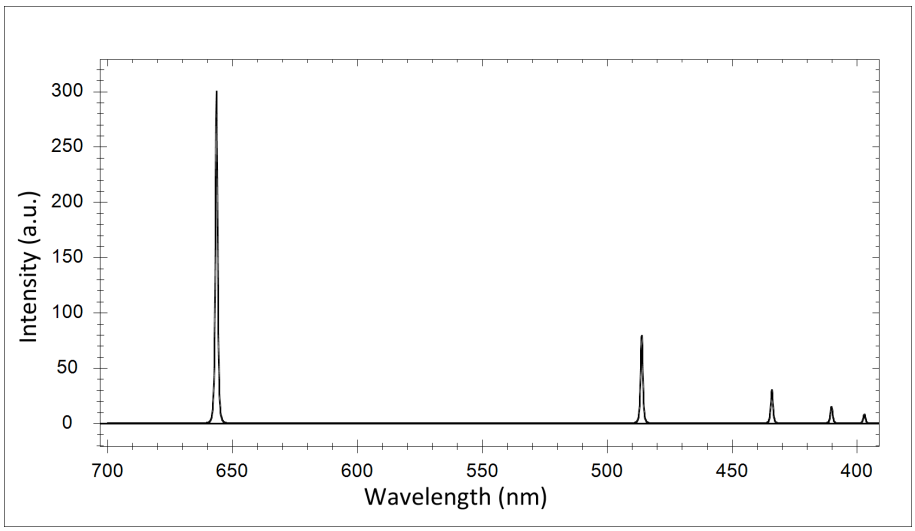

Figure 2.2: Example of the common visual representation of 1-dimensional spectral information. The unit and resolution of the x-axis are defined by the spectroscopic technique and the measuring hardware respectively while the y-axis usually defines intensity information in non-specified arbitrary units (a.u.). The spectrum shown represents the lines emissions of hydrogen, the unit of measurement is nanometer in decreasing order to match the illustration in figure 2.1,

2.3 Spectroscopic techniques

Spectroscopic and spectrometric methods are numerous as many energy sources can potentially be used as a basis for spectroscopic analysis. Additionally there are many meaningful categories that allow grouping of different techniques that there is no one distinct order to spectroscopic methods or types. The measured structures are one possible criterion. Spectroscopic methods can be used to measure atomic structures, molecules or solid matter, especially crystal structures. The measured process can be roughly divided in absorption, emission, scattering - elastic or inelastic - and others. Absorption spectroscopy compares the properties of an exciting source before and after interaction with a sample to create a characteristic absorption spectrum. Spectroscopic methods based on characteristic emissions made by substances can be summarized under the term emission spectroscopy. Since most substances do not naturally emit energy an excitation source is used which then usually determines the type and name of the specific spectroscopic emission technique. Elastic scattering gains information by

studying the reflection of excitation energy by the sample material while inelastic scattering involves the shifts in the emission observed when the excitation energy interacts with a sample. Every variable in the spectroscopic technique naturally influences the possible applications and achieved results as not every technique is suitable for every application and in fact many spectroscopic techniques are highly specialized towards a very particular subset of identifiable compounds and do not work well if certain requirements to the compounds structure are not met. Some examples of spectroscopic methods based on different physical processes are given below, for references on the different spectroscopic techniques see [SW80]. More detailed overviews and introductions of specific spectroscopic techniques can be found here: Atomic absorption spectroscopy is presented in [Wal55], laser induced breakdown spectroscopy [RCSW97], infrared and Raman spectroscopy [CDW75], Raman spectroscopy [Lon77], Auger and x-ray photoelectron spectroscopy [SB83], Nuclear magnetic resonance spectroscopy [Fri88].

2.3.1 Atomic absorption spectroscopy

Atomic absorption spectroscopy (AAS) is used in analytical chemistry for the quantitative and qualitative analysis of chemical elements. Measured is the absorption of visible light caused by interaction with free atoms. Based on the method used to create the required gaseous state of the sample material the method consists of several subtechniques; Cold vapor AAS, graphite furnace AAS, flame AAS and high-resolution continuum-source AAS. Figure 2.3 is showing an exemplary Infra-red absorption spectrum.

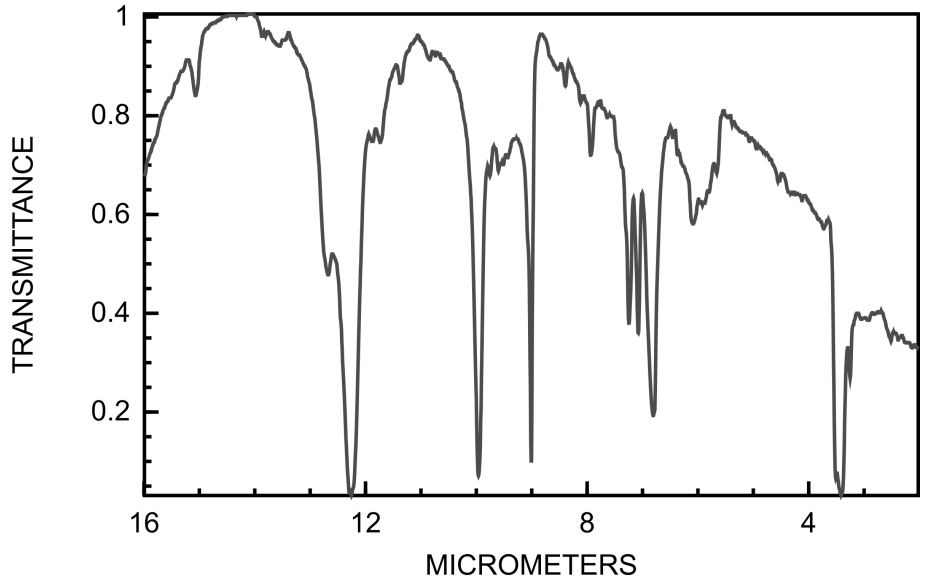

Figure 2.3: Exemplary Infra-red absorption spectrum (of Ferrocene) taken from the NIST chemistry Web Book.

2.3.2 Atomic emission spectroscopy

Atomic emission spectroscopy is used in the quantitative and qualitative analysis of solid, liquid and gaseous materials. The method is destructive and based on the observation that excited atoms emit radiation on frequencies characteristic for the chemical element. Methods are usually distinguished by the means to create and deliver the excitation energy to a sample. Examples are flame, arc, plasma and spark. Laser induced breakdown spectroscopy (LIBS) uses a laser beam to create plasma on the sample which ablates some of the surface. Other plasma based techniques include Inductively coupled plasma (ICP) which uses a plasma torch that operates at temperatures of 7000 - 9000 degree Kelvin to dissolve an evaporate the sample, for comparison the average surface temperature of the sun is about 5800 degree Kelvin. An example for a LIBS measurement is given in figure 2.4 showing the spectral response of aluminum.

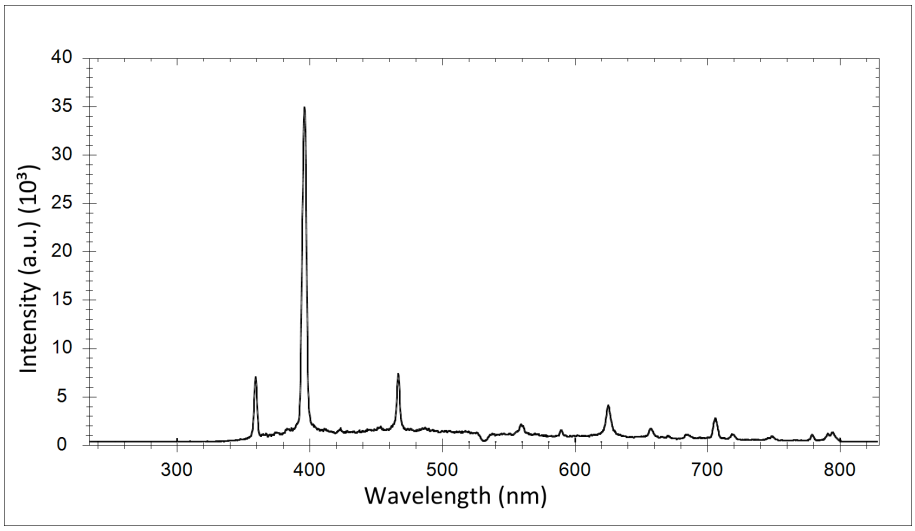

Figure 2.4: Spectrum of aluminum acquired using Laser Induced Breakdown Spectroscopy. Aluminum is an example of a substance that can be analyses using LIBS but does not produce a spectral response as part of a Raman experiment.

2.3.3 Electron spectroscopy

Electron Spectroscopy is based on the manipulation of the number of electrons in the outer shell of atoms and is used mainly for close contact surface analysis of solid objects. The analysis is carried out in vacuum and uses monochromatic x-rays as excitation. Sub-techniques are Auger electron spectroscopy, photoelectron spectroscopy, electron microprobe analysis and electron energy loss spectroscopy. Figure 2.5 shows an example of an X-ray photoelectron spectrum in which several peaks have been marked according to the corresponding atomic structure.

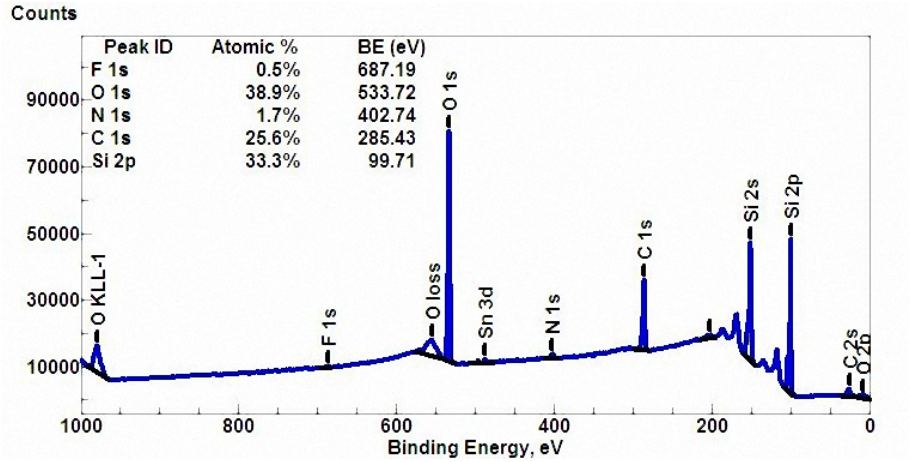

Figure 2.5: Wide scan X-ray photoelectron spectrum. Some elemental peaks are marked and signal contributions are given. Baseline corrections are marked in black.

2.3.4 Vibration spectroscopy

Vibration spectroscopy e.g. infrared spectroscopy uses low energy radiation with wavelengths of 2500 nm to 16000 nm to induce a vibration to bonded groups of atoms. These vibrations are determined by the molecular structure and the bonds between atoms and are in fact characteristic for atoms and/or small groups as their frequency is determined by the mass of the involved atoms and the strengths of the bonds. When exposed to infrared radiation virtually all organic substances absorb portions of the infrared spectrum that is characteristic for the frequencies corresponding to the contained molecular structures.

A single molecule can have several types or modes of vibrations depending on the number of atoms contained. Each mode results in one or more characteristic absorption peaks in the spectrum. These vibrational modes are usually given descriptive names like twisting, rocking, scissoring or starching to be able to better discern them. Another techniques based on the vibration of molecules is Raman spectroscopy which uses visible light to excite the sample and measures the inelastic scattering of photons. An example for spectra obtained using Raman spectroscopy is given in figure 2.6.

2.3.5 Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance spectroscopy is a non-destructive technique based on the responses of materials within a strong magnetic field when excited by a lateral electromagnetic pulse. Since not all nuclei show a response to a magnetic only selected nuclei are suitable and are name giving for the type specific of NMR, the most commonly used nuclei are ^{1}H and ^{13}C other also suitable nuclei are^{19}F and ^{31}P all of which have an integral spin of $I = \frac{1}{2}$. The energy difference between the two possible spin states

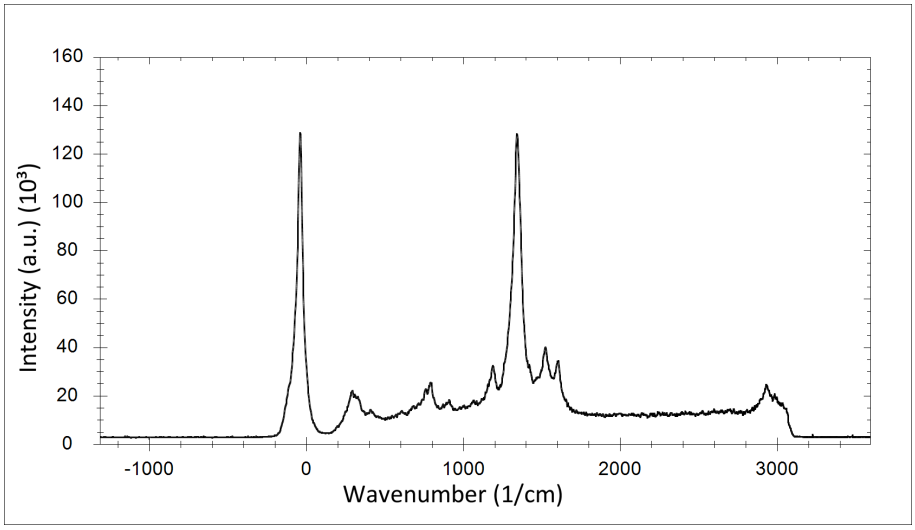

Figure 2.6: Typical Raman spectrum of an explosive substance. The measured response is based on Raman scattering effect allowing conclusions about the molecular composition of the analyte.

$I = -1/2$ and $I = +1/2$ is dependent on the external magnetic field, the stronger the field the higher the energy difference. Modern NMR spectrometers use magnetic fields up to 20 Tesla. Within the external magnetic field the magnetically active nuclei form a weak magnetic field themselves, the direction of this field is changed by a lateral pulse and the measured process is the free induction decay that occurs when the direction of the induced magnetic field returns to the direction of the external field. The lateral pulse frequency is tuned to excite only a particular Nuclei, e.g. ^1H, however the recorded signal is not only affected by the Nuclei itself - which would result in only a single response function - but also by the molecular bonds of the excited nuclei present in the probe substance resulting in a superposition of slightly different FID functions making field homogeneity an important factor, [SB13]. The same principal is also used in medical applications such as magnetic resonance tomography (MRT).

2.3.6 X-ray spectroscopy

X-ray spectroscopy is the general term for spectroscopic techniques utilizing radiation of wavelengths between 0.01 and 10 nanometers. Spectroscopic techniques are further distinguished by the energy of the radiation used. Radiation of high energy levels - and correspondingly small frequencies - is called hard, while radiation of lower energy levels is called soft. Classes of x-ray spectroscopy are: photoelectron spectroscopy (XPS), X-ray fluorescence (XRF), energy dispersive X-ray spectroscopy (EDS or EDX) and wavelength dispersive X-ray spectroscopy (WDS or WDX). EDS uses a particle beam of constant energy to excite a sample and observed the characteristic radiation emitted

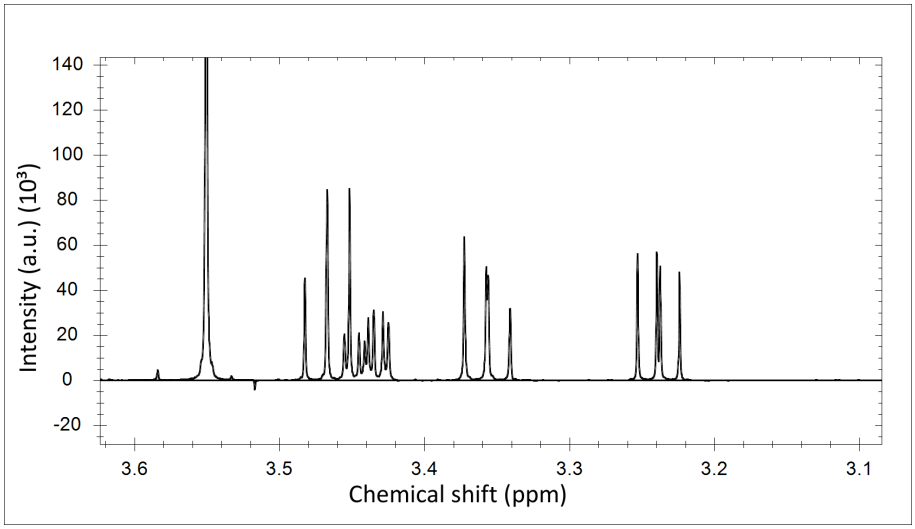

Figure 2.7: Nuclear magnetic resonance spectrum of a glucose mixture. Show is only a partial spectrum because NMR spectra are too large to achieve adequate readability in normal formatted images.

by the sample while WDS only observes a single frequency or wavelength. The detection threshold of the WDS method is magnitudes lower than that of EDS; however the EDS method allows simultaneous observation of numerous elements.

2.4 Optix

Optix is an acronym for "**O**ptical **T**echnologies for the **I**dentification of E**x**plosives", [IND13]. As a capability project the aim was to develop a portable system for the standoff diction of explosives at a distance of twenty meters using three optical technologies - Laser induced breakdown spectroscopy, Raman spectroscopy and absorption Infra-red spectroscopy - either alternatively or simultaneously. Other solutions for the detection and identification of explosives exist but no system fully meets the operational capabilities desired, which are:

- Standoff detection and identification of explosives.
- Fast and reliable (very low false alarm rates) detection and identification.
- Large operational availability.
- High degree of automation allows for operation by non-expert personnel.
- Short deployment time.
- Adaptability to variable environmental conditions.

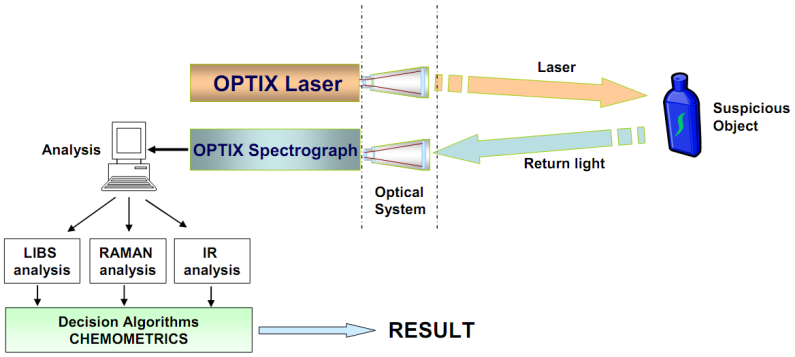

Figure 2.8: Schematic illustration of the final Optix system

- Adaptability to new threats such as (new explosives).
- Detection of very small or trace amounts of explosive material.

The most prominent and sought after feature is the capability for standoff detection and identification of explosives, in order to be able to anticipate the threat from a safe distance and avoid entering into the lethality area of an Improvised Explosive Device (IED) as until now there are no systems capable of detecting and identifying explosives from a distance, and in a real scenario, with the required minimum reliability. A target or reference distance of 20 m for several reasons:

- It is well in the boundary between the three different technologies: LIBS and Raman have been proven (not fully developed) at distances around 50 m, though samples reported were traces and bulk respectively, whereas IR needs much further development to achieve the target. A distance of 20 m represents a compromise between the capabilities of the target technologies.
- It represents a distance long enough to exceed any current capabilities and, at the same time, ensures that conclusive results are expected with minimized risk.
- Represents a range where hardware compatibility is supposed to be adequate; this is, optical systems compatible with 2 or 3 of the technologies can be employed and the design optimized, lasers requirements are much alike, etc.
- The degree of portability of the final system will be adequate; longer ranges would imply bigger optical and laser systems.

As improvised explosive devices (IEDs) propose a complex scenario in which it may be hidden in any place, may be camouflaged with virtually any shape carrying any kind of explosive charge, etc..., is not to be expected that a single technology is be

able to deal with every possible scenario. Therefore, the project was aimed towards the development of systems that combine different technologies in order to improve upon the performance of each individual technology, to reach an acceptable performance in a variety of situations.

The spectroscopic techniques selected for the Optix project use lasers as a means to deliver necessary excitation energy to a target sample. Laser beams are uniquely qualified to deliver high amounts of controlled energy to a target without requiring physical contact and as such offer the best potential for standoff detection. Due to the similarity of the experimental set-up consisting of laser, telescope, detector and computer hardware the optical spectroscopy techniques (LIBS, Raman, and Infra-red) are good candidates for the integration into a single standoff explosive detection system and, as many of the necessary instruments can be shared.

2.4.1 Challenges

Tests using standoff distance between the laser and detector instruments and low pressure absorption cells were able to show that the some explosive substances possess distinctive spectra in the mid-infrared. No stand-off systems for detection of explosives with low vapor pressure are realized based on infrared spectroscopy. The reason for this is the very low vapor pressures of most explosives, which means that only trace amounts of explosives are present in the gas phase. Exposing the gas cloud to uncontrolled environmental conditions outside of a laboratory environment aggravates this problem further. This means that the quality of IR measurements acquired at standoff distance outside laboratory environments is likely very low and advanced analytical techniques are needed to extract reliable information.

The limitations of standoff LIBS technique lie in the difficulties of transmitting enough energy at long distance, in the limited light gathering capability of the light collection system and in the light attenuation with distance under severe weather conditions (heavy fog, heavy rain, snow). The energy density at the target is a function of the distance (inversely proportional) and of the focusing mirror (or lens) diameter (directly proportional). Furthermore the energy required to create the plasma necessary for LIBS is dependent on the temperate of the sample, low temperatures will prevent plasma creation unless the energy output is severely increased. LIBS measurements are destructive, meaning that no replicate measurements of trace amounts are possible. Background interferences that are unavoidable in real scenarios outside a laboratory environment are more problematic for LIBS as an atomic detection technique than it is for molecular techniques. Automatic analysis approaches have to consider background signals and ideally compensate for their influences on measurements made under different conditions.

Fluorescence interferences and the extremely weak nature of the Raman effect make

the use of Raman for trace detection of explosives problematic unless resonance or surface enhancement methods are used. Harmonic generation modules for 532, 355 and 266 nm studied to in an attempt to use multiple excitation wavelengths to achieve an increased probability to excite resonance. For Raman technology, the ultra-sensitive spectrometer is a must, while for LIBS the timing and gating of the detectors is the most important feature. Fluorescence effects are unpredictable especially in in scenarios where the target is unknown, and must be removed in order to enable automatic identification methods.

Other challenges are directed towards hardware and engineering. The three optical technologies (LIBS, Raman, and Infra-red) use different wavelengths for emission and reception. The final Optix prototype integrating the three technologies will have to use an optical system dealing with these different requirements. This means a custom optical system has to be developed and build. The requirements towards the optical system are different from one technology to the other. Diffractive optics cannot be employed to deal with so many different wavelengths in transmission. Light collection where UV, visible and mid infrared radiation has to be analyzed has to be developed. A reflexive optical system with a large aperture to enable collection of low intensity radiation coming from the sample is required.

Given this information a major focus of the analytical techniques developed for the project is the robust and reliable enhancement of the achieved measurements in order to reduce the distortions introduced by the standoff scenario and the highly variable conditions under which the systems has to operate.

Chapter 3

Noise, denoising and smoothing

Noise distortions are present in almost all forms of real measurements and therefore the removal or reduction of noise is a common task to enhance the quality of measurements. As it is common in practice it is assumed that options to reduce noise influences like calibrating and optimizing the measuring hardware have been exhausted and that data is presented as is, without further information or access to external parameters. In this chapter an overview of the nature of noise including several different types of noise is given and common methods to reduce noise influences in measurements after they have been recorded are discussed. Since digital imagery is a popular and widely known field of application in which noise and the reduction of noise plays an important role it is not surprising that many denoising techniques are often developed for data that represents images. An excellent reference in this field is the book Digital Image Processing by Gonzalez and Woods [GW92] which is suggested to readers as a source of further information on filtering.

3.1 Noise distortions

3.1.1 Definition of noise

> *Noise*: In acoustics, any undesired sound, either one that is intrinsically objectionable or one that interferes with other sounds that are being listened to. In electronics and information theory, noise refers to those random, unpredictable, and undesirable signals, or changes in signals, that mask the desired information content.
> *Encyclopædia Britannica*

The term noise is based on audible acoustic interferences caused by irregular behavior in electrical currents. The corresponding phenomenon in the context of video data is also called snow. The term was adapted by other techniques and today noise is a synonym for random interferences characterized by a wide unspecific frequency spectrum.

Noise plays a large role in the transmission of data as all parts of the transmission system - sender channel and receiver - are generally affected by noise.

3.1.2 Types of noise

Noise types are often defined by their characteristic spectrum and named for the proportions of different frequencies analogous to visible light spectrum. High frequencies are blue or violet lower frequencies correspond to reddish tones. Some well-known noise types are given here

White noise

The most common noise is thermal noise which can be approximated by a normal distributed or Gaussian random process, this type of noise is also called white noise because it covers all frequency bands to an equal degree. The discrete probability density function is described by the normal law. Let events occur at an average rate of a and let a be large then the probability that an event n occurs in a given interval of time is given as

$$P(n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(n-a)^2}{2\sigma^2}}. \tag{3.1}$$

For $a = \sigma^2$ the normal law is also called Gaussian law. Because of its universal characteristics and the central limit theorem, see below, white noise is generally used as reference if no specific noise characteristics are known.

Shotnoise

Shot noise is the noise caused by the quantification of currents due the electrical charge of electrons. Shot noise can be characterized by the Poisson distribution. The discrete probability density function is described by the Poisson law. Assuming that individual events are independent and occur at random at an average rate of a then the probability $P(n)$ that an event n occurs in a given unit of time is defined as

$$P(n) = \frac{e^{-a}a^n}{n!}. \tag{3.2}$$

For large values of n the Poisson law reduces to the Normal law. Furthermore practically all fluctuations generated in electrical devices can be described by a normal distribution function. The reason for this is that these fluctuations are caused by a large number of independent random variables and the central limit theorem holds.

Pink noise

Another type of noise is so called Pink or $1/f$ noise is characterized by a power distribution that drops with increasing frequency. An example for (approximate) pink noise it the sound of a waterfall where noise characteristics are caused by the different sized water drops as they collide with each other and the water surface.

Red noise

Red noise - also known as Brownian noise or $1/f^2$ noise - has a characteristic spectrum similar to pink noise but the energy drops off faster towards higher frequencies. Brownian noise is based on the Brownian motion of molecules described by the botanist Robert Brown in 1827 who studied to the motion jittering motion of pollen in water.

Central limit theorem

In everyday language the central limit theorem states that given a sufficiently large number of random variables the sum of those variables is approximately normally distributed. Or more specifically:

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and a finite variance σ^2. Then the sum $Y_n = \sum_{i=1}^{n} X_i$ has mean $\mu_Y = n\mu$ and variance $\sigma_Y^2 = \sigma^2 n$ and the distribution function of the random variable $C_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$ approximates the normal distribution $N(0,1)$ for large n.

This theorem is also the main reasoning behind the model assumption of normal distributed noise and in fact models based on the normal distribution lead to good results in estimation and simulation.

3.1.3 Sources of noise

The sources of noise are various and often the observed noise is a combination of many different internal and external noise sources. A complete list of all noise sources cannot be given and it is likely that not all sources are known. Internal noise sources are often caused by specific electrical components, the connection between different components or the variability of currents in a circuit. Examples include generation-recombination noise of transistors, ionization noise, shot noise and thermal noise which is caused by the random motions of current carriers in all electrical instruments. Quantification noise in electrical systems happens due to the fact that currents themselves are not continuous but quantified as sums of charges of single electrons. External sources of noise are e.g. cosmic noise which is caused by the electromagnetic fields emitted by cosmic objects such as our sun or distant quasars, noise created by phenomena in earth's atmosphere like lightning discharges, rain, air pressure and temperature changes, or man-made

influences like electromagnetic fields of electronic devices, pollution and stray light of artificial light sources.

3.2 Real space filter

Real space filter operate on a signals given as set of intensity values corresponding to equally spaced samples. Spectra are a typical example for one dimensional data while images acquired with a digital camera are common example for discretely sampled two dimensional data.

3.2.1 Mean filter

One of the most simplistic filters that are used to reduce the influence of high frequency noise is the mean or moving average filter. As the name suggests data values x_i of the original measurement are replaced by a new value that is the average value taken from a neighborhood containing $K + 1$ elements. Let K be even then the filter can be mathematically described as

$$g(i) = \frac{1}{K+1} \sum_{k=-K/2}^{K/2} x_{i+k}. \tag{3.3}$$

Since frequency is often used as a defining characteristic of noise it is sensible to look at the filters characteristics with regard to frequencies to achieve a better understanding of its effect on filtered signals. Since the mean filter as defined in equation 3.8 describes a linear time invariant (LTI) system its frequency response can be expressed as the discrete time Fourier transformation of its impulse response. Given an impulse response function $h(m)$ and j with $j^2 = -1$, the frequency response function $H(\omega)$ is generally described by

$$H(\omega) = \sum_{m=-\infty}^{\infty} h(m) e^{-j\omega m}. \tag{3.4}$$

The impulse response function of the moving average filter is dependent on the size of the neighborhood or with regard to sampled signals on the number of values that are used to calculate the average. Let K be the number of sample points used for the average filter then the impulse response is given by

$$h(m) = \begin{cases} \frac{1}{K} & for \quad m < K \\ 0 & else. \end{cases} \tag{3.5}$$

Because of the finite impulse response (FIR) characteristic of the mean filter its frequency response function can thus be reduced to a finite sum

$$H(\omega) = \frac{1}{K} \sum_{m=0}^{K-1} e^{-j\omega m}, \tag{3.6}$$

the summation can be rewritten using the geometric sum criterion to yield

$$H(\omega) = \frac{1}{K} \frac{1 - e^{-j\omega K}}{1 - e^{-j\omega}}. \tag{3.7}$$

The frequency response function reveals the low-pass characteristics of the moving average filter but also unwanted, non-monotonous suppression behavior.

3.2.2 Gaussian filter

Visual inspection of the frequency response function, shown in figure 3.2, reveals that the mean filter exhibits low pass characteristics meaning that frequencies in close to zero are not attenuated while higher frequencies are, but the relation between frequency and attenuation is not monotone. In fact while some frequencies are completely eliminated other, higher frequencies pass the filter and are attenuated to a smaller degree. The cause for this behavior is the fact that the mean filter itself possesses sharp edges, since a data-value is either contributing with a weight of one or zero. In that context the Gaussian filter can be regarded as an enhanced version of the moving average filter that takes the frequency characteristics of the filter into account to avoid non monotonous behavior of the frequency response. In contrast to the mean filter which applied constant weights to all data-values used for the calculation, the Gaussian filter uses a filter kernel that assigns weights based on the Gaussian distribution thus emphasizing the influence of closer data-values and assigning less weight to data-values that are farther away from the middle of the filter window. The kernel function $g(x)$ of the Gaussian filter is given by

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \tag{3.8}$$

and the frequency response function of the ideal windowed Gaussian filter

$$g(\omega) = e^{-\frac{\omega^2}{2\sigma_\omega^2}}. \tag{3.9}$$

Among standard mean and Gaussian filter many other possible kernel function exist that can be used for the purpose of signal smoothing some examples include the Triangular or Bartlett kernel, Hamming, cosine and Kaiser kernel. Each kernel describes a characteristic weight distribution of the data-points within the window function resulting in slightly different filter characteristics.

3.2.3 Percentile filter

The percentile filter is a nonlinear technique that uses a window function similar to the mean filter to define a neighborhood; filtered data values are calculated based solely on this user defined neighborhood. The difference to mean, triangle or Gaussian filter is the use of the rank or cumulative histogram function instead of the actual data values. A percentile parameter of $p = 0.5$ thus models a median filter. Mathematically the filter can be described as

$$P_p(x) = \inf\left\{t|H_{f|N_x}(t) \geq p|N_x|\right\}, \tag{3.10}$$

where f is the signal, H is the cumulative histogram function and N_x the neighborhood of x. The filtered data value is the value of the element that is at the position in the ordered neighborhood defined by the neighborhood size and the parameter p. Only the rank of elements is considered for the selection not the actual values. The approach works well with data that is corrupted by intense but infrequent distortions. It is very robust against outliers but renders a more angular result than filters working directly on the measured data. While the computational effort of the filter is generally higher than that of filters operating on the data values because the rank function needs to be calculated for every step Duin et al. [DHZ86] have shown that some simple optimizations can reduce the computational effort considerably.

From the theoretical standpoint the percentile filter is neither linear nor a morphological filter. Because of the rank based characteristic it is not possible to define a response or transfer function analogue to the mean or Gaussian filter. While a percentile of 0.5 is the most natural choice for data smoothing the filter can also be used to create upper and lower envelopes of a signal by setting the percentile parameter to 1 and 0 respectively. If used on higher dimensional data the aperture of the filter can take various shapes similar to morphological filters. Filter shapes and sizes are the

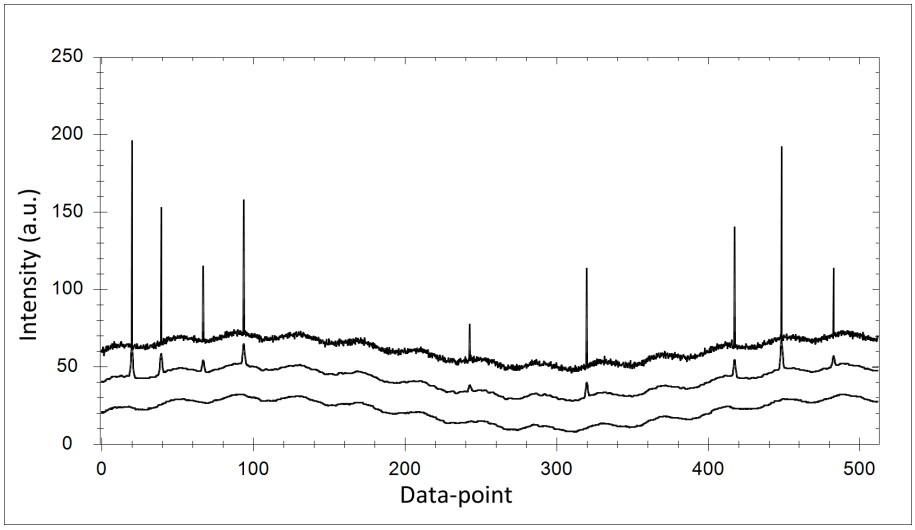

Figure 3.1: Example of a synthetic noisy signal ($\Delta x = 0.25$) with intense outliers (top) filtered with a Gaussian filter, window-size $w = 8$ (middle) and a percentile filter, $p = 0.5; w = 8$ (bottom). A different constant offset was added to each data-series for better visualization after the filtering process.

essential parameters to achieve the desired effect and need to be adapted by a user to suit the filtered data.

3.2.4 Savitzky-Golay filter

The Savitzky-Golay filter is based on the works originally presented by Abraham Savitzky and Marcel J. E. Golay in 1964 [SG64]. Observing that moving average filters have a flattening (signal loss) and widening (line broadening) effect on spectral signals Savitzky and Golay presented their ideas as an alternative which kept the smoothing capabilities of the moving averages filter but avoided its main problems. The idea of the filter is to perform a local polynomial regression for each data-point and then return the value of the regression curve at this point as the smoothed result. The characteristic feature of the method - and arguably the reason for its success - lies in the fact that the whole calculation was broken down into coefficient vectors. Thus the application of the Savitzky-Golay filter can be achieved in the same fashion as the mean and Gaussian filter by simply using the vectors supplied by Savitzky and Golay as kernel functions. Table 3.1 shows an extract of the first table given in the original paper for various filter sizes.

The filter scheme requires a sampled input with equally spaced sample points to produce correct results. These requirements are usually implicitly satisfied by modern measurement instruments. Key to the success of Savitzky and Golay's method is the observation that the coefficients of the least squares fit with a polynomial function

Number of Points	25	23	21	19	17	15	13	11	9	7	5
-12	-253										
-11	-138	-42									
-10	-33	-31	-171								
-9	62	-2	-76	-136							
-8	147	15	9	-51	-21						
-7	222	30	84	24	-6	-78					
-6	287	43	149	89	7	-13	-11				
-5	343	54	204	144	18	42	0	-36			
-4	387	63	249	189	27	87	9	9	-21		
-3	422	70	284	224	34	122	16	44	14	-2	
-2	447	75	309	249	39	147	21	69	39	3	-3
-1	462	78	324	264	42	162	24	84	54	6	12
0	467	79	329	269	43	167	25	89	59	7	17
1	462	78	324	264	42	162	24	84	54	6	12
2	447	75	309	249	39	147	21	69	39	3	-3
3	422	70	284	224	34	122	16	44	12	-2	
4	387	63	249	189	27	87	9	9	-21		
5	343	54	204	144	18	42	0	-36			
6	287	43	149	89	7	-13	-11				
7	222	30	84	24	-6	-78					
8	147	15	9	-51	-21						
9	62	-2	-76	-136							
10	-33	-31	-171								
11	-138	-42									
12	-253										
Norm	5175	805	3059	2261	323	1105	143	429	231	21	35

Table 3.1: Weights and norm of one possible Savitzky-Golay smoothing filter for different window sizes

can be pre-computed without the need for actual data. Since these coefficients are universal they can be used with any sampled data-series in a later step, reducing the computational load dramatically. The coefficients can be calculated as follows: Let $n_L + n_R + 1$ be the size of the moving window with n_L as the number of points to the left of the "current" data-point i and n_R the points to the right. If f is the vector of data values and the polynomial of degree M to fit the values $f_{-n_L}, \ldots, f_{n_R}$ is given by $a_0 + a_1 i^1 + \ldots + a_M i^M$ then the vector of a_j's in matrix notation is given by

$$a = (A^T \cdot A) \cdot (A^T \cdot f), \tag{3.11}$$

with $A_{ij} = i^j$, $i = -n_L, \ldots, n_R$ and $j = 0, \ldots, M$. The original paper by Savitzky and Golay contained several errors in equations and tables that were corrected by Steiner, Termonia and Delour [STD72]. Furthermore Madden commented on simple methods to generate larger vectors than those given by Savitzky and Golay if needed [Mad78]. A general overview of the calculation including asymmetric filter windows and source code is given in [SS89].

3.3 Fourier space filter

Other than real space filter Fourier space filter work in the frequency domain thus usually requiring the data to be transformed into a representation that encodes information in amplitude and phase information of basis frequencies. Signals in the Fourier domain are represented as sums of sine waves and often the Fourier transformation is viewed as a decomposition of a signal into its encompassing frequencies even if that simplification is not accurate.

3.3.1 Low pass filter

Low pass filters are frequency filters and as such process signals in the frequency domain. Since data is usually not given as frequencies the application of frequency filters consists of three steps: Fourier transform, filter application to the transformed signal and inverse Fourier transform. Low pass filters work under the assumption that high frequencies contain fine detail information and noise while low frequencies contain general trends and larger, slow changing information. If an appropriate filter kernel can be found, frequency filters can also be implemented as spatial filters. Filter application in the Fourier or frequency domain is achieved by multiplying the transformed signal with the appropriate filter function

$$G(k) = F(k) \cdot H(k), \tag{3.12}$$

where $F(k)$ is the signal in the frequency domain, $H(k)$ the filter function and $G(k)$ the filtered signal in the frequency domain. A main reason to use frequency filters is the computational complexity of the convolution operation. Convolution in the spatial domain is a costly operation especially for higher dimensional data - images as two dimensional data are the most popular example - however a convolution in the spatial domain is equivalent to a multiplication of the transformed signal and the transformed filter in the frequency domain. For higher dimensional data e.g. images, Fast Fourier transform makes the three step application of the frequency filter (transform, filtering, inverse transform) less computationally complex and thus faster than the application of the filter with the same effect in the spatial domain. The ideal low pass is one of the most simple filters, it sets all coefficients of the Fourier transformed signal to zero that belong to frequencies above a given threshold T. This behavior also immediately explains the name of the filter; it lets low frequencies pass through unchanged while blocking all high frequencies:

$$H_{LP}(k) = \begin{cases} 1 & for \quad k \leq T \\ 0 & for \quad k > T. \end{cases} \tag{3.13}$$

The ideal filter possesses a sharp edge results in a series of over and undershoots, so called ringing artifacts based on the Gibbs phenomenon [Gib98], [Gib99]. These effects can be reduced by designing a filter that drops off smoothly - thus approximating the behavior of real low pass filters - or when speaking in terms of the spatial domain by modifying the window function so that it drops off more smoothly. Common low pass filters are filters that possess an exponential drop off characteristic towards higher frequencies e.g. Gaussian shaped filters or Butterworth filter. These filters are not ideal as they also change lower frequencies and do not block all high frequencies but usually provide visually better results as artifacts introduced by sharp filter apertures are avoided. The Gaussian filter is a good example to illustrate the connection between frequency filters and spatial filters. The frequency response function of a Gaussian filter is again a Gaussian function - using a different factor. This means that in place of a spatial Gaussian filter is also possible to use a frequency Gaussian filter and achieve the same effect. For other filters the corresponding filter kernel in spatial and frequency domain do not share the same characteristics but still describe the same effect on the filtered signal. Figure 3.2 gives an example of the frequency response functions of a Gaussian and a mean filter that illustrate their effect on different frequencies. As mentioned before the characteristic frequency response function of a Gaussian smoothing filter corresponds to a Gaussian function in the frequency domain. The maximum of the frequency response function is one meaning that no frequencies are intensified and it reaches this maximum for a frequency of zero which means that a the lowest frequency with regard to the filter aperture is not changed. In other words, a linear function passes the filter without attenuation. Other, higher frequencies are suppressed by a factor that is proportional to the frequency. The second frequency response function shown in figure 3.2 belongs to a moving average filter. The overall characteristic of the moving average filter with regard to frequencies is similar to the Gaussian filter. Higher frequencies are suppressed to a higher degree than lower frequencies thus defining a low pass; however the suppression behavior is not monotonous. The frequency response function reveals several zeros introduced by the sharp filter aperture which result in alias effects.

3.3.2 Threshold filter

Pruning or threshold filters are nonlinear filters operating in the frequency domain. The idea behind pruning is simple: Assuming noise has a characteristic constant power

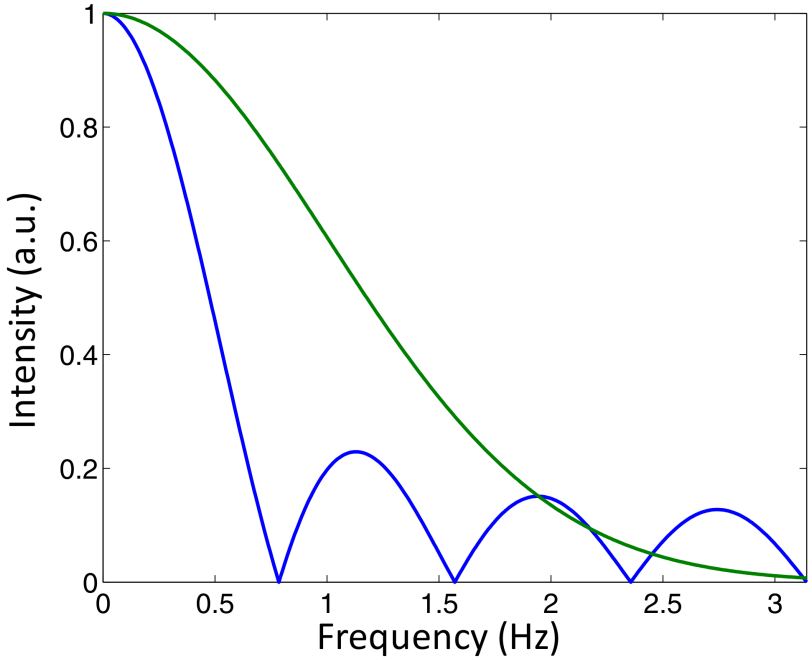

Figure 3.2: Illustration of the frequency response functions of moving average filter (blue) and Gaussian filter (green). The parameters for the specific functions shown are $K = 8$ for the moving average filter and $\sigma = 1$ for the Gaussian filter.

then eliminating frequencies from the signal that are below that power threshold should eliminate a considerable portion of noise from the signal. The mathematical description of the pruning filter is given by

$$F_{Prun}(k) = \begin{cases} \sqrt{|F_R(k)|^2 - \lambda E(F_R)^2} & for \quad |F_R(k)| > \lambda E(F_R) \\ 0 \quad else. \end{cases} \tag{3.14}$$

where R is the mean amplitude of noise distortions in the signal and $\lambda \geq 1$ a user defined correction factor. Higher values of λ result in stricter filtering conditions since all coefficients are attenuated by the pruning filter to a degree defined by the mean noise amplitude and the correction factor. The idea of pruning filters in the frequency domain is similar to threshold filters in the wavelet domain which are discussed in chapter 6. Figure 3.3 illustrates the effects of pruning on a signal in the frequency domain. Since many signals possess a characteristic coefficient distribution with high coefficient values in low frequency regions and low coefficient values in high frequency regions the results seem similar to a low pass filter. However with regard to the representation chosen in figure 3.3 an ideal low pass describes a vertical threshold that separates the transformed

spectrum in left (low frequencies) and right (high frequencies) while a pruning filter can be seen as a horizontal threshold that separates the coefficients in high (absolute value above threshold) and small (absolute value below threshold).

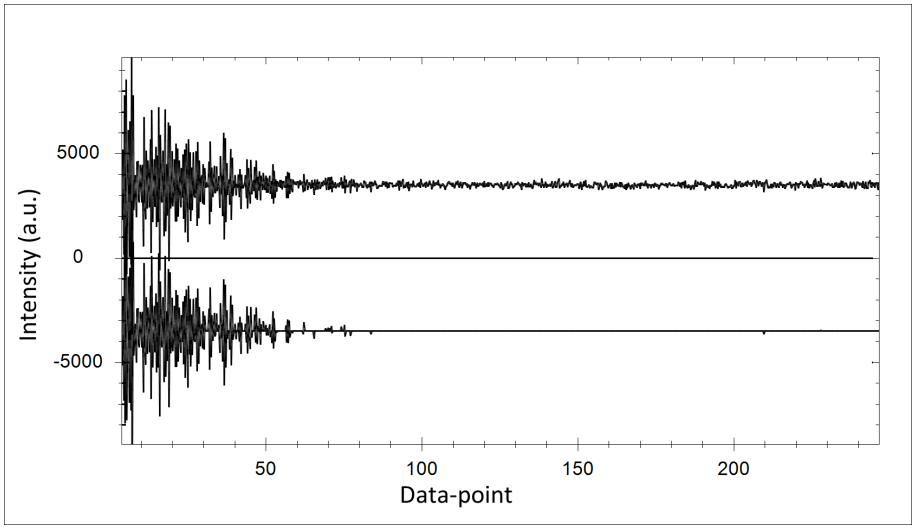

Figure 3.3: Illustration of a typical signal (real part) in the frequency domain. Shown are the original unaltered real coefficients (top) and the coefficients of the signal after a pruning filter with $\lambda = 3$ was applied (bottom). Constant offsets were applied to both signals for better visual representation.

3.4 Other denoising filters and methods

3.4.1 Repetition averaging

Iterative averaging is a simple but powerful method to reduce noise in static measurements. When adding two random variables X and Y with known variances σ_x^2 and σ_y^2 the variance of the resulting function $\hat{\sigma}^2$ is defined as

$$\hat{\sigma}^2 = \sigma_x^2 + \sigma_y^2 + 2Cov(X, Y), \tag{3.15}$$

where $Cov(X, Y)$ describes the covariance between X and Y. Assuming that noise and signal are uncorrelated, that signal is constant and noise is random with constant variance σ^2 and an expected value of zero, noise encountered in repeated measurements is uncorrelated. This means that the covariance of two noise variables belonging to different measurements is zero. Then the noise variance of the summation can be expressed as

$$\hat{\sigma}^2 = \sigma_x^2 + \sigma_y^2. \tag{3.16}$$

Since noise variance is constant meaning $\sigma_x^2 = \sigma_y^2$ the standard deviation of the summation is $\hat{\sigma} = \sqrt{2\sigma^2}$. Regarding the ratio between signal S and random noise with constant parameters N this means that signal contributions grow faster than noise contributions when calculating the sum of measurements. Given k replicate measurements the summation of signal contributions is obviously $k \cdot S$, while the noise contributions are characterized by $\sqrt{k\sigma^2}$. The gain achieved by averaging is immediately apparent when examining the ratio of expected signal S and expected noise values N:

$$\frac{S_{sum}}{N_{sum}} = \frac{kS}{\sqrt{k\sigma^2}} = \sqrt{k}\frac{S}{\sigma}. \tag{3.17}$$

Under the described assumptions an averaging of k measurements enhances the signal to noise ratio by a factor of $\sqrt{k}$.

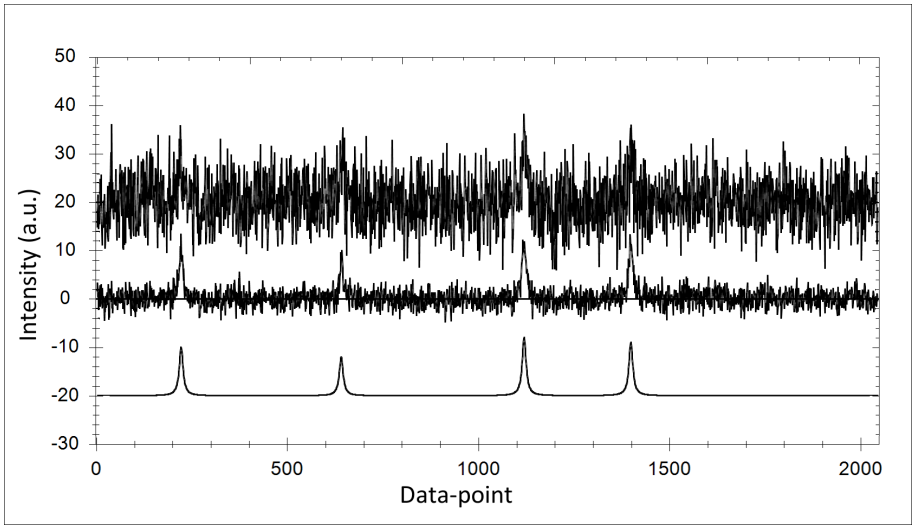

Figure 3.4: Illustration of the signal to noise ratio enhancements achieved by signal averaging. Synthetic signal afflicted with normal distributed noise $\mu = 0$, $\sigma = 5$ (top), Average noisy signal, $k = 10$ (middle) and pure synthetic signal without noise for reference purposes (bottom).

3.4.2 Wiener filter

The Weiner filter can be interpreted as an inverse filter coupled with a filter based on spectral density. The theory was developed by Norbert Wiener and published in 1949 [Wie64]. The filter is designed to reduce noise in a given stationary signal using statistical information of signal and noise. An optimal noise reduced signal is calculated

using a mean squared error approach. Furthermore the Wiener filter is related to Kalman filter [Kal60] in the sense that Kalman solves the filtering problem similar to the Wiener filter but operates on non-stationary signals. The advantage of the Wiener filter when compared to pure inverse or pseudo inverse filters is its ability to reduce random noise and blur interferences while direct inverse filtering can reduce the interferences introduced by a transfer function g but not the random influence of noise. The signal model which describes the input signal for the Wiener filter is assumed as

$$z_k = (h_k \cdot ((g_k \cdot s_k) + n_k)), \tag{3.18}$$

where z_k is the observed signal, s_k is the noise free signal and n_k the additive noise, g_k an internal filter function (e.g. the characteristics of the measuring instrument) and h_k the Wiener filter function. Let ξ be the noisy obscured signal at the input and let the correlation functions R_s, R_ξ and R_y be known. Furthermore let P denote the Fourier transform of the correlation functions or the power spectral density, then the relation between power spectral density of the noise signal P_ξ, the Wiener filter function and internal filter function in the Fourier domain H and G and power spectral density of pure signal and noisy signal $P_{s\xi}$ is described by

$$P_\xi(z)H(z) = P_{s\xi}, \tag{3.19}$$

and

$$P_{s\xi}(z) = P_s G(\frac{1}{z}). \tag{3.20}$$

The transfer function of the wiener filter can be described as

$$H(z) = \frac{P_s(z)G(1/z)}{P_n(z) + P_s(z)G(z)G(1/z)}. \tag{3.21}$$

For noise power densities $P_n(z) \to 0$ the Wiener filter is a simple inverted filter $G(z)^{-1}$. For values $P_s(z)G(1/z) << P_n(z)$ the filter values H(z) approach zero, see [OV10].

3.4.3 Whittaker smoother

The Whittaker filter is based on an algorithm published by E.T. Whittaker in 1923 [Whi22]. The same principals and procedures are also used by the Hodrick-Prescott filter [HP97] a well-known filter in the field of economics. The discrete penalized least

squares method is used to calculate a smooth series z to a set of data points y the algorithms by balancing two conflicting criteria: 1) Minimize the overall deviation from the original data. 2) Minimize the roughness of z. The deviation from the original series can be described by the sum of squares of the differences between z and y

$$S = \sum_i (z_i - y_i)^2, \tag{3.22}$$

while the roughness can be measured by the sum of a derivative function, e.g. the discrete first derivative

$$R = \sum_i (z_i - z_{i+1})^2. \tag{3.23}$$

In his original paper Whittaker used third order differences, $R = \sum_i (z_i - 3z_{i-1} + 3z_{i-2} - z_{i-3})^2$, to represent the roughness penalty. Practically it is possible to use the order of difference as an input parameter of the filter. Both terms S and R are balanced using the user defined real parameter $\lambda \in \mathbb{R}$

$$Q = S + \lambda \cdot R. \tag{3.24}$$

Larger values of λ will increase the influence the R leading to smoother series z at the expense of the overall fit to the original data y. A positive aspect of the algorithm is the possibility to easily modify it to handle missing data. To do so a weight vector w is introduced with $w_i = 0$ for missing values and $w_i = 1$ otherwise. The missing y_i are set to an arbitrary value - commonly zero - and added to the series. By leaving R unchanged and adjusting the calculation of S to

$$S = \sum_i w_i (z_i - y_i)^2. \tag{3.25}$$

Missing values of y are automatically smoothly interpolated in z. This feature is also used by airPLS which in introduced in [ZCL10]. A notion of optimal smoothing - for constant noise - can be achieved by using cross validation to choose the value of λ. In principal the cross validation process calculates smooth predictions $\hat{y}_i$ for every y_i using only the remaining data-points. This leads to the cross validation standard error

$$e_{cv} = \sqrt{\frac{1}{m}\sum_i (y_i - \hat{y}_i)^2}, \tag{3.26}$$

where m is the number of data-points in y. The optimal smoothing parameter λ^* is then defined as that λ which minimizes e_{cv}. For a detailed discussion on the efficient calculation of e_{cv}, see [EB05].

3.5 Estimation of variability

An important task in data analysis is the estimation of spread or variability in a given data set [Tuk77], [HMT83]. The gained information can be used to derive characteristics of contained signal and distortions in order to optimize detection, fitting or suppression techniques. In the context of automatic enhancement of spectroscopic measurements the estimation of noise is important factor because automatic noise reduction techniques require a parameter to determine the strength or degree of suppression. In order to determine a suitable degree it is consequently important to reliable estimate distortion characteristics encountered in real measurements. Well known measures of variability are standard deviation, variance, interquartile range, average absolute deviation and median absolute deviation. Let $Y = \{y_1, y_2, \ldots, y_N\}$ represent a given dataset, then definition of each measure is given here.

Variance

The Variance σ^2 is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - E(Y))^2. \tag{3.27}$$

Standard Deviation

The standard deviation σ is defined as the square root of the variance so

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - E(Y))^2}. \tag{3.28}$$

In the absence of outliers the standard deviation returns precise results however it lacks robustness in case the underlying data contains particular large and/or small values.

Average absolute deviation

The average absolute deviation σ_{AAD} is defined as

$$\sigma_{AAD} = \frac{1}{N} \sum_{i=1}^{N} |y_i - E(Y)|. \tag{3.29}$$

Deviations or distances are not squared for the calculation of the average absolute deviation which means that outliers with extreme values have less influence on the result than on variance and standard deviation. However even a single outlier can severely influence the result.

Interquartile range

The interquartile range σ_{IQR} is the value of the 75th percentile minus the value of the 25th percentile. Like the median is uses a rank function and is consequently more robust against outliers than methods operating directly on the data values. The so called breaking point of the interquartile range is at 25% of the original data.

Median absolute deviation

The median absolute deviation σ_{MAD} is defined as

$$\sigma_{MAD} = med(|y_i - med(Y)|, \ldots, |y_i - med(Y)|), \tag{3.30}$$

where $med(Y)$ is the median or 50th percentile of a dataset Y. The main argument that makes the median absolute deviation appealing for a wide array of practical applications is its high degree of robustness against outliers. This feature is inherited from the median and allows for sensible results to be obtained even if 50% of the data is affected by intense outliers. A drawback of σ_{MAD} when compared to standard deviation or average absolute deviation is the additional computational complexity that comes with the calculation of the rank function. While the mean of a population can be found by addition and a single division $O(N)$, simple implementations of the median sort the data and then calculate the median from the sorted values, resulting in a worst case complexity of $O(N \log(N))$. More advanced methods like e.g. Quickselect achieve an average case complexity of $O(N)$. Further improvements to σ_{MAD} were proposed in [RC93]. In practice σ_{MAD} is often used to estimate the standard deviation of normal distributed noise that is given as a set of data which contains an unknown number of outliers. Therefore the relation between σ_{MAD} and σ in a outlier free environment is used, which is given by

$$\sigma_{MAD} = 0.6745 \cdot \sigma \tag{3.31}$$

where 0.6745 is the 75th percentile of the normal distribution. This also means that $\sigma_{MAD} \cdot 1.4826 = \sigma$. Both numbers 0.6745 and 1.4826 are often found in the context of estimating the parameters of random distributions, sometimes masked by an additional factor, without being explicitly linked to the normal distribution or σ_{MAD}.

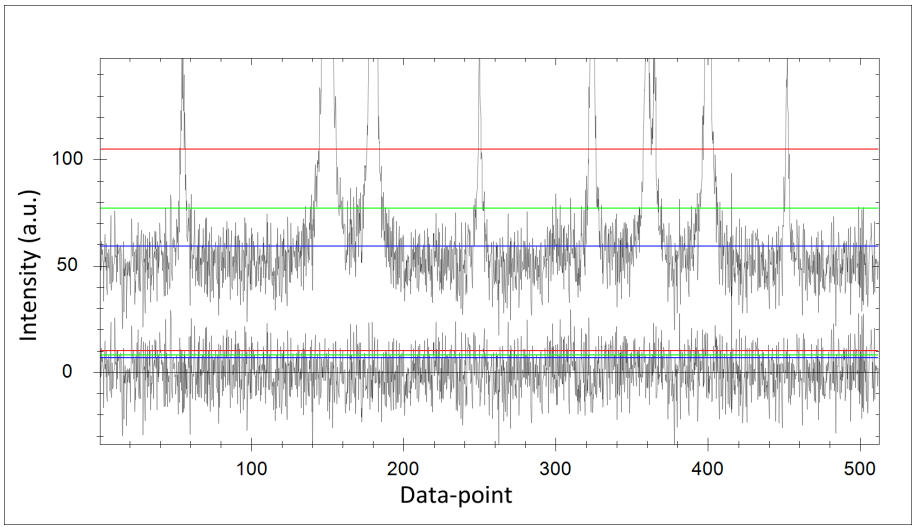

Figure 3.5: Illustration of variability estimator results on pure noise data and on synthetic data simulating a mixture of signals and noise. Shown are variance (red), average absolute deviation (green) and median absolute deviation (blue). The data of the mixture (signal and noise) and the corresponding estimator results are shifted for better readability.

	composite signal	pure noise signal	relative change
Standard deviation	54.88	10.10	443.36%
Average absolute deviation	27.08	8.08	235.15%
Median absolute deviation	9.39	6.76	38.90%

Table 3.2: Results of variability estimators on pure noise data and on synthetic data simulating a mixture of signals and noise.

The relative and absolute changes of the variability measures illustrated in figure 3.5 are given in table 3.2. Each of the three estimates is affected to a significant degree due to the fact that Lorentzian peaks, given by equation 8.3, possess wide tails and thus change a significant amount of data-values. If the goal is to estimate the noise characteristics from a mixed data-series containing signal and noise σ_{MAD} returns the most robust results although limitations of the approach are clearly visible given a relative change of nearly forty percent. Average absolute deviation σ_{AAD} and variance

σ^2 are naturally better suited to describe the composite signal as a whole and are consequently less suitable for noise estimation in a measurement containing signal and noise.

3.6 Discussion

Averaging of acquired signals requires a stable environment in which variations to the sample are minimal, a sample that is relatively stable, a non-destructive measurement technique and timing constrains that allow replicate signal acquisitions. If all these requirements are met repetitive signal averaging is an amazingly simple and effective method to drastically enhance signal to noise characteristics. Achieved results are independent of signal characteristics as long as unavoidable shot to shot variations are small in comparison to signal parameters. In fact the only reason not to use repetitive signal averaging is that requirements cannot be met. Timing constrains are often critical in real applications, either because multiple signal acquisitions consume too much time with regard to the application or because access to the sample is limited, e.g. because it is moving.

The classical moving average filter formulates a first intuition of a smoothing filter that works on a single data set. However examination of the frequency response function quickly reveals the prime reason why it is used as an example of simple smoothing operators, but has basically no place in practical applications. While the filter possesses low pass characteristics in a sense that it generally attenuates high frequencies to a greater degree than low frequencies the frequency response is not monotone. Some higher frequencies are attenuated to a lesser degree than higher ones. Additionally the filter completely eliminates some frequencies while leaving others, higher as well as lower, intact. This somewhat erratic behavior is caused by the filter aperture, specifically the sharp edges which themselves describe high spatial frequencies and thus introduce oscillating effects to the frequency response function. The Gaussian filter represents a refinement of the moving average filter that reduces the problems caused by these high spatial frequencies. It is equivalent to the multiplication with an exponential decaying function in the Fourier domain and therefore perhaps the most basic smoothing filter that has practical relevance. A drawback that the Gaussian filter inherited and which remains valid even with the adjusted filter aperture is an effect that is known as line broadening or signal degradation. In essence it is simply the effect the averaging nature of the filter has on relatively sharp signals. Sharp signals are usually defined by relatively steep edges or in other words high frequencies, these frequencies are attenuated by the low pass characteristics of the Gaussian filter and thus filtered signal edges become less sharp or steep. Since the Gaussian filter does not change the overall integral of a filtered signal, meaning the area under the curve for filtered and

unfiltered signal stays constant, the observed effect can be described as sharp features becoming smaller and gaining the corresponding area in width. The percentile filter with parameter $p = 0.5$, also called median filter, represents the percentile filter best suited for the reduction of noise with expected value of zero. Other than mean and Gaussian filter the percentile filter is a non-linear filter. It is based on the rank function of discretely sampled values which makes it robust against extreme valued outliers. The results achieved by median filtering are often angular as no, or depending on the precise implementation, only minimal interpolations are calculated that could provide smooth transitions between median values. Given these observations the median filter is more suitable as a pre-filter to e.g. a Gaussian filter as it does not return a smooth result but is able to eliminate intense outliers that cannot be addressed by Gaussian filters, see figure 3.1.

Savitzky-Golay filters are in essence least square local regressions which are pre-processed into weight matrices. The computational complexity of the filter application is thus no different than that of a Gaussian filter, the main workload of the regression is encoded into the weight coefficients which as Savitzky and Golay have shown are dependent on the regression function and window size but independent of the specify signal. Savitzky-Golay filters are commonly used in practice e.g. in various fields of spectroscopy and analytical chemistry because their average signal loss is lower and broadening effects are not as prominent as that of Gaussian filters. Figure 3.6 gives an example comparison of the effects of Gaussian and Savitzky-Golay filters. Signal degradation effects for the Savitzky-Golay filter are visibly lower than in the case of a Gaussian filter but still considerable.

The Weiner filter describes a different class of filters called inverse filter as it incorporates knowledge of the characteristics of the measuring instruments and requires additional information about signal and noise and uses this information to calculate an optimal smooth result. In a sense mean, Gaussian and even Savitzky-Golay filters are blind, or using a positive term, independent from the signal. The Wiener filter on the other hand adapts its behavior based on information about noise and signal characteristics. Starting from a single signal without additional knowledge the inverse part of the Wiener filter is essentially without effect. Furthermore, regardless of the inverse filtering, a challenge for practical application of the wiener filter is the estimation of required additional information used to calculate a least squares optimal solution, since actual noise characteristics are usually not known.

The Whittaker filter calculates a smoothed signal by balancing smoothness measures via a derivative function and squared error between distorted source signal and the calculated smooth approximation. This implicitly assumes that a signal can be characterized by a global maximal derivative function and that the balance between noise distortions and smoothness is also global. While Eilers suggests cross valida-

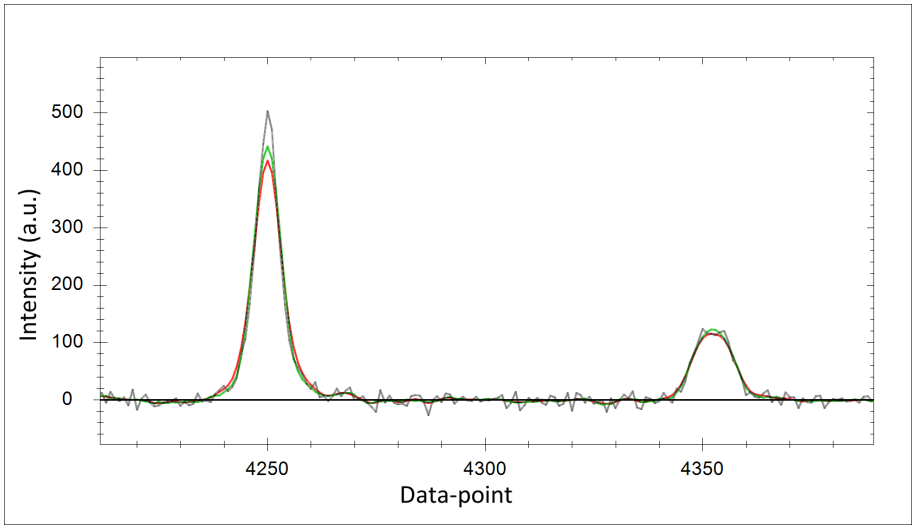

Figure 3.6: Illustration of signal-loss and line-broadening effects introduced by Gaussian and Savitzky-Golay Filters. The unfiltered synthetic signal is shown in black, the results of Gaussian filtering in red and Savitzky-Golay quadratic polynomial filtering in green. Window size of both filters was set to 5. Sample points of the series are equidistant with $\Delta x = 1$.

tion as a possible method to automatically choose the right balancing parameter the achieved results are not reliable. For automatically chosen parameters the smoothed curve often follows the original too closely and only for sub sampling of the original data a smoothing effect is achieved. This in turn means that the problem of choosing the right balancing parameter simply becomes the problem of choosing the correct sub-sampling rate. Finally it is advised to carefully consider automatic choices for real applications.

Filters that operate in Fourier space are sometimes thought to be superior to real space operators because noise is considered to be best separated from signal by frequency. However cutting off the transformed signal at a given frequency threshold introduces unwanted oscillations into the inverse transformed signal. Similar effects appear when an amplitude threshold is applied to the transformed signal. The key reason for this behavior is the global nature of the sine function. A Fourier transformed signal is represented by real and imaginary coefficients which represent amplitude (real) and phase (imaginary) of sine functions. Since the sine function is an infinity wave every point in a signal is the linear combination of all sine functions specified by its amplitude and phase in the transform domain. By removing the influence of high frequency or low amplitude sine waves the influence of the remaining functions on the entire domain is pronounced. The effect is visible in areas of the real space domain where no signal is present and an optimal corrected signal should consist of a flat line. However due to the global nature of the since function this is difficult to achieve and would inadvertently result in severe signal loss due to the flattening of high frequency peak

tops. Furthermore while the transformed signal allows an estimation of noise intensity based on the high frequency coefficients the estimation is necessary global as locality is lost in the transform process. In real measurements noise levels may be variable and information about changing noise intensities can drastically enhance signal estimation performances.

Summarizing it can be said that filters possessing low pass characteristics are aimed mainly to improve visual smoothness and less suited to truly reduce or remove noise. Based on the assumption of normal distributed random variables as adequate noise approximations noise power is constant over all possible frequencies, removing or reducing only high frequency portions therefore only removes a portion of the actual noise. The degree of smoothing is variable for all presented filters but for real space filters it is difficult to determine the optimal value in practice. Fourier filter offer the ability to estimate global noise influences using the highest frequency coefficients of the transformed signal which usually contain almost no characteristic signal influences. Removal of these noise influences presents serious problems due to the global nature of the sine function. Possible solutions are techniques that offer localization properties in the transformed signal like short time Fourier transform and wavelet transform.

Chapter 4

Baseline estimation

4.1 Baseline concept

An important task in the field of spectroscopy is the comparison of known signals with measurements obtained from an unknown sample. If the unknown and known signals show characteristics that are sufficiently similar and the characteristics are specific enough to describe the known sample with a high enough certainty then information about the unknown sample can be gained. Spectral characteristics are described by sets of intensities and position distributed over the spectral domain. Variable offsets, can reduce the effectiveness of automatic comparison methods severely. Using a metaphor one could say known signals are locks and measurements are keys, the task of identifying an unknown measurement is not unlike finding the lock which it matches. Measurement containing baseline distortions then represent bent keys and in order to fit the key the distortions introduced by bending have to be removed. This chapter presents several ways to estimate and remove baseline distortions based on the different concepts of model-building, polynomial fitting and conditional smoothing and discusses their respective strengths and weaknesses.

4.1.1 Spectral baseline definition

The baseline is the notional zero, the frame of reference from which all data is measured [MM97]. In an ideal scenario the baseline would simply be of zero intensity making signal intensities directly comparable, however experimentally obtained measurements often contain distortions that influence the correct reference frame. The task of estimating and correcting these distortions is called baseline estimation or baseline correction, respectively.

A human observer can usually identifiy the signals illustrated in figure 4.1 as identical except for a low frequency background signal; the correction for the changing offset is done automatically and almost subconsciously by the human brain's powerful pattern

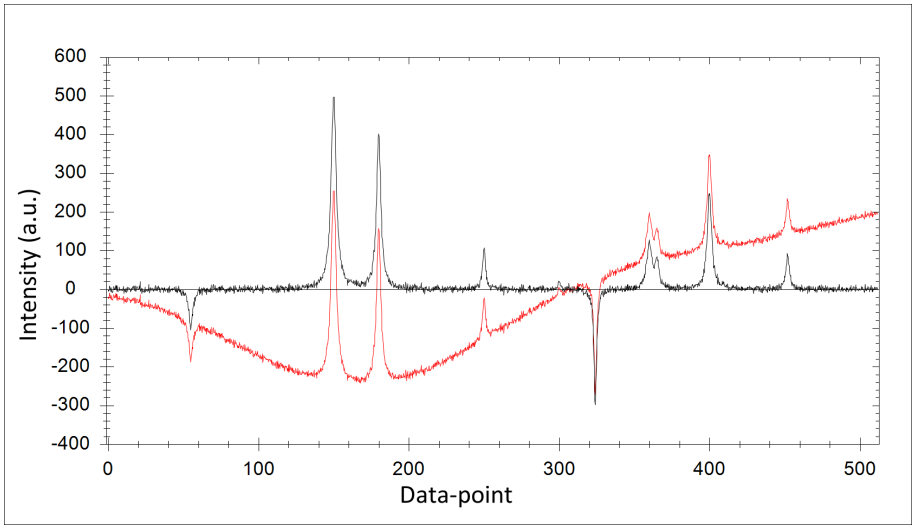

Figure 4.1: Visual comparison of a synthetic noisy signal influenced by baseline distortions and without baseline distortions.

recognition. However when an automated technique is comparing the intensity values of both signals point by point the analysis will generally not result in a classification that identifies both signals as similar or equal. Baseline estimation or baseline correction describes the process that enables the human observer to separate background and actual signal with the goal of enabling an automated analysis routine to perform the same correction before calculating mathematical measures that describe similarity of data-sets.

4.1.2 Baseline distortions

Baseline distortions have a variety of names; baseline drift, drift noise, background signal or simply baseline are synonyms for a class of problems that occur in many analytical techniques. The shared essence of these problems is an unwanted non-constant alteration to the measurement reference value by distortion effects. Without correction these distortions skew quantitative analysis and degrade overall comparability of measurements. The magnitude and sources of baseline distortions are highly dependent on the nature of the experiments, the measurement hardware involved in the data acquisition process as well as transformations or other processing steps used on the raw data. Common, known sources include instrumental factors like source intensity instability (flicker), detector response variations, temperature fluctuations as well as spatial correlations in the detection sensors or physical variations in the sample [BVMW00]. Generally a measurement y with regard to baseline estimation is modeled using

$$y = s + b + \varepsilon, \tag{4.1}$$

with the relevant information described by the signal features s, the background or baseline b and the noise distortions ε. Due to its multiple sources and often sample dependent characteristics baseline distortion can lead to baseline intensities that vary dramatically with the domain of a measurement and are impossible to estimate theoretically [PH96]. Some techniques possess unique characteristics which can lead to certain effects taking a predominant role as the cause of baseline distortions. Raman spectroscopy is an example of such a case. Experimental Raman measurements of biological materials are often influenced by intense fluorescence that can be orders of magnitude greater than the desired signal [LMJ03]. Since these fluorescence effects are so intense they dominate the background signal, baseline distortion and fluorescence are sometimes used as synonyms in the field of Raman spectroscopy.

Without a reliable physical or mathematical explanation for baseline distortions the most common and most important working hypothesis about baseline distortions is that they are smooth and vary relatively slow. Or in terms of frequency: Frequencies attributed to baseline distortions are found in the low end of the measurement frequency band. While there is no theoretical proof that supports this hypothesis it matches with observed measurements and is the accepted basis for all works in the field. Brown characterized baseline distortions more carefully as colored noise with low frequency dominance in the noise power spectrum, [BVMW00]. This implies that baseline distortions also appear in higher frequency bands but on the other hand affirms the working hypothesis by stating that their main energy is located in low frequencies. Regardless of the difficulties offering a comprehensive definition for exact baseline characteristics the removal or - to formulate more carefully- the reduction of baseline distortions is a central problem in post-acquisition data processing [JSY$^+$04] since it is a problem that is ubiquitous with spectroscopy [RJFD01]. Examining publications discussing the field of baseline correction or background removal approaches can be roughly categorized into four different groups : Manual selection, model extraction, polynomial fitting and conditional smoothing. These four categories are not official but they manage to grasp the basic differences of existing methods.

4.2 Manual selection and interpolation

Baseline correction via manual correction and interpolation is an option that is available in many commercial products. The idea of the method is to aid the user in finding a suitable baseline rather than estimating the baseline semi-automatically. For that purpose the system lets the user determine points or areas of the measurement to be corrected that show baseline distortions but no relevant signals. A complete baseline is calculated by interpolating the areas between those user defined regions. The method naturally allows several variations such as the interpolation method (e.g. linear, poly-

nomial or spline based) and the number of selected points or regions. The process can also involve an iterative component which allows adding and removing sample points after calculation of the initial baseline approximation. Results obtained by manual selection and interpolation are generally regarded good and most importantly very robust. Since the most important and critical task which is the selection of sample points and as such the separation of baseline and signal features is directly controlled by the user the margin of error that can be attributed to an algorithm is slim. The workflow usually contains three steps. Step one which might be the most important but also the most elusive, encompasses an expert user examining a measurement visually. The user has in depth knowledge about the measurement technique and is familiar with the instruments and uses that knowledge to form a model that separates the measurement into its components (baseline, noise and signal). In step two the user transfers the baseline component of that model by marking a number of sample points in the measurement that are suitable to describe it. Then in the final step the computer generated baseline model is compared to the user model and is either accepted or refined by modifying the sample points. The final result therefore is conform to the user belief and classified as correct or at least sufficiently close to the best known solution. Since the exact shape of a distorted baseline can generally not be determined in real measurements expert knowledge and implicit models based on that knowledge are the only grounds for comparison. In fact manual selection and interpolation usually is the standard against which automatic methods are compared and many mechanisms and ideas used for automatic baseline estimation techniques are based on experiences made with manual techniques and partly even mimic the workflow.

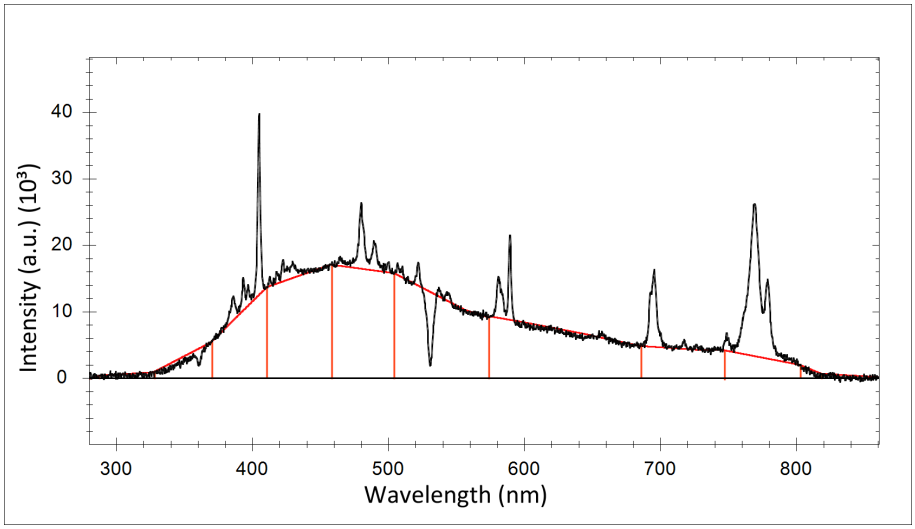

Figure 4.2: Example of linear interpolation between user-selected baseline points. Sample points and the linear interpolation are marked in red.

4.3 Model building and fitting

Baseline correction via model subtraction is a very effective and elegant solution to enhance analytical measurements. The general method uses a model of the signal and the baseline distortion or just the baseline distortion which is fitted to the actual measurement to eliminate the influence of baseline distortions. Examples for model based baseline correction include the approach by Phillips and Hamilton [PH96] who used two unspecified functions $N(v)$ and $W(v)$ to model the narrow line shapes and wide backgrounds of FT IR spectra respectively. Both functions are iteratively matched to the observed measurement while residual errors are allocated primarily to the wide baseline component and only those that cannot be attributed to the baseline are assigned to the signal component. The iteration stops if residual levels agree with a previously specified noise level. Additionally Phillips and Hamilton maximize the entropy between the two components to reduce the number of possible solutions generated by the iterative process. Boelens, Dijkstra, Eilers, Fitzpatrick and Westerhuis proposed a method to remove baseline distortions or background spectra in measurements acquired with different spectroscopic techniques [BDE$^+$04]. The proposed method uses principal component analysis to obtain a limited number of principal components which are used to describe the baseline or background. Asymmetric least squares regression is used subsequently to fit the model to the observed measurements in order to approximate contribution. Xu, Sun and Harrington proposed a method for Gas and Mass Spectrometry data that uses singular value decomposition to construct an orthogonal basis and employs a novel regularization parameter that prevents over-fitting [XSH11].

4.3.1 Background removal by Boelens

Boelens, Dijkstra, Eilers, Fitzpatrick and Westerhuis presented a model based method to eliminate the background spectrum (EBS) for liquid chromatography, infra-red and Raman spectroscopic measurements [BDE$^+$04]. The model building step uses a number of principal components which are extracted from a training set of spectra containing only background. Then asymmetric least squares regression is used in the model fitting step on data containing background and signal information with respect to the extracted background model.

Method

Let X_b $(n \times k)$ be the matrix of k spectra containing only background information so that each column contains one spectrum with n data-points. Principal Component Analysis (PCA) is used to model the matrix X_b with a number n_{pc} of principal components.

$$X_b = PK_b + E_b. \tag{4.2}$$

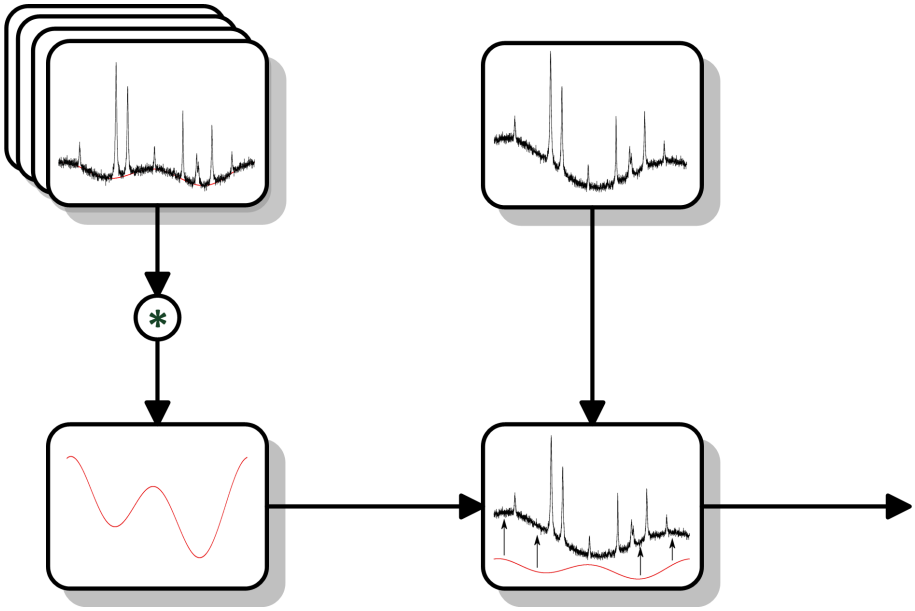

Figure 4.3: General schematic of baseline estimation via model fitting. Critical steps are model extraction and to a lesser degree model fitting. The $*$ operator indicates that multiple reference measurements are used to generate the baseline model.

The orthogonal basis that describes the variance in X_b is formed by matrix P. The authors call the space spanned by the column vectors in the P matrix $B - space$. K_p are the corresponding coefficients that allow reconstruction of the background spectra using the principal components in P. E_b represents the changes that are not explainable using the principal components and should only contain noise influences. To determine the number of components used in P the authors used the method described in [NP87]. Alternatives to determine the number of principle components are also given in [Mal02] and [Jac91]. Assuming that the extracted basis is also valid for further measurements every obtained spectrum can be described as

$$x_{ab} = Pq_{ab} + s_a, \tag{4.3}$$

with q_{ab} as the unknown coefficients modeling the background contained in x_{ab} and s_a representing the signal information. In the model described by equation 4.3 the authors disregard any noise contained in the measurement. The vector q_{ab} is estimated using asymmetric least squares regression assigning different weights to positive and negative errors. The asymmetric least squares algorithm minimizes the function

$$f = \sum_{i=1}^{n} w_i (x_i - P_i q_{ab,i})^2, \tag{4.4}$$

where

$$w_i = \begin{cases} p & \text{if } x_i - P_i q_{ab,i} > 0 \\ 1-p & \text{else,} \end{cases} \tag{4.5}$$

and p with ($0 < p < 1$) as the user controlled asymmetry factor, determining the degree to which positive residuals are given less weight than negative ones.

Advantages and disadvantages

The method by Boelens et al. is a straightforward model building approach. Principal component analysis is used on selected spectra containing only variations of the target model. PCA is used to decrease the data volume needed to store the model. The reconstruction is enhanced by infusing knowledge about the problem - residuals after baseline subtraction need to be positive - into the model fitting step while influences of noise are disregarded during the fitting. The most critical aspect is the selection of spectra used to build the B-space basis. One has to ensure that all possible variations of background behavior are included into the selection; otherwise the fitting method will fail and create false results.

4.3.2 Procedure by Phillips and Hamilton

Phillips and Hamilton proposed a method for baseline estimation and correction for FT-IR (Fourier transform Infra-red) spectra based on modeling functions for single peaks signals and baseline shape using a maximum entropy criterion for optimization [PH96].

Method

Phillips and Hamilton model an FT-IR spectrum as the sum of three components: a smooth low frequency baseline, the actual signal - in this case a set of narrow absorption peaks - and noise, which conforms to the model given by equation 4.1. Given a measurement $Y(v)$ containing N data-points the proposed method does not try to extrapolate the three components from a single measurement but iteratively simulates a noiseless version of the measurement using given functions of signal and baseline. Let $A_j(v)$ be the estimated positive, nonzero component signal spectrum and $N(v)$ the mathematical function representing the instrumental line shape. Further let $B_j(v)$ be the positive, nonzero component baseline spectrum and $W(v)$ a relatively wide background shape function. The noise free spectrum is described as

$$C_j(v) = B_j(v) \otimes W(v) - A_j(v) \otimes N(v), \tag{4.6}$$

the difference of the convolution of $B_j(v)$and $W(v)$ and the convolution of $A_j(v)$ and

$N(v)$ with $j = 0, 1, 2, ...$ indicating the iteration count. Note that the example used by Phillips and Hamilton contains only absorption signals.The residual error function is calculated by comparison with the original measurement. The description of the method states that errors are allocated primarily to $B_j(v)$ and the remainder being assigned to the component signal spectrum $A_j(v)$. The iterative process is repeated until the mean squared error between simulated spectrum and measurement agrees with a specified noise level $\sigma(v)$:

$$N = \sum_{j=1}^{N} \frac{(Y(v_i) - C_j(v_i))^2}{\sigma(v_i)^2}. \tag{4.7}$$

For a second optimization criterion Phillips and Hamilton propose maximum entropy. The entropy S is calculated from the two components of their model as

$$S \propto \left(\sum_{I=1}^{N} A(v_i) \ln \left(\frac{A(v_i}{R} \right) + \sum_{I=1}^{N} B(v_i) \ln \left(\frac{B(v_i}{R} \right) \right), \tag{4.8}$$

with R as a small arbitrary value. R is also the convergence target in case no signal is present. Phillips and Hamilton describe that the results obtained by their proposed method are insensitive to starting estimates $A_0(v)$ and $B_0(v)$ but that good estimates reduce the number of iteration needed.

Advantages and disadvantages

Phillips and Hamilton discuss the results achieved by their method using variable parameters of σ and line-width of $N(v)$. Line width is identified as the most critical parameter since it is used to separate baseline distortions from signal components. The spectra shown in the article appear to be very good natured. All peak signals are very similar possessing seemingly identical line width parameters and show no overlap at all. Furthermore noise visible in the example spectra appears to be minimal, not influencing the shape of peaks in any noticeable fashion. The article also does not mention the mathematical description of either the signal modeling function $N(v)$ nor the baseline modeling function $W(v)$.

4.3.3 Baseline correction method by Xu

The baseline correction method proposed by Xu, Sun and Harrington has been developed for data of gas chromatography (GC) and mass spectrometry(MS) data, [XSH11]. It uses singular value decomposition (SVD) to construct an orthogonal basis with is used to approximate the real baseline distortions found in measured spectra.

Method

Using SVD a matrix of background spectra X, which were sampled from measurement regions void of analytical signal, is decomposed into a scores matrix U, a diagonal matrix S and a loadings matrix V which represents the orthogonal basis

$$X = USV^T. \tag{4.9}$$

The baseline correction optimizes three parameters: the number of background vectors, the number of basis vectors and the background subtraction error threshold. To prevent overcorrection between the basis and the measurement Xu, Sun and Harrington introduce two error threshold regularizations named *Smartbaseline*1 and *Smartbaseline*2. *Smartbaseline*1 uses two criteria which are executed until the stop criterion $x_c \geq e_{thres}$ is reached, where

$$x_c = x - \lambda(xV)V^T, \tag{4.10}$$

with

$$\lambda = \frac{\bar{x} - e_{thres}}{\bar{x} - \bar{x_c}}. \tag{4.11}$$

$\bar{x}$ denotes the average uncorrected measurement, $\bar{x}_c$ the average of the corrected measurement and e_{thres} is the error threshold. For *Smartbaseline*1 equations 4.11 and 4.10 are executed until $x_c \geq e_{thres}$. *Smartbaseline*2 is executed as long as the difference between the smallest positive peak in the original measurement and the background at its position s is smaller than the error threshold. The stop criterion for *Smartbaseline*2 can be written as $s \geq e_{thres}$ with

$$s = x(i_{min}) - x_b(i_{min}), \tag{4.12}$$

and

$$i_{min} = min((x(i) - e_{thres})/x_b(i)), \tag{4.13}$$

where i represents the indices of all positive peaks in x_b. The calculation of λ for *Smartbaseline*2 is given by

$$\lambda = (x(i_{min}) - e)/(x_b(i_{min})). \tag{4.14}$$

The projections of the measurement on the orthogonal basis x_b is given by

$$x_b = (xV)V^T. \tag{4.15}$$

For *Smartbaseline*2, equations 4.14 and 4.10 are executed until $s \geq e_{thres}$.

To define the response surface the authors propose two possible methods: Average SNR or projected differential response PDR [CHH$^+$04]. The obtained response surface is then modeled by the polynomial function

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_1x_3 + b_6x_2x_3 + b_7x_1^2 + b_8x_2^2 + b_9x_3^2 \quad (4.16)$$

with the response y, the polynomial coefficients b_i and the variables x_i. The authors do not specify which method is used to fit the polynomial but mention that all calculations were performed using Matlab R2010b. The standard Matlab "polyfit" function uses the Vandermonde matrix and least squares. To create synthetic data the authors used the Gaussian function to simulate peaks as well as the background noise. The parameters for the background distortion, simulated by a single Gaussian function, are given in the article however the parameters for the peak signals are not. Figures given in the article imply that peak widths are orders of magnitudes smaller and constant intra as well as inter measurements. The equation used to simulate background and signals is given by

$$f(x) = a \cdot \exp\left(-\frac{(x-b)^2}{2c^2}\right). \quad (4.17)$$

The parameters of the Gaussian function are amplitude a, position b which describes the center of the peak, and the standard deviation or width parameter c. To simulate the effect of column bleeding the baseline amplitude parameter a was replaced by

$$f(t) = k + A\frac{1}{1+\exp(-(t-B)/C)}, \quad (4.18)$$

which is changing along the retention time t. k is a constant and similar to the Gaussian function, A the amplitude, B the center position and C the variable that determines the slope of the bleeding.

Advantages and disadvantages

The method described by Xu, Sun and Harrington appears to be more complex than most other baseline correction procedures due to the nature of combined GC/MS data. The estimated baseline is actually a base surface as the data is two dimensional however the essence of the procedure remains similar to other model based baseline correction approaches. The synthetic datasets used by the authors appear to be extremely simplistic compared to real GC/MS data sets. The synthetic datasets shown in the article contain four clearly separated Gaussian peaks. The simulated baseline is highly symmetric and equally sharp defined by the Gaussian function given in Equation 4.17 and

Equation 4.18. Noise appears to be nonexistent in the synthetic spectra. The achieved results are accordingly good rendering correction values very near to 100%. The authors have also tested their method on real GC/MS data. The figures given for the real data are unfortunately too small to visually convey a sense of effectiveness or allow for closer comparison of signal components. Furthermore the article contains a table displaying accuracy values obtain using the described baseline correction method on real measurements. The table itself however mentioned an original synthetic data set. Furthermore a measure of accuracy must somehow describe the distance between correct result and achieved result however for real measurements the correct baseline as well as the pure signal component of the measurement are generally unknown.

4.4 Polynomial fitting

Polynomial fitting is an approach used by several correction methods to eliminate baseline distortions. The general idea of all polynomial fit methods is - as the name suggests - the usage of a polynomial function to approximate the low frequency distortions in a given measurement. Polynomial fitting can be achieved via least squares polynomial regression, [KKMN98]. The mathematical description of the problem given a series of n points (x_i, y_i) and a polynomial of degree k, $y = a_0 + a_1 + ... + a_k x^k$, can be formulated in matrix notation as $y = Xa$ and written as minimization problem

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} \rightarrow 0, \tag{4.19}$$

where $a_0, ..., a_k$ are the polynomial coefficients which are to be fitted. Fitting a polynomial to a series of data-points that represent a combination of signal features, noise and baseline distortions will of course not approximate the baseline but the combination of all superimposed components. For that reason baseline estimation techniques which use polynomial fitting generally also employ some form of filtering mechanism that is aimed to minimize the influence of signal features.

Examples for methods using polynomial fitting include the baseline correction algorithm by Dietrich, Rüdel and Neumann which uses a fifth degree polynomial function in combination with constrains based on the first derivative and pseudo power spectrum to select data-points suitable for baseline approximation [DRN91]. Another example is the method by Lieber and Mahadevan-Jansen which iteratively fits a polynomial of degree n to the measurement [LMJ03]. Fitted values that are greater than the original value are reset to the measurement value ensuring that a baseline estimate stays below the measured data. The next iteration then fits the polynomial of degree n to the

resulting series instead of the original data. This process is repeated until convergence. Graphically this process can be described as iteratively cutting off positive peak signals of a given measurement until only the baseline remains. Gornushkin, Eagan, Novikov, Smith and Winefordner proposed polynomial fitting with a maximum degree of ten in combination with constrains using minor and major minima in the original data [GEN+03]. The constrains used in the algorithms by Dietrich et al. as well as Gornushkin et al. act as an automatic selection of relevant of suitable points (or rejection of not suitable points) with the goal to remove most non-baseline points from the fitting process similar to the manual selection methods but based on mathematical criteria instead of expert knowledge. The algorithm by Liber and Mahadevan-Jansen achieves similar results by the rejection of points that lie above the approximated baseline during the iteration process.

4.4.1 Baseline recognition by Dietrich

In 1990 Dietrich, Rüdel and Neumann proposed a baseline recognition technique for one and two dimensional NMR spectra based on peak recognition and polynomial interpolation [DRN91]. To eliminate peak signals that would lead to false results in the polynomial approximation the derivative and power spectra are used then the baseline it iteratively estimated by polynomial fitting.

Method

Dietrich identifies peak recognition as a key factor for the proposed baseline estimation approach. Before calculating the discrete derivative the measurement is treated with a standard mean filter to reduce noise influences which are generally amplified by derivative calculation. For measurements represented by series of N data-points Dietrich suggests a mean filter of width $w = N/256$. The filter mentioned in the original article can by described by

$$y'_i = \frac{1}{2w+1} \sum_{k=i-w}^{i+w} y_k. \tag{4.20}$$

Note that is this definition the actual window size of the filter is $2w + 1$. After filtering the discrete first derivate is calculated and from there the power spectrum. It is not specified how the power spectrum is calculated but from figures it can be assumed that the authors did not use Fourier transform but instead simply used $y_{pow} = y^2$ to create a data series containing only positive values. This would yield the described power spectrum by using the mean filtered data series as input of

$$\hat{y}_i = (y_i - y_{i+1})^2. \tag{4.21}$$

The threshold by which data-points belonging to signal peaks are separated from those not containing signal is determined by iteratively calculating the mean and standard deviation of the power spectrum. In each iteration step q the threshold is given by

$$\hat{y}_{thres} = mean(\hat{y}^q) + 3\sigma^q. \tag{4.22}$$

Points that exceed the threshold are dropped in the next iteration. The iteration is stopped when the set of points that exceed the threshold is empty. Before fitting a polynomial of a fixed degree of five to the points obtained during the iteration intensity values of dropped points are interpolated linearly. In a second iteration loop the linear interpolated sections are replaced by the values of the polynomial approximation. The second iteration loop is repeated until the residuals between two consecutive baseline approximations are below an unspecified tolerance value ϵ.

Advantages and disadvantages

The method proposed by Dietrich is optimized and tested for NMR spectra. Some assumptions made are not necessarily true for other spectra. The implicit degree of smoothing corresponding to the mean filter is not adjustable to varying noise levels between spectra. Mean and standard deviation are not robust against outliers and spectra of dense signal peaks. A fixed polynomial degree of five might not be sufficient for more complex baseline shapes. The peak recognition is dependent on sharply defined signal features and only allows minor signal overlap before breaking down.

4.4.2 Fluorescence removal by Lieber and Mahadevan-Jansen

Lieber and Mahadevan-Jansen proposed a method for the automatic subtraction of fluorescence in biological Raman spectra that is based on a modification of least-squares polynomial fitting. The method fits a polynomial to a measurement given as a series of data-points without the need for pre-filtering. Peak influences are eliminated by iteratively removing data-points that lie above the approximated polynomial baseline.

Method

Given a measurement y as a series of data-points (x_i, y_i) the method proposed by Lieber and Mahadevan-Jansen fits a polynomial directly to the given series without need for filtering or selection of specific areas. The polynomial degree is not fixed for the proposed method but Lieber and Mahadevan-Jansen suggest that for biological Raman degrees of four to five are reported to yield the best approximation results. The polynomial fit problem can then be described by equation 4.19 with $k = 5$. The initial fit approximates the measurement which is a combination of baseline and signal and noise. To reduce the influence of signal features, which are assumed to be exclusively facing

upward Lieber and Mahadevan-Jansen replace all data-points that exceed the estimated baseline by their estimated values. Let p^t be the polynomial baseline approximation in iteration t and $\hat{y}^t$ the modified spectrum then the data-points in $\hat{y}^t$ are defined by

$$\hat{y}_i^t = \begin{cases} y_i & \text{if } p_i^t > y_i \\ b_i^t & \text{else.} \end{cases} \tag{4.23}$$

The idea behind the iteration is to gradually eliminate high frequency peaks without the need for user dependent peak recognition. Lieber and Mahadevan-Jansen mention that in an ideal case the iteration process would stop when no more points need to be reassigned. Since reassignment only happens to those points that lie above the polynomial approximation that means that all data-points are beneath the polynomial. The authors further suggest that due to noise the iteration process can be stopped if the number of affected data-points converges.

Advantages and disadvantages

The iterative polynomial fitting method proposed by Lieber and Mahadevan-Jansen is very simple and does not require any a priori knowledge about signal features - other than upwards orientation - or additional information from derivatives or data transforms. However the model might be too simplistic for general purpose use. The assumption Lieber and Mahadevan-Jansen mention in the ideal case is erroneous. Approximating a series y containing N data-points by polynomial fit generally does not yield a polynomial curve $\hat{y}$ that fulfils the condition $\hat{y}_i \geq y_i$ for $i = 1...N$. In most cases the polynomial will possess several overshoots as well as undershoots since the least squares method does not weight positive and negative errors differently. The same observation also poses a problem in the non-ideal case. The proposed methods offer no solution to undershoots - regions where the polynomial lies below the target series - that happen not due to signal features but as part of imperfect approximation. In fact, since points above the polynomial approximation are replaced undershoots potentially amplify and become more and more intense with every iteration.

4.4.3 Baseline approximation by Gornushkin

Gornushkin, Eagan, Novikov, Smith and Winefordner proposed an automatic baseline approximation and correction method for laser induced breakdown spectroscopy (LIBS) and Raman spectroscopy that employs polynomial fitting of functions with polynomial degrees of one to ten, [GEN+03]. The algorithm uses so called major and minor minima to break down the spectrum into an optimal number of sections then pre-selects data-points in each section that are used for the polynomial approximation in order to eliminate the influence for signal features on the baseline estimation.

Method

The algorithm by Gornushkin et al. contains of two phases, the sampling phase which divides the measurement into a variable number of sections, approximates several baselines based on those sections and calculates their corresponding standard deviations based on selected data points and the correction phase which divides the measurement into the previously determined the optimum number of sections and calculates the optimal baseline approximation. The sampling phase starts with a maximum number of section $N_{max} = 64$, each subsequent iteration in the sampling phase reduces the number by 1 until $N = 1$, giving the phase a constant loop with 63 iterations. To pre-select points that are used for the polynomial fit Gornushkin et al. propose major and minor minima. Major minima are selected as the 6% points with the lowest intensity in each section. Let σ_{min} be the standard deviation of the major minima within a section then the minor minima are those pixels within that section that satisfy the condition

$$|y_i - \hat{y}_{min}| \leq 3 \cdot \sigma_{min}, \tag{4.24}$$

where y_{min} is a major minimum. The original article explicitly states "within three standard deviations of the major minima" suggesting that $\hat{y}_{min}$ is either the expected value of all major minima in the region or the major minimum with the largest differential in intensity to y_i. A polynomial of previously determined degree is then calculated for the entire measurement using the major minima of each section. Let $M^t = m_1^t, ..., m_k^t$ be the set of data-points containing all major and minor minima of all sections and $\hat{y}^t$ the polynomial curve in iteration step i of the sampling phase then σ_i is calculated according to

$$\sigma_i = \frac{\sum_1^k (m_i^t - \hat{y}^t)^2}{k}. \tag{4.25}$$

Each iteration of the sampling phase thus yields a value σ_i that is used in the correction phase to determine the optimal number of subsections for the final polynomial fit. The optimal number of subsection N_{opt} is the number corresponding to the smallest value σ_i. The final step approximates the polynomial using both, major and minor minima of the sections defined by N_{opt} .

Advantages and disadvantages

The method by Gornushkin et al. was tested on LIBS and Raman spectra containing 2048 each. Arbitrary numbers as the number of maximum subsection and could be directly related to the total number of data points and the frequency of usual signal features. Furthermore the choice of three times the standard deviation appears to be error prone given the small sample size of points (only two major minima per section

for $N = 64$ and 2048 total points) and the known effects of outliers on mean values. Since the algorithm is exclusively focused on minima any effects of artifacts in the original spectrum that introduce negative distortions or errors are critical and cause the method to fail.

4.5 Conditional smoothing

Filters aimed to remove high frequent influences from the original measurement can also be used to approximate baseline distortion. While smoothing alone usually does not yield satisfactory results several methods have been proposed that combine a set of constrains with smoothing techniques to ensure that the smooth curve remains true to the measurement and limits the influence of signal peaks on the baseline extrapolation. Xi and Rocke proposed a baseline correction method derived from a parametric smoothing model that uses a smoothing component in combination with a negative penalty to ensure that the smooth baseline approximation does not exceed the original measurement [XR08]. The parameters responsible for the influence of both components are derived from the noise standard deviation. Zhang, Chen and Liang proposed the adaptive iteratively reweighed Penalized Least Squares algorithm (airPLS) which enhances the baseline correction with asymmetric Least Squares smoothing method introduced by Eilers and Boelens [EB05] by adding a variable iteration number [ZCL10]. The method by Eilers and Boelens uses the Whittaker smoother in combination with asymmetric least squares which penalizes positive deviations from the original data to a higher degree than negative ones. While this asymmetry appears to be reversed at first glance closer inspection reveals that it actually ensures that that the smooth curve will not exceed the original data. Another example of conditional smoothing is the automatic iterative moving averaging (AIMA) technique for baseline correction presented by Prakash and Wei [PW11]. The algorithms works in two stages called iterative averaging and iterative averaging smoothing that are designed to mark peak areas an then replace the marked areas by linear interpolations between the remaining unmarked points that are assumed to be baseline points. The design of the iterative averaging step ensures that local minima are preserved thus the technique following the same idea of approximating the baseline from 'underneath'.

4.5.1 Algorithm by Xi and Rocke

The baseline correction technique proposed by Xi and Rocke is an example of parametric smoothing. It employs a score function that balances smoothness and asymmetric distance while locally weighted scatterplot smoothing is used to estimate noise in the measurement to automatically determine error tolerances for the model. Results are demonstrated for NMR spectra.

Method

The measurement model used by Xi and Rocke is similar to the general model given by equation 4.1 but extends the model by adding a second set of random errors that are signal dependent. A measurement y_i is represented using the baseline b_i, the signal μ_i and the random errors η_i and ε_i,

$$y_i = b_i + \mu_i e^{\eta_i} + \varepsilon_i. \tag{4.26}$$

Xi and Rocke futher postulate that ε and η are normally distributed with mean zero and variance 1 and auto-correlated. The baseline is estimated using the score-function

$$F(b) = \sum_i b_i - A \sum_i (b_{i+1} + b_{i-1} - 2b_i)^2 - B \sum_i (b_i - y_i)^2 g(b_i - y_i), \tag{4.27}$$

with g representing the Heavyside step that implements the asymmetric error function to treat the baseline differently depending on its relative position to the original measurement,

$$g(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else.} \end{cases} \tag{4.28}$$

The score function consists of two parts, the smoothness penalty term and the negativity penalty. The smoothness penalty implements the discrete integral of the quadratic second derivative thus penalizes strong curvatures in the baseline while the negative penalty implements the quadratic error between baseline and spectrum while only counting those errors where $b_i > y_i$. Both terms are weighted with a factor A for the smoothness penalty and B for the negative penalty. The optimal baseline b_{opt} represented by a set of points b_i that maximize $F(b)$ while choice of A and B in connection with the original data determines the baseline. The authors introduce another version of their score function that uses static factors A^* and B^* in combination with the noise standard deviation. The only variable factor in this score function is the noise standard deviation which is proposed to be estimated via locally weighted scatterplot smoothing. The adjusted score function takes the form of

$$F(b) = \sum_i b_i - \frac{CA^*}{\sigma} \sum_i (b_{i+1} + b_{i-1} - 2b_i)^2 - \frac{B^*}{\sigma} \sum_i (b_i - y_i)^2 g(b_i - y_i), \tag{4.29}$$

with $A^* = 5 * 10^{-9}$, $B^* = 1.25$ and $C = n^4$. The noise variance is estimated by dividing the measurement in regions of 32 data-points and uses the minimum of the lowess regression of variance values of each region. The noise standard deviation is assumed

as the square root of the estimated minimum variance. Note that the adjusted score function using the estimated noise standard deviation is robust against scaling by a scalar k under the assumption that $\sigma' = k\sigma$.

Advantages and disadvantages

The baseline correction method by Xi and Rocke only needs a single parameter, the noise standard deviation, which on first sight makes it very easy to handle. The suggested method for baseline estimation is plausible assuming that the lowest variance found in a spectrum is a good indicator for a constant noise variance of the entire spectrum. However the baselines presented in the article are relatively flat and the spectral features of the NMR measurements shown similar in each spectrum. The factors A^* and B^* could be and probably are different for other types of spectra. Furthermore the method assumes a constant degree of baseline curvature over the entire spectrum. A more interesting scenario would be a truly variable baseline and signal features that show significantly different characteristics.

4.5.2 Baseline correction by Eilers and the airPLS algorithm

The adaptive iteratively reweighted Penalized least squares (airPLS) algorithm for baseline correction proposed by Zhang, Chen and Liang [ZCL10] extends the algorithm by Eilers and Boelens [EB05]. Eilers and Boelens key idea is the use of the Whittaker smoother [Whi22] combined with asymmetric weights [Efr91], [NP87] to approximate the baseline of spectra with positive peak signals.

Method

Since the Whittaker smoother is a key aspect of the algorithms it's properties are summarized here, for more details on the Whittaker-filter's application for smoothing see section 3.4.3. Let y be the measurement represented by a series of data-points and let z be a series of data-points approximating y then the Whittaker smoother minimizes the penalized least squares function

$$S = \sum_i w_i (y_i - z_i)^2 + \lambda \sum_i (z_i - 2z_{i-1} + z_{i+2}). \tag{4.30}$$

The weight function w is formulated so that negative residuals - which correspond to z_i being greater than y_i - are assigned higher weights than positive residuals:

$$w_i = \begin{cases} p & \text{if } y_i > z_i \\ 1-p & \text{else.} \end{cases} \tag{4.31}$$

Usual choices for p and λ are in the ranges $0.001 \leq p \leq 0.1$ and $10^2 \leq \lambda \leq 10^9$ respectively. The problem contains a mutual dependency of smooth function and weights that can be solved iteratively. Given a measurement y and variables λ and p the following Matlab code computes the baseline approximation proposed by Eilers and Boelens in ten iterative steps, see listing 4.1.

Listing 4.1: Eilers' Matlab code 2005

```
function z = baseline(y, lambda,p)
%Estimate baseline with asymmetric least squares
m=length(y);
D=diff(speye(m),2);
w=ones(m,1);
for it =1:10
    W= spdiags(w,0,m,m);
    C= chol(W + lambda * D' * D);
    z= C \ (C' \ (w .*y));
    w= p*(y > z) +(1 - p) * (y < z);
end;
```

AirPLS uses the same core algorithm but extends it by implementing an alternative weight function w and an exit criterion that stops the iteration process if the absolute sum of negative residuals becomes smaller than a threshold value. The airPLS weight function is given by

$$w_i^t = \begin{cases} e^{\frac{t(x_i - z_i)}{|d|}} & \text{if } y_i < z_i \\ 0 & \text{else,} \end{cases} \tag{4.32}$$

while the termination criterion is described as

$$|d| < 0.001 \cdot |x|. \tag{4.33}$$

The vector d contains the negative residuals between the original measurement and the fitted curve z in the previous iteration step. This means that the algorithms stops if the absolute value of negative errors - original data-points below the estimated baseline - is small in relation to the number of data-points or if the predetermined maximum number of iterations is reached. The airPLS algorithm also accepts a user defined order or difference which is used in the calculation of the smoothness measure. The default order is 2 meaning that the second derivative is used for the calculation of the smoothness parameter and the default maximum number of iterations is 20.

Advantages and disadvantages

The use of the Whittaker smoother with iteratively fitted weights gives Eilers' baseline correction algorithm and its variant airPLS the ability to adapt to a given measurement. Zhang, Chen and Liang demonstrated the algorithms effectiveness using numerous examples of real and synthetic spectra. The key point of the approach remains the parameter λ which controls the balance between smoothness and fidelity terms of the weighted penalized least squares problem. This parameter needs to be tuned by hand and is absolutely essential in order to obtain good results. Furthermore both smoothness and fidelity parameter are global, meaning they are the same for every data-point in the spectrum which implies constant baseline and noise parameters across the spectral domain.

4.5.3 AIMA algorithm by Prakash and Wei

The Automated iterative moving average (AIMA) algorithm for baseline correction proposed by Prakash and Wei uses linear interpolations between automatically selected measurement segments to approximate baseline distortions [PW11]. The algorithm uses two stages in which peak signals are first identified and then maximized.

Method

The algorithm uses two steps called iterative averaging (IA) and iterative averaging smoothing (IAS). Given a measurement y represented by a series of points $y_i, i = 1...N$ the iterative averaging step recalculates even and odd data-points separately according to

$$y_{i+1} = min(y_{i+1}, (y_i + y_{i+2})/2). \tag{4.34}$$

The procedure is repeated $k = N/4$ times for even and uneven data-points, dropping the first and last data-point in each iteration. The iterative averaging smoothing step is a modification of the first step. It uses the same separation into even and odd points and the same scheme of iteratively dropping data-points from both ends of the interval. The equation

$$y'_{i+1} = \frac{(y'_i + y'_{i+2})}{2} \tag{4.35}$$

describes the intensity updates for each iteration and is done separately for even values of i and for odd values of i analogous to step one. After the averaging part of step 2 is finished selected segments of the updated spectrum are replaced by linear interpolations between the leftmost and rightmost data-point of the segment. The segments are defined as interval where $y_{dif} = 0$,

$$y_{dif,i} = |y_{max,i} - y_{min,i}|, \tag{4.36}$$

$$y_{max,i} = max(y'_i, y_i), \tag{4.37}$$

$$y_{min,i} = min(y_{max,i}, y_i). \tag{4.38}$$

For the calculation of $y_{dif,i}$, $y_{max,i}$ and $y_{min,i}$, even and odd indices are no longer treated separately so that $i = 1, 2, ..., N$ for the corresponding equations. The result of step 2 of the AIMA algorithm is a modified array y'. Step 2 is repeated with $y''_t = min(y, y^{'})$ until the termination criterion given by

$$\frac{\sum_i |y''_{i,t-3} - y_i|}{\sum_i |y''_{i,t-2} - y_i|} > \frac{\sum_i |y''_{i,t-1} - y_i|}{\sum_i |y''_{i,t} - y_i|}, \tag{4.39}$$

is satisfied, where t indicates the iteration number of step two.

Advantages and disadvantages

The AIMA algorithm is presented as a fully automatic method for baseline correction and works without any parameters. The authors compare their algorithm with airPLS, the asymmetric least squares method by Eilers - which is the basis for airPLS - and the parametric method by Xi and Rocke. The numbers given by the authors suggest that AIMA is overall superior in performance and additionally does not require fitted parameters. The achieved results are however difficult to confirm as the reasoning behind the different steps and the separation of even and uneven data points is not discussed.

Figure 4.4 which is taken from the original article [PW11] contains visible hints at the weak points of the method. While the linear approximation agrees with well with the human impression of the correct baseline in the left region of the shown measurement the high frequency behavior between channel 6700 and 6900 does not agree with the common assumption of a slowly changing and smooth baseline. Furthermore the data series in 4.4 appears to be almost free of noise and all other samples shown in the original article possess very favorable noise characteristics. Since the algorithm is based on minor and major minima which naturally increase with higher noise intensities the performance under such conditions is unclear. Like other baseline estimators AIMA is by design not able to deal with negative distortions or errors.

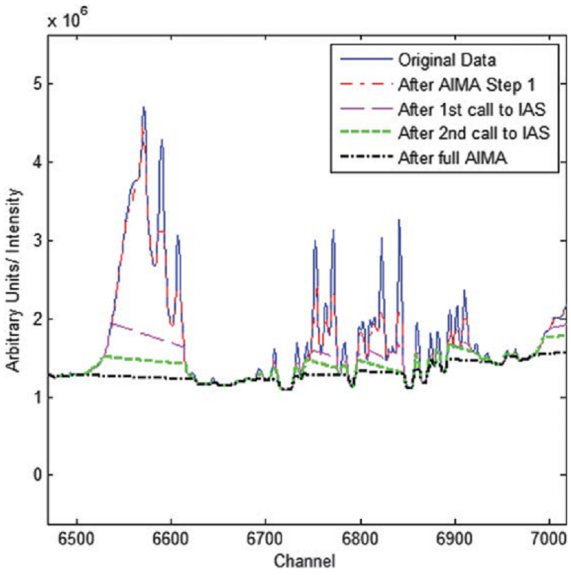

Figure 4.4: Example of AIMIA baseline correction

4.6 Discussion of estimator characteristics

Baseline approximation and correction algorithms possess specific characteristics depending on their class - model based, polynomial or conditional smoothing - but also shared weaknesses that are inherent to the problem that is to be solved and the available information to do so.

4.6.1 Manual selection

Manual selection is perhaps the most powerful baseline correction methods available but it has several inherent weaknesses. The main drawbacks of the approach are unsurprisingly directly linked to its dependence on user interaction. It is time very consuming in larger scales. While single measurements are often unproblematic regarding time consumption even relatively small quantities quickly become expensive when done by a properly trained expert. Another main drawback is the requirement of expert knowledge. Since the method itself relies on the user to determine which data points within a measurement are suitable for baseline estimation the result is naturally dependent on the quality of the chosen points. As such the procedure is only advisable for users possessing in depth knowledge of measurement and acquisition device physics and characteristics. Another potential drawback that is directly related to the usage of expert knowledge is the subject of bias. Ten different experts might extrapolate ten different baselines for a single data set. In practice these baseline are usually very similar but for particularly difficult measurements results could alter in key aspects. The prevalence of manual selection methods in the field of baseline correction can also be interpreted

as a testimony to the difficulty of the general problem of identifying and eliminating baseline distortions.

4.6.2 Model functions

Baseline correction using model functions are naturally dependent on the quality of the model or simply the general availability of a model. Since the processes that lead to baseline distortions are most often unknown or at least too underdetermined to create a reliable model the area of application for model based baseline correction methods has to be carefully researched and tested. Faulty models or simply models that cannot express the full range of possible baseline variations inevitably lead to baseline estimations that introduce errors into a corrected signal. Factors that can complicate the generation of suitable models are for example changing environmental conditions such as temperate, humidity, atmospheric pressure, intensity of sunlight or wind which might influence the monitored reaction and/or the technical instruments. Variations in analyte composition or impure analytes that represent a mixture of different substances pose another considerable risk for model based methods if the model baseline is dependent on the examined analyte. If analytes create fundamentally different baselines then an encompassing baseline model becomes unspecific and faces the same problems as polynomial fitting while a separation into different model functions creates the need for methods that determine the correct model for unknown measurements. Furthermore model functions need to be trained or extracted usually using a number of measurements that represent pure background, which is not applicable for every measurement technique. Summarizing one can say that model functions for baseline estimation can work very well for selected techniques and controlled environments but are not a general solution to the problem of baseline distortion.

4.6.3 Polynomial fitting

Polynomial fitting suffers from two major drawbacks. Since the degree of the polynomial used to fit the baseline needs to be known in order to perform the fitting the right choice of becomes crucial. Low grade polynomials cannot approximate necessary baseline changes sufficiently while high grade polynomials produce excessive over- and undershoots that distort and corrupt the measurement. Most methods using polynomial fitting accept this drawback arguing that the polynomial degree is pseudo-fixed for specific applications. The fixed polynomial degree can be seen as a class of modeling functions similar to predetermined models used in model function approaches. The main difference is given by the fact that a fixed polynomial degree usually allows greater degrees of freedom than a model function that was extracted from actual measurements in a pre-processing step.

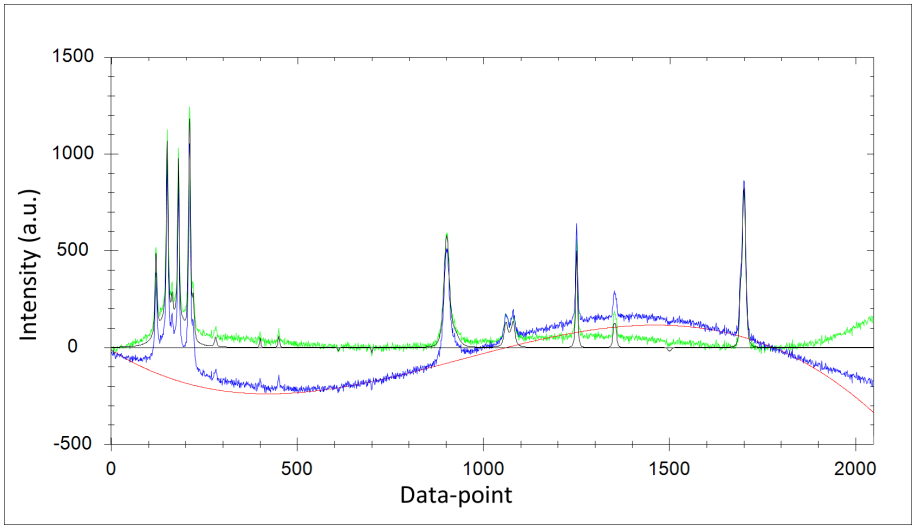

Figure 4.5: Illustration of polynomial baseline estimation using the method of Lieber and Mahadevan-Jansen. Shown are a synthetic measurement with 2048 data-points containing baseline distortions (blue), the estimated baseline using the method by Lieber and Mahadevan-Jansen (red) and the corrected synthetic measurement(green)

The focus of most polynomial fitting methods for baseline correction is the treatment of signal features that hinder the correct baseline approximation. The problem of signal and baseline forming a mixed signal is common to all baseline estimation methods. In the case of polynomial fitting the polynomial needs to be fit to the baseline while only information about the composite signals in available. Dietrich et al. determine which points are suitable prior to polynomial fitting while Lieber and Mahadevan-Jansen approach the problem iteratively introducing the additional condition that data-points of lowest intensity must belong to the baseline - note that the condition is not explicitly stated in the original article but the iterative design of "cutting off" from the top clearly implies the idea. Gornushkin et. al. also use an iterative method based and the low intensity baseline claim but in their approach the number of iterative results is fixed and each result rated individually allowing the algorithm to escape local optima in favor of the global optimum. Assuming that all signals are positive peaks - directed upwards - is a common constraint which is backed up by conditions found in real measurements. Signals have either emission character or absorption character but usually only one type appears in a measurement of a given technique so methods assuming all positive peak signals can be adjusted to perform in a similar matter when dealing with measurements containing only negative peaks. However, distortions and features introduced by filters can introduce comparatively high frequency negative features into a measurement containing otherwise only positive signal peaks. These features have a sometimes dramatic deteriorating effect on baseline estimation methods. In the case of polynomial fitting single errors are often unproblematic as the entire

baseline is approximated as a whole. However adding an iterative estimation that assumes only positive peaks will gradually push the baseline down. Figure 4.6 shows an example of the effect caused by a negative distortions on the method by Lieber and Mahadevan-Jansen. The polynomial for $i = 1$ offers a rough estimation of the overall trend of the curve, the influence of datapoints above the curve is removed and following esimates are only influenced by the remaining data and interpolated data values. As a result the negative distortion gradually pushes the estimate down.

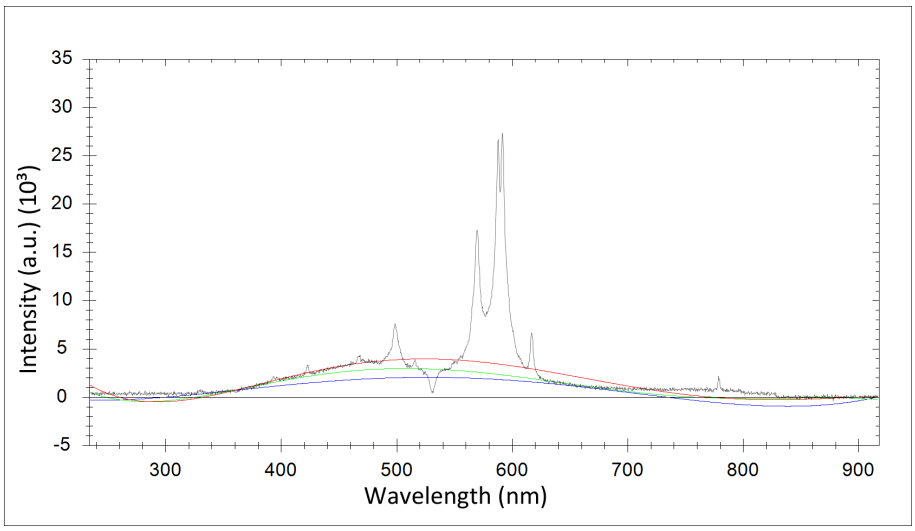

Figure 4.6: Effects of negative distortions on baseline estimation using the method by Lieber and Mahadevan-Jansen. Shown are the original data and three baseline estimations at different iteration counts $i = 1$, $i = 5$ and $i = 15$ and constant polynomial degree $d = 6$.

As baseline estimation and the separation of signal features from baseline features are inherently circular problems, an iterative approach is categorically more reasonable than a pre-selection without feedback information. The reason for this is simple assumption: If a sufficiently accurate separation of baseline and signal is possible without actually estimating the baseline the estimation would be redundant. Instead of estimating the baseline it would then be more sensible to simply select all signal features and separately set their minimum intensities - assuming positive signals - to zero to obtain comparable intensities. Iterative approaches on the other hand offer a more robust estimation although they also introduce new risks. The main drawback of polynomial fitting is the necessity of a fixed, global polynomial degree which makes over- and undershoots nearly inevitable in complex scenarios. Furthermore while baselines are assumed to be smooth they are generally caused by a combination of numerous sources, which implies that a single polynomial function with constant parameters for the entire measurement domain might not be the best solution. Figure 4.7 illustrated the effect of imperfect polynomial fit. The polynomial with degree of $d = 8$ contains

several intense over and undershoots as the polynomial degree appears to be too high for the actual baseline. The effects are less intense in the case of a polynomial degree of $d = 5$ but nonetheless still occur. Often these oscillations introduce low frequency errors into otherwise region flat regions. For practical purposes the oscillations are usually ignored if their intensity is close to the noise intensity.

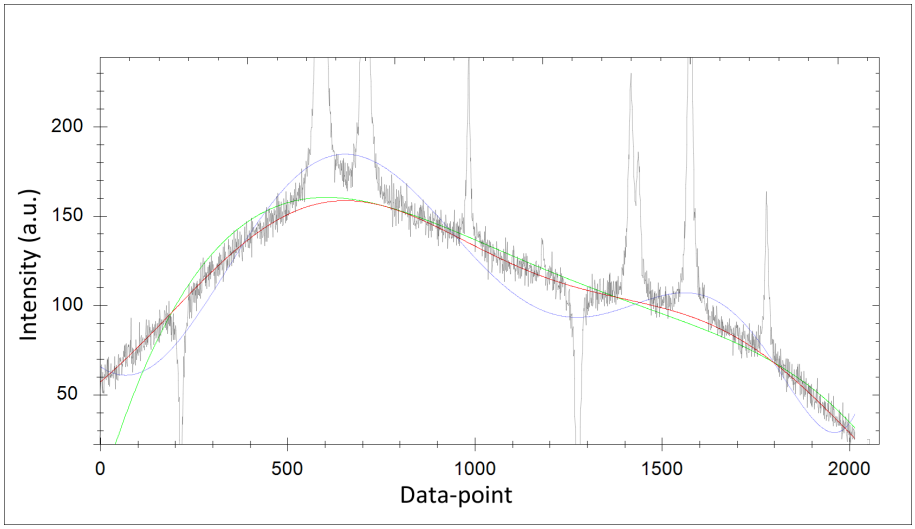

Figure 4.7: Illustration of over and undershoots as typical errors introduced by polynomial approximation. Shown are a synthetic data set, two polynomial approximations with degrees $d = 5$ (green) and $d = 8$ (blue) and the underlying synthetic baseline created using scheme 1 (red) as a reference.

4.6.4 Conditional smoothing

Conditional iterative approximation methods do not require a derived model or a polynomial function to fit the baseline but use the measurement itself to build a baseline model. Different smoothing techniques or other methods that possess low pass characteristics are used to remove signal features leading to a variable model. Similar to some polynomial fitting methods conditional smoothing methods work under the condition that the baseline can be approximated from below, which implies that signal features are exclusively positive. The algorithm by Xi and Rocke as well as the algorithm by Eilers and its extension airPLS work by balancing baseline smoothness and distance to the measurement. Asymmetric least squares which weights distances between estimated baseline and original measurements differently depending on whether the baseline is above or below the measurement is used in both case to ensure the baseline follows the bottom trend and is not effected by positive peaks. Similarly the AIMA algorithm iteratively cuts off positive peak signals. The principal of cutting off peaks resembles the algorithm by Lieber and Mahadevan-Jansen but instead of a

polynomial function that approximates the entire baseline AIMA uses piecewise linear interpolations. AIMA has no parameters making it - at first glance - the only fully automatic baseline approximation method; however the algorithm uses affixed smoothing parameter in its iterative averaging step and also in the iterative averaging smoothing step. Furthermore none of its internal parameters adapts to the treated measurement, thus any method can be made fully automatic by simply fixing all variable parameters to constants. The performance of the so created fully automatic methods of course depends strongly on the treated measurement. Apart from its non-adaptive nature major problems of AIMA are the non-smooth nature of piecewise linear interpolations and the already observed high frequency nature of the estimated baseline seen in 4.4. Both points are in direct conflict to the general assumptions on baselines, namely that they are of low frequency and smooth.

The algorithms by Xi and Eilers by definition always produce smooth baseline estimation. The degree of smoothness depends on the parameter balancing the two criteria used by both methods but even for parameters heavily favoring baselines that are very close to the original measurement the resulting estimation may contain higher frequencies but will never result in a rough or angular baseline. The main problem for both methods is the determination of the most optimal balancing factor between smoothness and fidelity criteria.

Xi proposes a method that determines the balancing automatically by simply inputting the noise standard deviation but examples are only given for extremely well behaved NMR measurements which possess very flat baselines and sharp signal features. The results discussed by Xi mention weights of the order of 10^7 for the smoothness property and 10^{-4} for the distance property which coincides with the trend towards very low frequency baseline estimations. The airPLS algorithm was tested using a wider range of measurements in the article by Zhang et al. For the tests the balancing parameter was determined by hand and the achieved results compared to similar correction methods that were also hand tuned to the data. The tests show good results given a fitting balance parameter. The adaptive weights allow the airPLS method to ignore signal features and automatically interpolate the so created 'holes' with a smooth curve. Challenging scenarios for the algorithms by Xi and Eilers (and by extension also airPLS) are broad peaks or superposition of many signals manifesting as a singular, wide feature in combination with non-flat baselines. Usually the problem of a baseline following wide peaks too closely can be addressed by increasing the balance parameter λ, however this also influences the general behavior of the baseline estimation. Figure 4.8 illustrates the effect of strict smoothness demands, $\lambda = 10^6$. The baseline estimate does not follow wide peak but also loses the ability to appropriately follow the overall baseline shape. The selection of the balance parameter is critical as it affects the estimated curve in a similar fashion as the polynomial degree does for polynomial fit

approaches. Strict smoothness demands generally tend to generate estimates that do not follow the original data close enough while relaxed smoothness demands can lead to estimates that follow signal features too closely.

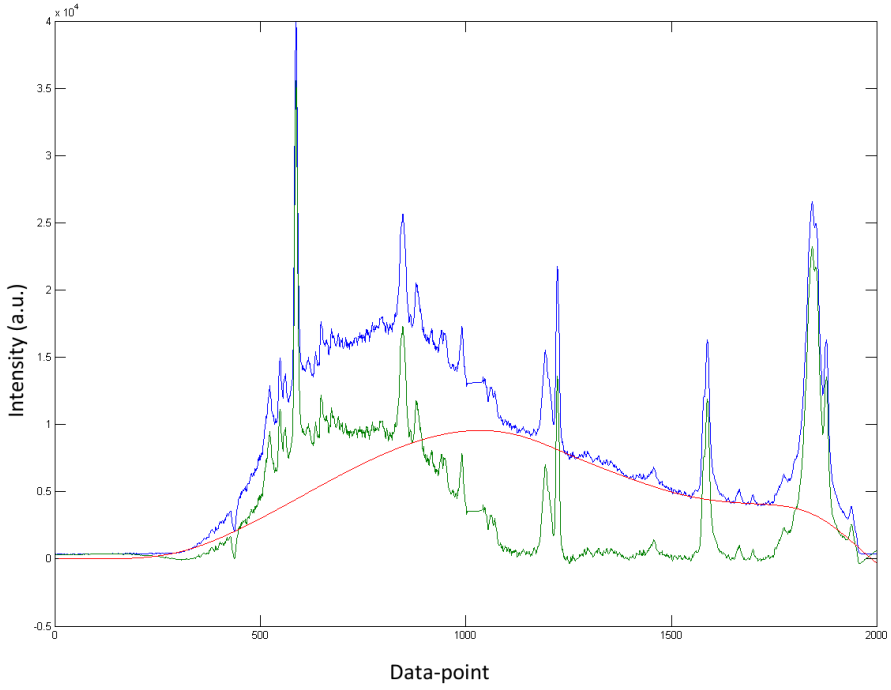

Figure 4.8: Illustration of strict smoothness factors used in airPLS that can lead to false approximations if baselines are distinctly curved. Shown are the original data, the baseline estimate via airPLS using $\lambda = 10^6$ and the corrected data.

Given the wide variety of possible signals and baseline shapes it is unrealistic to assume a baseline technique to be fully automatic and adaptive to all scenarios. The distinction between signals, noise and baseline distortions is extremely dependent on the measuring technique and measuring instruments and observed features that are classified as baseline in one case may be regarded completely different in a different instrumental setup. The goal of good estimation techniques must be to offer an intuitive and easy to adjust mechanisms to adapt to different conditions and reduce the influence of hard constrains to offer enhanced robustness.

Chapter 5

Wavelet transform

Wavelet transform is a powerful tool that can be used for sampled data of various sources. Wavelet transform describes a family of transformations which use basis functions, so called wavelets, to decompose data into distinct subspace scales of different detail [BGG98]. In contrast to Fourier transform, wavelet transform uses finite basis functions and thus retains a sense of local information in the transformed data. Scales are nested satisfying the multi-resolution analysis requirement, meaning that the space that contains finer scales also contains larger, coarser scales. Since the characteristics of the transform are a key component in the treatment of modulated noise the basics of wavelet transform are presented in this chapter. Wavelet theory is based on scaling spaces that are defined by scaling functions which are nested,

$$V_j \subset V_{j+1} \, \forall j \in \mathbb{Z}. \tag{5.1}$$

In the continuous case the finest scaling space is $V_{-\infty} = \{0\}$ and the coarsest $g \in V_{+\infty}$, with g representing the regarded function in its entirety. In the discrete case the finest scales usually span two sample points while the coarsest scale spans the entire signal. The wavelet transform of a given signal uses a set of functions $\psi_{j,k}(t)$ called wavelets, that span the differences between the spaces spanned by the scaling functions. Each scaling space V_j can thus be described by the scaling space V_{j-1} and the wavelet subspace W_{j-1} which describes the differences between V_j and V_{j-1},

$$V_j = V_{j-1} \oplus W_{j-1}. \tag{5.2}$$

By extending this idea one can transform the relation of cascading scaling subspaces to render

$$R = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_{n-1} \oplus W_n, \tag{5.3}$$

with R denoting the space of presentable functions, V_0 denoting the coarsest scaling

space and W_j the wavelet space corresponding to scale j. A representation of the discrete wavelet transform of a signal $f(t)$ is closely related to the above relation of subspaces and can be described by

$$f_k(t) = c(k) \cdot \phi(t-k) + \sum_j d_j(k) \cdot 2^{j/2} \psi(2^j t - k), \tag{5.4}$$

with

$$f(t) = \sum_k f_k(t). \tag{5.5}$$

$\phi(t)$ is called the scaling function with the corresponding coefficients $c(k)$ while $\psi(t)$ represents the wavelet function with corresponding coefficients $d_j(k)$ and $k \in \mathbb{Z}$ describes the translation along the t axis. In practice this means that given a specific basis function or mother-wavelet a function f can be fully characterized by the coefficients given in $c(k)$ and $d_j(k)$, similar to the representation as coefficients of the discrete Fourier transform.

A common application of discrete wavelet transform can be found in the field of data compression, for while the DWT does not change the amount of data the coefficients are easier to compress since often relevant information is stored in only very few coefficients. However the DWT is not shift invariant, meaning that shifting a given signal and then calculating the DWT will result in a set of coefficients that is different to those rendered by the DWT of the non-shifted signal and that no simple or intuitive relation between the coefficients of the shifted and non-shifted signal can be given. It is however possible to compute a shift invariant wavelet transform at the cost of additional computational complexity $O(N \log(N))$ instead of $O(N)$ [BGG98], [Bey92], [She92]. This variant of the DWT is called stationary discrete wavelet transform (SWT), redundant discrete wavelet transforma(RDWT). Please note that shift invariant in this context does not mean that the transformations of shifted and non-shifted signal are identical, but that the values of the coefficients shift in relation to the signal without changing unexpectedly. In contrast to the DWT the SWT of a signal S contains redundant data as each scale regardless of coarseness contains exactly n coefficients, where n is the number of sample points of S. While the additional redundant information added in the transform process makes SWT less suited for data compression it also adds robustness, thus reducing the risk of introducing unwanted effects by operating on the coefficients [Dau92]. Also Mallat's scheme to approximate the continuous wavelet transform is equivalent to the algorithm à trous, which describes the calculation of the redundant wavelet transform [She92].

One major advantage of wavelets compared to Fourier transform is the fact that wavelet coefficients remain localized. In other words, while the sense of locality is lost during Fourier transform and only regained by inverse Fourier transform wavelet trans-

form coefficients only effect ranges that are defined by the span of the corresponding wavelets. This means that the effects of a localized signal within in the original spectral range effects only wavelet coefficients that add to that particular area. As a result information in finer detail scales, where wavelet span small areas, is highly localized while coarser scales are harder to localize as they are span larger areas of the original spectral range.

5.1 Wavelet theory

What is known as wavelet transform today is the combination of methods and ideas that have been developed in parallel in the different fields most notably mathematics and physics over years. One of the most well-known sources detailing the underlying ideas and mathematical theory of wavelets was formulated by Ingrid Daubechies in 1992 [Dau92]. Wavelet based methods are a mathematical tool similar but not identical to Fourier analysis and as such are used in a variety of different fields such as the natural sciences biology, chemistry and physics and many applied disciplined such as economics, engineering and computer science. Wavelet transform describes data in terms of set of (often orthogonal) basis functions named scaling function ϕ and wavelet function ψ. Wavelet transform is closely related to the better known Fourier transform - or to be precise to the Fourier series. There are terminological differences between Fourier and wavelet transforms that are mostly due to the fact that wavelet transform was developed in scientific fields which offer different views on similar problems. The method generally known as discrete wavelet transform actually describes a series expansion similar to the Fourier series while the discrete time wavelet transform transforms a discretely sampled signal into a set of coefficients analogous to the discrete Fourier transform. The relationship between discrete wavelet transform and discrete time wavelet transform is equally close as that between Fourier series and discrete Fourier transform. Signals are expanded using the discrete wavelet transform but the implementation uses the discrete time wavelet transform. The analogies between discrete and continuous input and output can be given as a table [BGG98].

	Discrete Time	**Continuous Time**
Discrete Frequency	Discrete Fourier Transform	Fourier Series
Continuous Frequency	Discrete Time Fourier Transform	Fourier Transform

Table 5.1: Fourier transforms for discrete and continuous input and output

Discrete wavelet in table 5.2 indicates discrete values for the scaling and translation parameters. Since wavelets and wavelet transform has been developed in parallel in numerous scientific fields there are several different views on the subject. It is possible

	Discrete Time	**Continuous Time**
Discrete Wavelet	Discrete Time Wavelet Transform	Discrete Wavelet Transform
Continuous Wavelet	Discrete Time Continuous WT	Continuous Wavelet Transform

Table 5.2: Wavelet transforms for discrete and continuous input and output

to start with a continuous representation for the theoretical background of wavelets and later incorporate the filter bank approach that efficiently calculates and interprets transform coefficients offering a more practical oriented view sometimes preferred by electrical engineers. Both approaches use the concept of resolution which in terms of wavelets is represented by the scaling function ϕ. Scaling functions and wavelet functions - which are derived from scaling functions -are essential parts of wavelet transform similar to the sine wave function in the context of Fourier analysis. However other than in the Fourier context which is clearly defined since there is only one sine function wavelet transform can be seen as a family of transformations distinguishable by different scaling and wavelet functions.

5.1.1 Signal spaces

To talk about signals that are represented by scaling and wavelet functions some terminology from the field of functional analysis is needed. An important vector-space in signal processing is the Lebesque space $L^2(\mathbb{R})$, the space of all square-integrable functions which also represents the only hilbert space in the class of L^p spaces. The inner product of tow function $f(x)$ and $g(x)$ in $L^2(\mathbb{R})$ is defined as

$$\langle f(x), g(x) \rangle = \int f^*(x) g(x) dx, \tag{5.6}$$

where $f^*(x)$ denotes the complex conjugate of f. An expansion set for a vector space V is a set of functions $\phi_k(t)$ if all functions in V can be expressed as $f(t) = \sum_k a_k \phi_k(t)$, where a_k are coefficients. A basis is a an expansion set that is also unique. Regarding the same context from the other direction the space that is defined by a set of functions is called the span of a basis or expansion set. The closure of an expansion set includes all signals that can be expressed by linear combinations of the expansion set and also those that are the limit. Given a set of scaling function that is defined in terms of a basic scaling function and an integer k denoting a translation

$$\phi_k(t) = \phi(t - k), \tag{5.7}$$

with $\phi \in L^2$. The subspace related to this set of functions is defined as

$$V_0 = closure\,\{\phi(t)\}\,. \tag{5.8}$$

Since $\phi \in L^2$, V_0 describes a subspace of $L^2(\mathbb{R})$. To talk about multi-resolution a two dimensional set of functions is generated by scaling and translation of the basic scaling function.

$$\phi_{j,k} = 2^{j/2}\phi(2^j t - k), \tag{5.9}$$

leading to the subspace V_j over k

$$V_j = closure\left\{\phi_{j,k}(t)\right\}, \tag{5.10}$$

for $k \in \mathbb{Z}$. This means that all functions in V_j can be expressed as

$$f(t) = \sum_k a_k \phi(2^j t + k). \tag{5.11}$$

For different j the size and translation steps of $\phi_{j,k}(t)$ differs leading to representations in different detail. Wider scaling functions are equivalent to a coarse representation of information while smaller functions can represent finder details. Furthermore the spaces V_j are nested, meaning that each higher resolution space also contain those of lower resolutions,

$$V_j \subset V_{j+1} \quad \forall j \in \mathbb{Z}. \tag{5.12}$$

For the limits $j = \infty$ and $j = -\infty$ the spaces V_j are defined as

$$V_\infty = L^2,\ V_{-\infty} = \{0\}. \tag{5.13}$$

5.1.2 Scaling function

Using the definition of scaling functions the statements made for spaces can be also used to formulate dependencies for $\phi(t)$. Since

$$f(t) \in V_j \ \Leftrightarrow\ f(2t) \in V_{j+1}, \tag{5.14}$$

each $\phi(t)$ can also be expressed by shifted and weighted $\phi(2t)$,

$$\phi(t) = \sum_{n\in\mathbb{Z}} c(n)\sqrt{2}\phi(2t - n)\ , \tag{5.15}$$

where $c(n)$ represent the vector of scaling function coefficients while $\sqrt{2}$ is used to keep the norm of the scaling function constant. Different wavelet families satisfy equation 5.15 for different $c(n)$ e.g. the Haar wavelet scaling function uses $c(0) = 1/\sqrt{2}$ and $c(1) = 1/\sqrt{2}$.

5.1.3 Wavelet function

As a benefit of defining scaling functions and their properties in detail, motivation of wavelet functions becomes very simple. Wavelet functions or wavelets $\psi_{j,k}(t)$ can be described as a set of functions that span the differences between the spaces spanned by the scaling functions of neighboring scales. Wavelet and scaling function do not need to be orthogonal but in practice they usually are designed to satisfy the orthogonal criterion since calculation of expansion coefficients becomes simpler and allows for separation of signal energy of the coefficients in the wavelet domain. If $\psi_{j,k}(t)$ is orthogonal to $\phi_{j,k}(t)$ then

$$\int \phi_{j,k}(t) \cdot \psi_{j,l}(t)\ dt = 0, \tag{5.16}$$

for all $j, k, l \in \mathbb{Z}$. Incorporating the idea of wavelets spanning the differences between spaces spanned by scaling function the subspace W_0 is defined so that it satisfies

$$V_1 = V_0 \oplus W_0. \tag{5.17}$$

By extension this gives

$$L^2 = V_0 \oplus W_0 \oplus W_1 \oplus W_2 \ldots \tag{5.18}$$

This shows that any $f \in L^2$ can be represented using one scaling space and a number wavelet spaces. The wavelet function can be defined by the sum of the weighted and translated scaling function,

$$\psi(t) = \sum_{n \in \mathbb{Z}} c_1(n) \sqrt{(2)} \phi(2t - n), \tag{5.19}$$

The generated function is called mother wavelet $\psi(t)$ of a class of functions that have the form

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k). \tag{5.20}$$

2^j characterizes the scaling of the wavelet and $2^{-j}k$ the translation in t. The coefficient set $c_1(n)$ in 5.19 defines the wavelet and since the wavelet function is usually required to be orthogonal to the scaling function the coefficient vector representation of the scaling function $c(n)$ and the coefficient vector representation of the wavelet function $c_1(n)$ are related via

$$c_1(n) = (-1)^n c(1 - n). \tag{5.21}$$

5.1.4 Wavelet transform

Combining the set of scaling functions $\phi_k(t)$ and wavelet functions $\psi_{j,k}(t)$ any function $g(t) \in L^2(\mathbb{R})$ can be expressed in the form

$$g(t) = \sum_{k=-\infty}^{\infty} c(k)\phi_k(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d(j,k)\psi_{j,k}(t). \tag{5.22}$$

The first sum in equation 5.22 describes a coarse approximation of $g(t)$ while for increasing j the second sum adds finer details. In practice where a signal is given by a series of sampled values the coarsest and finest scales are given by the sampled signal itself. While j_0 which sets the coarsest scale can be chosen freely between zero and the highest resolution the finest scale j_{max} is equal to be sample level of the given signal. The discrete wavelet transform of a signal $g(t)$ are the coefficients of the wavelet expansion

$$g(t) = \sum_{k} c_{j_0}(k)\phi_{j_0,k}(t) + \sum_{k}\sum_{j=j_0}^{j_{max}} d(j,k)\psi_{j,k}(t). \tag{5.23}$$

In other words the discrete wavelet transform of a signal of length 2^n is a vector of length 2^n which is composed of subsequences corresponding to the different detail scales

$$d_{n-1}, d_{n-2}, d_{n-3}, \ldots, d_1, d_0, c_0. \tag{5.24}$$

An important feature for the practical use of wavelets and wavelet transform is the existence of scaling and wavelet function that are orthogonal and are nonzero over a finite interval. This feature known as compact support was shown by Ingrid Daubechies [Dau88], [Dau93]. Describing signals based on their wavelet coefficients thus offers the same features as short time Fourier analysis and Gabor analysis using scale as a new variable. However while for orthogonal short time Fourier transform there is a trade-off between time of frequency resolution, meaning that for good time resolution the frequency resolution is poor and vice versa this is not the case for wavelet transform.

5.1.5 Stationary wavelet transform

A drawback of discrete wavelet transform, namely that is not shift invariant, has already been mentioned. The problems can be described as follows, when comparing the transform coefficients of a signal and the same signal which has been circularly shifted on the x-axis prior to the transform both sets of coefficients are of course different, however the relation between both sets is not simple as one could expect given the nature of the shift operation.

Shift invariant discrete wavelet transform (SWT) uses a non-decimated filter bank to create a redundant set of transform coefficients that are shift invariant[LGO$^+$95],

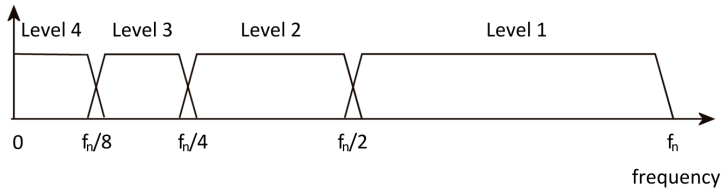

Figure 5.1: Schematic illustration of discrete wavelet transform (DWT) coefficients in relation to spatial frequency of the source data.

[LGO+96]. The idea behind SWT is to calculate the DWT of all possible circular shifts of a given finite signal. Given a signal of length N the a single DWT as also N coefficients and a computational complexity of $O(N)$. Since there are N different circular shifts for a signal of length N the complexity of the overall wavelet transform then is $O(N^2)$ and the required storage space also N^2.

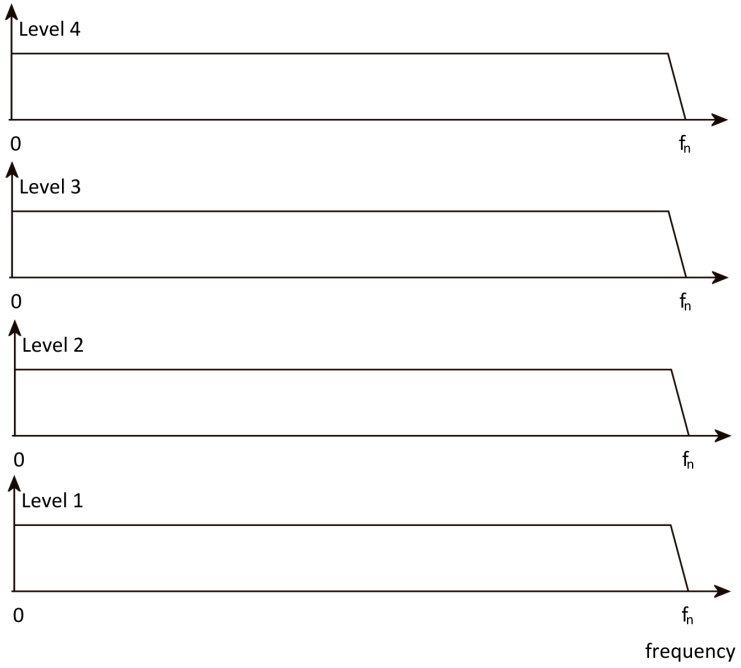

Figure 5.2: Schematic illustration of shift invariant discrete wavelet transform (SWT) coefficients in relation to spatial frequency of the source data.

The initial method was later enhanced independently by Shensa, Beylin and others to a computational complexity of $O(N \log(N))$ [Bey92] [She92]. The enhancement is based on examining the decimating filter structure of the DWT in connection with the circular shift operator. If for a non-shifted signal all even indexes are dropped by the decimating filter then a shift by one will result in all odd indexes to be dropped. However for a shift of two the same values as for the non-shifted signal are dropped. Observing all even

and odd shifts reveals that there are only $2N$ different wavelet coefficients after the first of the shift invariant version of the wavelet transform. Since only half of the values are propagated to the next stage of the filter bank structure one can prove a computational complexity of $O(N \log(N))$. Table 5.3 shows the SWT coefficients of synthetic normal distributed noise. Each scale (or level) consist of $N = 2048$ data-points. One has to be careful with the term of frequency in a wavelet context, however generally speaking smaller scale numbers correspond to small wavelet functions and thus usually to high frequency information while higher scale numbers contain coefficients of dilated wavelet functions which usually represent lower spatial frequencies.

5.2 Short time Fourier transform

Many ideas realized in wavelet transform are related to Fourier transform, which is why the idea of locating a signal on the time frequency plane was also - and with some restrictions - realized using Fourier transform. The classic Fourier spectrum separates information into different frequencies but contains not information about the time (or location) domain of a signal. The reason for this is the global nature of the functions $e^{j\omega x}$ or $\sin \omega t$ and $\cos \omega t$. Global in this context means that any change of the function at a point on the t-axis influences every point on the ω -axis. A simple but effective way to obtain local frequency information from a larger signal is to remove the desired region from the larger context and calculate the Fourier transform only for the smaller separated region. This method is called windowed Fourier transform or short time Fourier transform (STFT) and was introduced by Gabor in 1946 [Gab46]. The function used to separate the region of interest from the full signal context is called window function $g(t)$ and should possess finite time and frequency variance. This means that given $g(t) \in L^2(\mathbb{R})$,

$$k \cdot g(t) \in L^2(\mathbb{R}), \tag{5.25}$$

and that

$$\frac{dg(t)}{dt} \in L^2(\mathbb{R}). \tag{5.26}$$

The STFT of a signal $f(t)$ is defined as

$$Sf(u, \xi) = \int_{-\infty}^{\infty} f(t) g(t-u) e^{-i\xi t} dt, \tag{5.27}$$

with $|g| = 1$ for any (u, ξ) in $\mathbb{R}^2$. The window function g behaves as a low pass filter. Since the window function essentially cuts irrelevant portions of the signal the STFT, unlike the Fourier transform, does not need to know the entire signal but only the

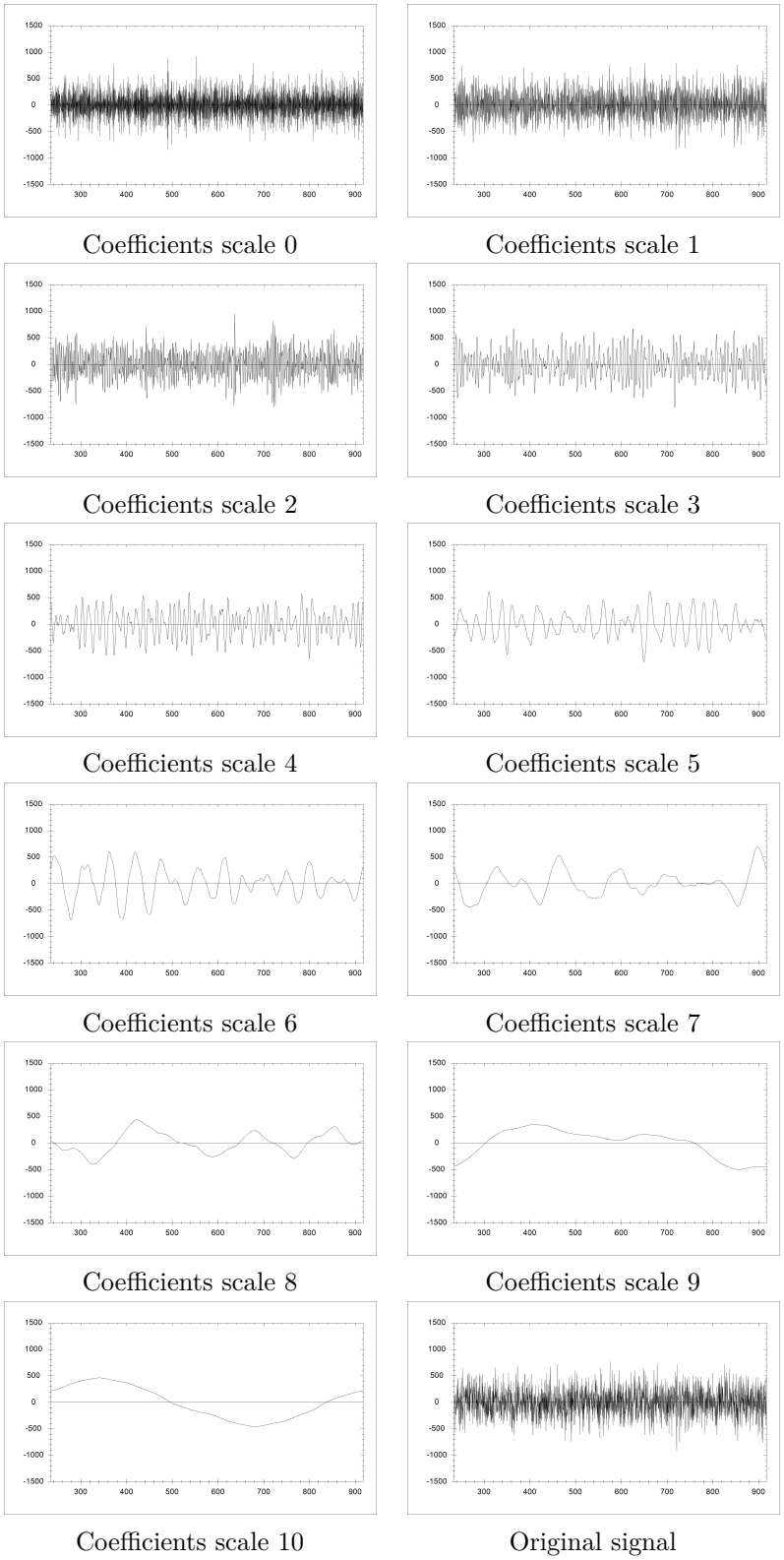

Table 5.3: Shift invariant discrete wavelet transfrom (SWT) coefficients of simulated normal distributed noise, $N = 2048$.

portion for $g(t-u)$ is nonzero. So since essentially short time Fourier transform can be described as Fourier transform coupled with a window function the properties of the window function are of particular interest. Different window functions will result in different transform results as they define the modulation applied to the signal before the Fourier transform is calculated. A drawback of STFT compared to wavelet transform is the fixed nature of the window function. The window function is chosen in advance and the transform is calculated according to the chosen window. However a window and especially the size of the window influence the properties of the transform. Small windows possess good time (or location) resolution as the region which is analyzed is narrow but the frequency resolution is consequently poor. On the other hand larger windows offer better frequency resolution but obviously loose localization properties, meaning the time resolution becomes consequently poorer for large windows. Graphically speaking the window that is moved thought the time-frequency domain is fixed. Wavelet transform introduces the concepts of translation and scaling to counter this drawback of short time Fourier transform.

Chapter 6

Noise reduction utilizing wavelet transform

After the concepts of wavelet transform have been discussed in chapter 5 the following chapter presents the application of wavelet transform in the context of noise suppression and analyzes achieved results with regard to those achieved by classical techniques presented in chapter 3. When Wavelets first reached a status that could be called widely known after the publication of Daubechies in 1988 practical applications for the new technique were rare [Dau88]. Later nonlinear filtering for the purpose of denoising data, which is largely based on the ideas of Donoho and Johnstone together with Coifman and Beylkin, not only improved the general understanding of wavelets, but also helped to formulate and solve important practical problems. Where traditional Fourier transform filter seek to find a linear separation between signal and noise in the frequency domain wavelet based non-linear filtering separates noise and information based on the amplitude of coefficients.

6.1 Thresholding

The thresholding scheme for noise suppression was first proposed by Donoho in 1995. The original scheme contains two different modes that describe a kill-or-keep and kill-or-shrink approach respectively. Hard thresholding sets each coefficient with intensity smaller than a given threshold to zero and leaves other coefficients unchanged [DJ95],

$$T_h(Y,t) = \begin{cases} Y & if \quad |Y| \geq t \\ 0 & else. \end{cases} \tag{6.1}$$

Soft thresholding also sets coefficients with intensity smaller than the threshold to zero,

but also diminishes every other coefficient by the threshold value,

$$T_s(Y,t) = \begin{cases} sgn(Y) \cdot (|Y| - t) & if \quad |Y| \geq t \\ 0 & else. \end{cases} \tag{6.2}$$

This reduces the risk of introducing oscillation artifacts into the reconstructed signal but also creates effects of signal loss and line broadening similar to those that can be observed for Gaussian filters. The application of a thresholding based filter to a given signal S is defined by three steps. First the wavelet coefficients of the signal are calculated based on a chosen wavelet family. The choice of wavelet family naturally influences the corresponding representation in wavelet space and some wavelets might be more suitable to represent certain types of signals and also influence the observed performance of the denoising approach however the general idea of thresholding filters is not influenced by the choice of wavelets. In the second set the thresholding scheme is applied to the coefficients. New coefficients are calculated according to equation 6.1 or 6.2 one at a time on each scale of the transformed signal. Finally the coefficients created by applying the thresholding scheme are used in the inverse wavelet transform to render a reconstructed signal which is similar to the original but exhibits reduced noise characteristics corresponding to the chosen threshold value. As the description of the method suggests the key to successful thresholding is a suitable threshold value that is able to suppress noise to a high degree while leaving the signal features intact and ideally unchanged. Donoho and Johnstone have shown that the universal threshold

$$\lambda = \sigma \cdot \sqrt{2 \log_e(N)}. \tag{6.3}$$

produces asymptotically optimal solutions for noise of constant variance across the domain [DJ94]. The universal threshold is based on the heuristic that every wavelet coefficient contributes noise of variance σ^2 but only a small number of coefficients contain signal information. These assumptions are supported by the observations that signals which are generally spatially inhomogeneous functions are characterized by having their information concentrated in a small subset of wavelet coefficients. Additionally the transform of white noise is also white noise. The estimation of a suitable threshold thus becomes directly tied to the noise standard deviation. Having replaced one unknown variable by another does not appear to be a step towards a solution, however given some modeling assumptions about the expected noise allows the usage of established statistical estimators. Since wavelet coefficients of a noisy signal can be interpreted as a noisy version of the noiseless coefficients Donoho and Johnstone have proposed a robust estimator for the noise standard deviation for normal distributed

noise estimation based on the median absolute deviation. Assuming the median of al coefficient values is zero the standard deviation of normal distributed noise can be estimated based on the median absolute deviation, see chapter 3, by

$$\hat{\sigma} = med(|d_j|) \cdot 1.4826, \tag{6.4}$$

with d_j as the coefficients of the finest detail scale of the wavelet transformed signal. A combination of soft and hard thresholding was introduced by Gao and Bruce in 1997 [GB97]. The firm thresholding schemes aims to combine positive aspects of soft and hard thresholding without inheriting the negative ones. The firm thresholding scheme requires two separate threshold values t_1 and t_2 that essentially define when to use hard thresholding and when to use a modified soft thresholding. The firm thresholding filter function is described as

$$T_f(Y, t_1, t_2) = \begin{cases} Y & if \quad |Y| > t_2 \\ sgn(Y) \cdot \frac{t_2(|Y|-t_1)}{t_2-t_1} & if \quad t_1 < |Y| \leq t_2 \\ 0 & if \quad |Y| \leq t_1, \end{cases} \tag{6.5}$$

with $t_1 \leq t_2$. For $t_1 = t_2$, equation 6.5 describes the hard thresholding scheme while for $t_2 = \infty$ it is equivalent to soft thresholding.

6.2 Shift invariant discrete wavelet transform

Since the shift invariant wavelet transform of a signal consisting of N data-points uses more than N coefficients it obviously introduces redundancy to the representation. This redundancy adds numerical robustness to the transform e.g. against white noise added as a result of quantization. Standard DWT tends to concentrate relevant information in a small number of coefficients making it well suited for data compression. Intuitively operators applied to data sets that offer little to no redundancies are naturally more prone to introduce more severe errors. This can often be observed when applying hard thresholds to standard DWT coefficients as the results may include sharp localized oscillations. Soft thresholding on the other hand has a high probability that the filtered function is at least as smooth as the original function and does not contain sharp oscillations making it visually more appealing while hard thresholding yields better results in terms of root mean square l_2 error. Lang, Guo, Odegard and Burrus have shown that the redundancy in combination with Donoho's thresholding approach improves the noise reduction capabilities in terms of l_2 error and visual quality [LGO$^+$96], [LGO$^+$95].

In particular the negative side effects of hard thresholding are drastically reduced by the usage of redundant wavelet coefficients.

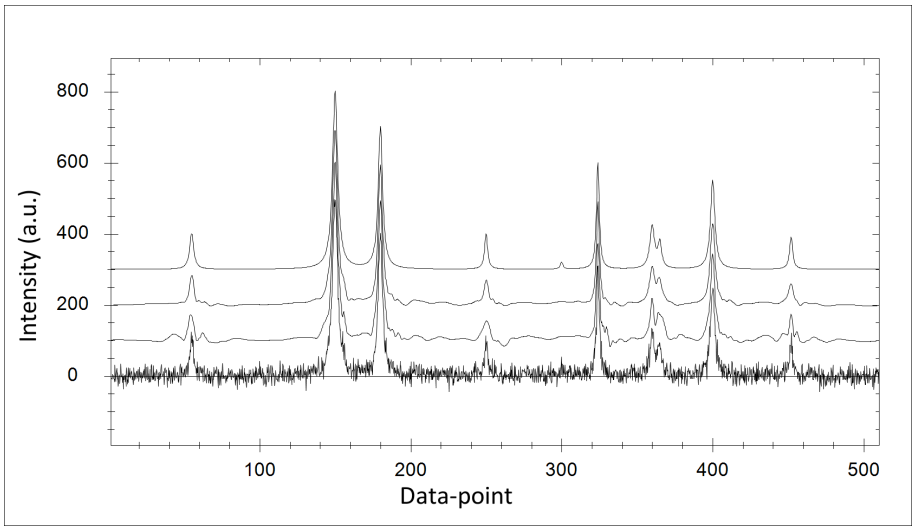

Figure 6.1: Comparison of noise suppression via hard thresholding on discrete wavelet transform (DWT) and shift invariant discrete wavelet transform (SWT) coefficients. Shown are (from bottom to top) the noisy synthetic signal. The noise suppressed data using DWT, the noise suppressed data using SWT and the original data without noise as reference.

6.3 Power distribution in the wavelet domain

White noise is characterized by equal distribution over all frequencies. While wavelet scales are not identical to frequency decompositions of the Fourier transform, the general characteristics are similar and the power distribution of white noise in the detail scales of the stationary wavelet transform is relatively constant while typical signal features generally possess higher mean power values in coarser scales. Figure 6.2 illustrates a typical example of power distribution of pure white noise and of a synthetic measurement containing noise and signal features. The corresponding spectra are given in figure 6.3.

The examination of power distribution in the wavelet domain shows two important facts. One, assuming white noise the finest detail scales offer the most reliable environment to estimate noise characteristics as the influence of signal features is at its lowest while noise characteristics are roughly constant throughout all scales and at their highest resolution in the finest detail scales. And two, limiting noise treatment to the high detail scales cannot remove noise influences as noise is present in all scales to an approximately equal amount. This also illustrates why low pass filters can reduce noise influences but not remove them as effectively as wavelet threshold filters.

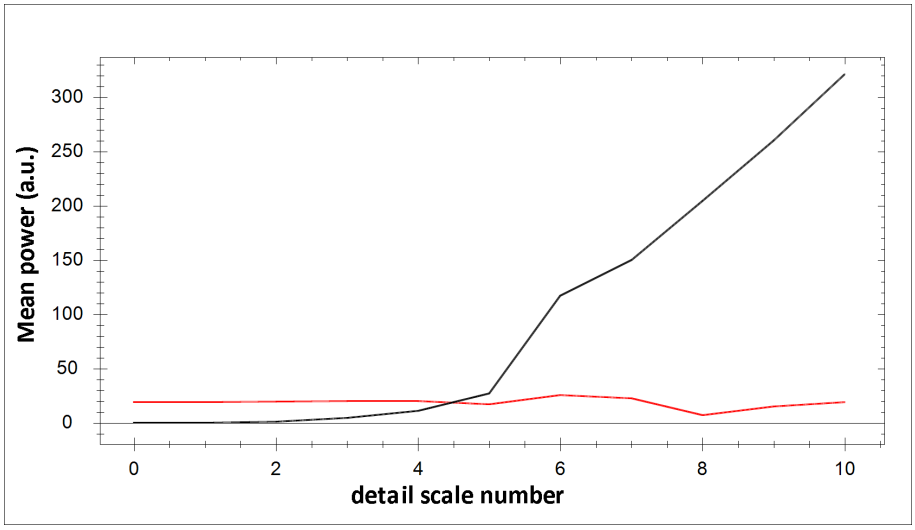

Figure 6.2: Example of mean power per scale in the wavelet domain. Power distribution of white noise (red) and a synthetic measurement containing noise and signal features (black). Mean noise power has been scaled up for illustration purposes.

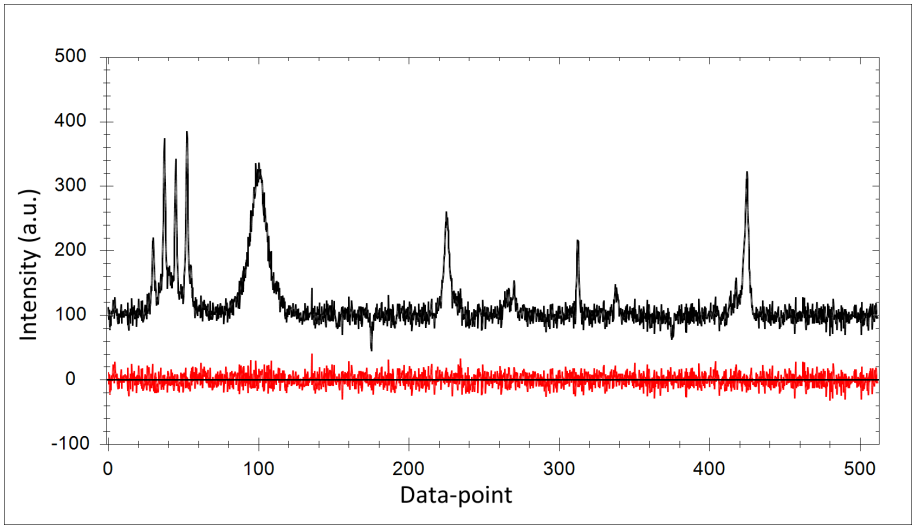

Figure 6.3: Illustration of the synthetic data-series used to illustrate typical power distributions in the wavelet domain. Normal distributed random variable simulating white noise (red) and a synthetic measurement containing noise and signal features (black). Examples contain 2048 data-points each with constant $\Delta x = 0.25$.

6.4 Error characteristics of different thresholds

When speaking of denoising or noise reduction it is important to develop an understanding of how filters work and especially in which ways imperfect noise suppression manifests in the filtered data. Donoho proposed the universal threshold as an asymptotically optimal solution; the threshold however depends on an estimation of the noise

standard deviation which - as any estimation - may differ from the actual value. Considering a data set D with $|D| = N$ and variable threshold value t_i the logical lowernbound for t_i is zero and the upper bound is naturally $t_i = \infty$. The upper bound can easily be reduced to $t_i = \max\left\{\sqrt{c^2}|c \in C\right\}$ with C being the set containing the wavelet coefficients of the wavelet transform of D. Since the threshold filter, independent from the specific scheme, sets all coefficients to zero that are not greater than the threshold any threshold greater or equal to the maximum absolute coefficient returns a constant zero signal.

At the other end of the possible extremes, filtering with a threshold value of $t_i = 0$ has no effect on the signal has the wavelet transform is loss-free and no coefficients are changed by the filter. Now, considering that that signal represented by D consist of signal components (S and noise components R thresholding works under the assumption that S and R are separated in the wavelet domain by their amplitudes and thereby by the value of their coefficients. In the ideal case the threshold can cleanly separate signal and noise while in any realistic scenario S and R overlap to a degree and the optimal threshold merely minimizes the error. With this in mind a threshold that is 'too low' leaves noise components in the filtered signal while a threshold that is 'too high' perhaps removes all noise influences but also damages the signal components to a higher degree. Therefore there is a trade-off between noise reduction and signal degradation. This tradeoff becomes visible when studying the filtered results of varying thresholds. Error measures usually decrease with increasing threshold values up to a point where the implicit classification made by the threshold value produces the overall lowest error. Increasing the threshold value further then also increases the error measures, as the removal of signal components have a more prominent effect than the removal of additional noise influences. This behavior can be seen in table 6.1 and 6.2.

6.5 Performance comparison Gauss smoothing and shift invariant wavelet transform thresholding

Gaussian filtering also corresponds to the multiplicative low pass filtering using and exponential function in Fourier domain, see section 3.3.1. These two simple techniques are well established methods to suppress noise in spectroscopic measurements and other data and therefore a comparison of performance can be used to assess the noise reduction capabilities of thresholding filters. Since Gaussian and the corresponding exponential Fourier low-pass are essentially producing the same results the comparison filter will be addressed as Gaussian filter, if a Fourier filter is used instead parameter changes of course need to be translated. When changing the window-size of a Gaussian filter the corresponding change for the Fourier filter is a change in the shape of the exponential curve that is used for the multiplicative filtering. The adjusting the width parameter

N[0;2]		N[0;5]		N[0;10]		N[0;15]	
Threshold	**MSE**	**Threshold**	**MSE**	**Threshold**	**MSE**	**Threshold**	**MSE**
0	3,98	0	24,87	0	99,47	0	223,8
5	0,6	5	17,51	5	92,31	5	214,72
10	0,47	10	5,5	10	69,64	10	195,24
20	0,79	20	2,25	20	20,88	20	111,18
25	1,19	25	2,27	25	10,39	25	71,14
30	1,46	30	2,6	30	7,66	30	45,2
35	1,9	35	2,95	35	7,2	35	28,64
40	2,39	40	3,6	40	7,11	40	18,91
45	2,98	45	4,05	45	7,64	45	15,67
50	4,02	50	4,83	50	8,37	50	15,15
60	6,75	60	7,23	60	10,5	60	15,7
8,18	0,44	19,41	2,22	38,46	7,05	57,69	15,37

Table 6.1: Exemplary mean squared error comparison of Donoho's hard threshold filter applied to SWT coefficients using different threshold values on signals with N=2048 Data-points, afflicted with normal distributed Noise $N[0;\sigma]$ with different noise levels σ. The last row shows the results obtained by estimating the noise level from the first scale and then using the Universal threshold.

N[0;2]		N[0;5]		N[0;10]		N[0;15]	
Threshold	**MSE**	**Threshold**	**MSE**	**Threshold**	**MSE**	**Threshold**	**MSE**
0	4,15	0	25,94	0	103,77	0	233,48
5	0,83	5	19,9	5	99,49	5	230,62
10	0,77	10	7,47	10	78,79	10	212,41
20	1,29	20	3,36	20	27,76	20	130,91
25	1,73	25	3,29	25	16,57	25	92,51
30	2,36	30	3,68	30	11,49	30	59,95
35	3,29	35	4,67	35	10,37	35	39,36
40	4,09	40	5,23	40	10,41	40	30,85
45	5,18	45	6,55	45	11,26	45	23,13
50	6,14	50	7,48	50	12,32	50	21,54
60	11,1	60	11,14	60	14,08	60	20,86
7,72	0,62	19,15	3,42	38,18	10,31	57,24	20,58

Table 6.2: Exemplary mean squared error comparison of Donoho's hard threshold filter applied to SWT coefficients using different threshold values on signals with N=1024 Data-points, afflicted with normal distributed Noise $N[0;\sigma]$ with different noise levels σ. The last row shows the results obtained by estimating the noise level from the first scale and then using the Universal threshold.

of a Gaussian filter generally produces a similar behavior in error measures than that which can be observed for gradually increasing thresholds and a typical measurement. Typical in this case means that a measurement contains signal and noise components, signal components dominate the flow frequencies and high amplitudes while noise is present in all frequency bands but possesses amplitudes that are significantly less intense than those of signal components. Under these conditions increasing the size of a Gaussian filter produces gradually better results in terms of error measures - be it mean squared error or l_2 error - until a point is reached where a further increase degrades the signal components to a degree that is higher than the gain of additional noise suppression.

Note that the idea behind Gaussian or Fourier low pass filters - attenuate high frequencies in order to remove noise influences - is fundamentally flawed as noise is usually present in all frequency bands. However since lower frequency components are often visually masked by signal features removing only high frequency noise components does improve measurements visually and in terms of error measures even if the filters are by design not suited to remove noise completely.

Table 6.3 shows error measures of Gaussian and thresholding filters using different parameters. In both cases the parameters are chosen to show the initial decrease and later increase in error measures. The table also shows that the overall error values achieved using shift invariant wavelet transform in combination with the hard thresholding scheme are superior to those achieved using traditional Gaussian filtering.

Filter type	**Parameter**	**MSE**
Without Filter	-	33.25
Gauss	Window-size = 3	10.03
	Windowsize = 4	8.24
	Windowsize = 5	8.12
	Windowsize = 6	9.49
SWT hard threshold	Threshold = 10	8.43
	Threshold = 20	2.18
	Threshold = 40	2.92
	Threshold = 80	10.18

Table 6.3: Comparison of mean squared error after noise suppression via Gauss filtering and SWT hard threshold filtering. Artificial signals were afflicted with uniform distributed noise U[-10,10].

Table 6.4 shows a similar evaluation for the second derivative of the filtered data. The second derivative can be used to describe curvature changes and therefore can serve as indicator for smoothness characteristics. The results show the same trend that can be observed in table 6.3 however the optimum for both methods is shifted towards stronger filtering - higher threshold and larger windows respectively. This behavior is unsurprising as numerically calculating derivatives amplifies high frequency noise.

Filter type	Parameter	MSE
Without Filter	-	189.81
Gauss	Windowsize = 4	1.42
	Windowsize = 5	0.58
	Windowsize = 6	0.35
	Windowsize = 10	0.49
SWT hard threshold	Threshold = 10	36.88
	Threshold = 20	0.29
	Threshold = 40	0.11
	Threshold = 80	0.3

Table 6.4: 2^{nd} derivative mean squared error comparison of gauss filtering and threshold filtering. Artificial signals were afflicted with uniform distributed noise $U[-10, 10]$ before derivative calculation.

As visual quality of denoising methods is another important characteristic the figure 6.4 and figure 6.5 offer a graphical comparison between smoothing results obtained using Gaussian filters and those obtained using SWT threshold filters.

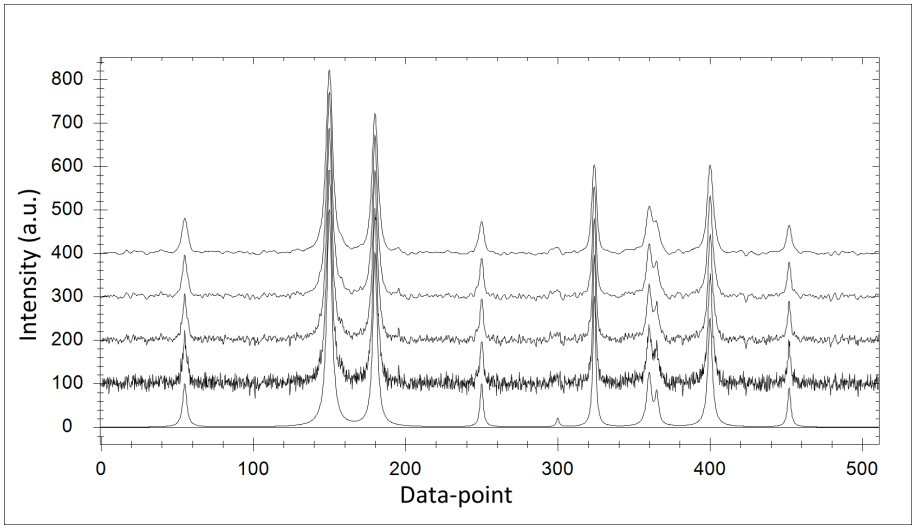

Figure 6.4: Illustration of noise suppression via Gaussian filter. Shown are, from bottom to top, the pure synthetic signal (2048 data-points), the signal afflicted with normal distributed noise $N[0, 10]$ and the results of filtering the noise afflicted signal using a Gaussian filters of window-sizes 3, 6 and 12 data-points.

As mentioned before Gaussian smoothing does not remove the noise but merely the high frequency components of noise, as a result regions that do not contain any signal as still affected by the lower frequencies of the noise spectrum and remain rippled. Achieving a relatively flat behavior in signal free regions often requires very wide filter apertures which in turn result in severe signal degradation. Choosing a threshold that is too low reveals the hit or miss philosophy of hard thresholding. Figure 6.5 shows an

example of noise partially passing the filter without being attenuated. Increasing the threshold value reduces the number of blips that remain as a result of noise passing the filter at the cost of attenuation of smaller signals.

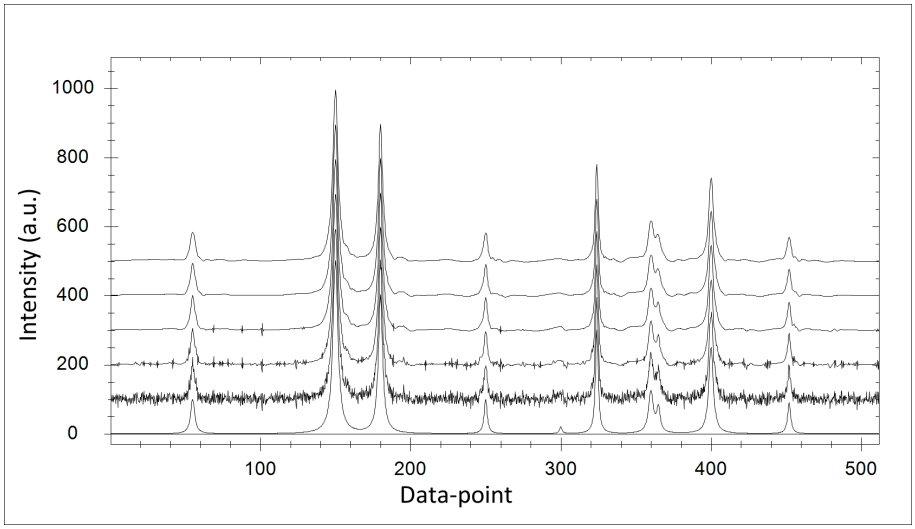

Figure 6.5: Illustration of noise suppression via SWT hard-thresholding filter. Shown are, from bottom to top, the pure synthetic signal (2048 data-points), the signal afflicted with noise $N[0, 10]$ and four results of filtering the noise afflicted signal using a thresholds of 20, 30, 39 and 50. The calculated universal threshold value is 39.05.

6.6 Modulated noise

The universal threshold approach was designed to estimate a suitable threshold to suppress normal distributed white noise in wavelet coefficients of noisy signals. However noise parameters encountered in real measurements are not always constant and methods that are suitable to approximate noise of roughly constant variance perform poorly in situations where the noise variance itself is variable. The new method presented here is based on the idea that noise intensity itself is a non constant function and accordingly tires to approximate a threshold function instead of a singular threshold value. Assuming variations of noise variance within a given signal can be expressed by constant noise σ - meaning that the variance of the distribution function is constant - and a modulation function f_{mod} then variable noise $\sigma(x)$ can be described as

$$\sigma(x) = \sigma \cdot f_{mod}(x). \tag{6.6}$$

To illustrate the problems arising from constant thresholds and variable noise it is possible to analyze simple examples. Arguably one of the simplest modulating functions

is a linear increasing straight line, $f_{mod}(x) = ax + b$. The universal threshold is not well suited to suppress noise of linear increasing intensity because one of the modeling assumptions implies that the noise variance σ is constant for the full signal domain. Using the median absolute deviation to estimate σ leads to a threshold value that is well suited to suppress noise with a variance that is approximately as high as the variance in middle of the finite signal domain. For noise values that are observed at x_i with $x_i < x_{N/2}$ the suppression is sufficient to remove noise influences from the signal but tends to be higher than actually necessary. On the other hand, for x_i with $x_i > x_{N/2}$ a constant threshold estimated via the median absolute is too small to achieve sufficient noise suppression. With increasing x_i increasingly large portions of noise pass the filter largely unsuppressed creating a two sided appearance in the reconstructed signal.

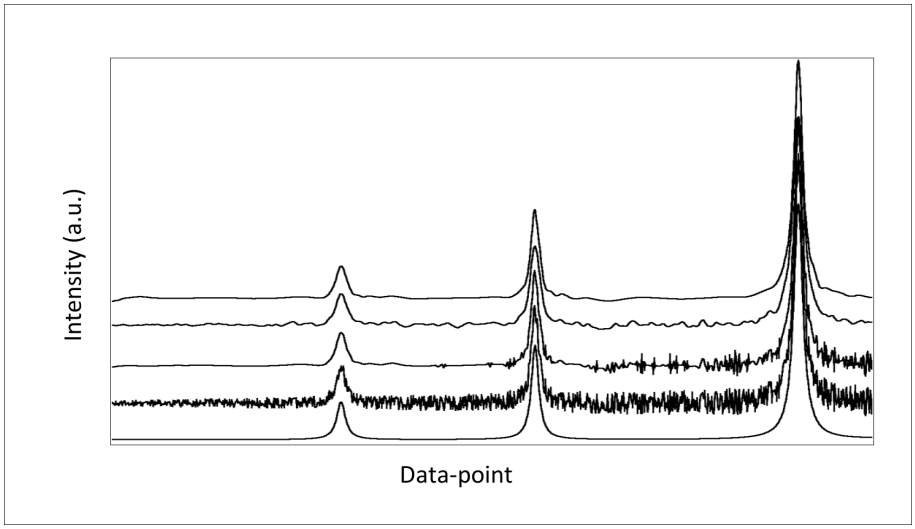

Figure 6.6: Noise suppression results on synthetic signal afflicted with normal distributed noise of linear increasing variance. Shown are, from bottom to top : 1)The pure synthetic signal. 2) Signal afflicted with noise. 3) The reconstructed signal after SWT hard thresholding using the universal threshold. 4) The result after applying a Gaussian filter to 2. 5) the reconstructed signal after SWT hard thresholding using a variable threshold function.

Figure 6.6 illustrates the effect of different noise suppression techniques on noise of linear increasing variance. The result of Gaussian filtering is given as a reference to show that while the Gaussian filter does not let noise pass unsuppressed it is unable to achieve the same smoothness as thresholding filters. The figure also illustrates the mixed results which are achieved using a constant threshold on variable noise. While on the left side of the signal noise is well suppressed the right side still contains severe noise distortions. Furthermore it can be observed that while the chosen threshold is too high for noise intensities encountered in the left part of the signal the resulting reconstructed signal is not severely distorted. The results shown in figure 6.6 might suggest that an upper bound estimation of changing noise parameters leads to satisfying

results regarding noise suppression and signal preservation. However while this is true for noise suppression - the threshold that sufficiently suppresses the highest contained noise intensities also suppresses lower intensities - signal preservation can suffer severely under upper bound threshold filtering. Considering a signal containing information in the form peaks of variable heights and noise of changing variance a constant threshold can severely diminish or under adverse conditions even eliminate entire peaks. Figure 6.7 shows an example designed to illustrate the worst case of complete signal loss due to the combination of variable noise parameters and constant threshold estimation. The signal domain is afflicted with noise that that displays slowly oscillating characteristic with strongly changing intensity. Several intense peaks are located in regions of high noise variance while a smaller peak is found in a region of comparatively low noise variance. Choosing an upper bound threshold to estimate a suitable threshold results in a relatively smooth reconstruction with suppresses the high noise intensities sufficiently but also eliminates the small signal. On the other hand a variable threshold allows different values to suppress noise in the regions of intense and minor noise. Thus the variable threshold achieves similar suppression results in regions of high noise variance but since the noise intensity in the region of the small peak is comparatively low no signal loss occurs. Note that a variable threshold does not prevent signal loss; if the small peak in the given example had been located in a high noise region it had been lost in the filtering process. If the characteristics of noise and signal are not sufficiently separable no filter that does not have access to additional information can prevent signal loss and suppress noise at the same time. However by allowing the threshold value to follow local noise intensities unnecessary signal degradation or loss can be minimized, [SH13], [SHML12].

6.7 Suppression of noise with modulated intensity

While the universal threshold is well suited to suppress normal distributed noise in wavelet coefficients of noisy signals, a changing noise distribution can lead to over and underperformance of the noise suppression characteristics as shown in figures 6.6 and 6.7. To approximate non-constant noise it is possible to utilize the wavelet transform's characteristic of keeping a sense of locality due to the fact that wavelets are finite functions. Since coefficients correspond locally to the noise characteristics of the original signal, usage of a threshold function in the wavelet domain can achieve superior suppression characteristics if noise parameters are changing within the signal domain. This is a new approach to noise suppression as noise characteristics are generally assumed to be constant within the signal domain. Allowing noise parameters to be non-constant the locality of wavelet coefficients becomes an essential feature in the attempt to reduce noise influences. Locality is limited by the size of the wavelets which directly

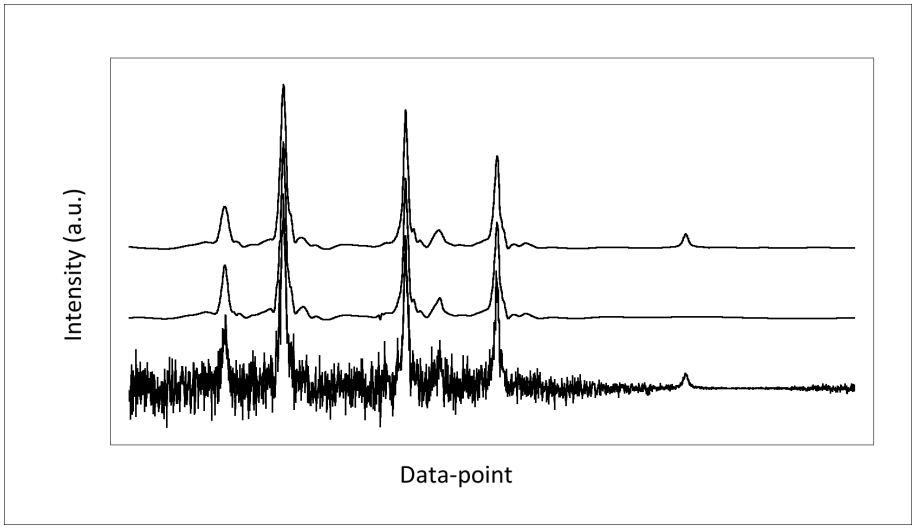

Figure 6.7: Visual example of a constant threshold in a variable noise environment. Shown are the original noisy data set (bottom), the reconstructed signal using the universal threshold (middle) and the result obtained using a variable threshold function (top).

corresponds to the scale and therefore the localization of features is most distinct in fine detail scales where scaled wavelets describe the influence on very small areas of the signal domain. In other words the scales that are most suitable to estimate noise parameters are also the ones that possess the best localization of information. With these facts in mind it is logical to attempt the estimation of noise that exhibits changing parameters across a given domain using the high detail scales of the wavelet transformed signal. Without prior knowledge about the modulation of noise the estimation of the correct threshold function is challenging. Two different approaches have returned good results in real and artificial scenarios, they are described in the following sections.

6.7.1 Windowed median

Starting from the estimation of σ for a constant noise environment using the median absolute deviation a natural approach towards variable noise parameters is a windowed median function that is not based on the entire signal domain but only on a local section. Assuming the variable noise intensity $\sigma(x)$ can be expressed by constant normal distributed noise σ and a modulation function $f_{mod}(x)$ according to equation 6.6 and that $f_{mod}(x) \approx const$ for small intervals $x_i, \ldots, x_j$, the estimated modulated standard deviation $\hat{\sigma}(x)$ can be expressed as

$$\hat{\sigma}(x) = med(|d_j(x, w)|) \cdot 1.4826, \tag{6.7}$$

with $d_j(x, w)$ as the coefficients of the finest detail scale of the wavelet transformed signal that fulfil $|x - j| \leq w/2$ for SWT. For DWT the coefficients fulfils $|x/2 - j \leq w/4|$, considering that the finest detail scale of the standard DWT contains N/2 coefficients. The quality of the approximation is directly dependent on the underlying function $f_{mod}(x)$ and the chosen window-size w for the median operator. Small window sizes reduce the robustness of the median estimation while for larger window-sizes the assumption of $f_{mod}(x) \approx const$ becomes increasingly problematic. Based on $\hat{\sigma}(x)$ the threshold function $\lambda(x)$ can be expressed analog to equation 6.3 as

$$\lambda(x) = \hat{\sigma}(x) \cdot \sqrt{2 \log(N)}. \tag{6.8}$$

Due to the nature of the wavelet coefficients of a noisy signal the windowed median operator will inevitably create a curve that displays irregular and angular behavior. Small window-sizes which enable the estimation to follow higher frequency modulating functions more closely amplify this behavior. For large window sizes the local median converges towards global median negating local influences of $f_{mod}(x)$. Furthermore an oscillating modulating function can lead to aberrant behavior of the derived threshold function given an adversely chosen window-size, see figure 6.8.

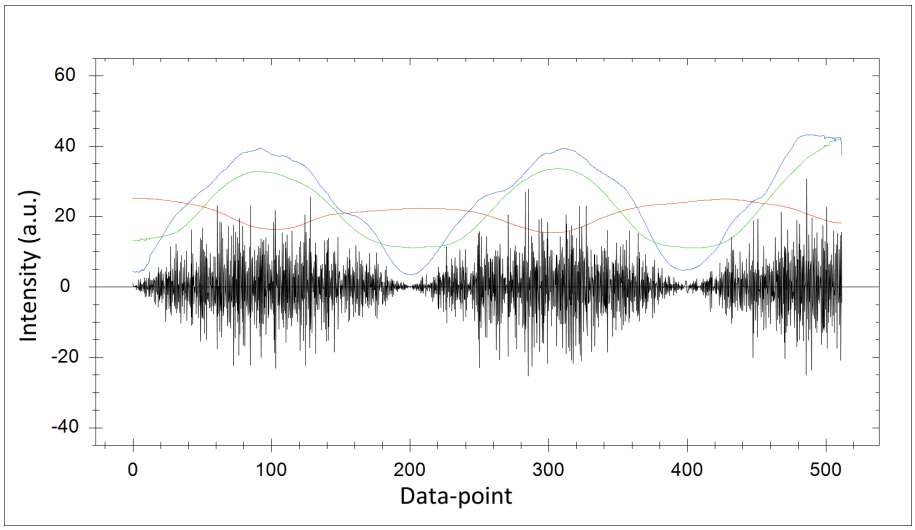

Figure 6.8: Noise threshold estimation using windowed median functions of window sizes $w = 50$ data-points (blue), $w = 200$data-points (green) and $w = 500$data-points (red). The shown signal contains 2048 data-points with $\Delta x = 0.25$

The presented example (figure 6.8) shows threshold functions based on the windowed median using different window sizes. While thresholds based on smaller windows manage to follow the trend of the noise, the threshold function based on the largest displayed window size displays behavior that is countercyclical to the correct behavior.

6.7.2 Enveloping functions

An alternative method to approximate the optimal threshold for modulated noise that converges towards the global maximum noise intensity uses enveloping functions, which are inspired by morphological operators. Direct application of morphological operators is not advisable since information of spectral measurements often differs dramatically in scale. Furthermore, often no fixed relation between the dimensions of different measurements can be given. Other than images that usually possess two dimensions that are fixed in a constant ratio, the intensity axis of spectral measurements is usually very variable and ratio changes between x and y axis are quite common when determining the ideal visual representation of data. Therefore the operators used to model enveloping functions use an interval - similar to the window size of the windowed median - that is defined in the x dimension and select a sample set from the existing data-points. The upper envelop selects the highest occurring data-points within the given interval while the lower envelop selects data-points with the smallest intensities, respectively. Let S be a signal with $S(i) = (x_i, y_i)$. Furthermore let $I(x, M)$ be a function that describes an interpolation method, based on a linear or higher ranking spline and

$$M_{high}(x_k, w) = \{(x_i, y_i) \in S | \forall j : (|k - j| \leq w/2) \wedge y_i \geq y_j\}, \tag{6.9}$$

which describes the subset of S that contains the highest points in regions defined by the window size w, then the enveloping function $Env_{upper}(x, w)$ is given by

$$Env_{upper}(x, w) = I(x, M_{high}(x, w)). \tag{6.10}$$

Accordingly the function $Env_{lower}(x, w)$ is defined as

$$Env_{lower}(x, w) = I(x, M_{low}(x, w)), \tag{6.11}$$

with M_{low} as the subset of S which contains the lowest points in regions defined by the window-size w,

$$M_{low}(x_k, w) = \{(x_i, y_i) \in S | \forall j : (|k - j| \leq w/2) \wedge y_i \leq y_j\}. \tag{6.12}$$

In most cases a simple linear interpolation combined with a Gaussian filter is sufficient for the purpose of threshold estimation. More information on interpolation and advanced splines are given in [Ack15], [ANW67], [Aki70], [Boo78] and [Hil87]. It should

be noted that the points relevant for the enveloping functions are not exclusively local maxima or minima of the enveloped signal, but rather local minima cannot be part of the data-point set chosen by the upper-enveloping function while local maxima cannot be part of the set chosen by the lower-enveloping function. The enveloping functions are built using interpolations between neighboring points in the respective sets of data-points forming an upper and lower envelop of the given signal. A major problem for the utility of upper- and lower enveloping function are outliers which both functions are sensitive to. The upper enveloping function contains positive or high intensity outliers while the lower enveloping function contains negative or low intensity outliers. To reduce the influence of outliers the enveloping operators can be executed consecutively similar to the morphological opening and closing operators. The function $UL(x, w)$ obtained by first applying the upper envelop filter followed by the lower envelop filter while the function $LU(x, w)$ is obtained by applying a lower enveloping filter followed by an upper enveloping filter,

$$\begin{aligned} UL(x,w) &= Env_{lower}(Env_{upper}(f(x),w),w), \\ LU(x,w) &= Env_{upper}(Env_{lower}(f(x),w),w). \end{aligned} \tag{6.13}$$

Applying the UL and LU functions the coefficients of a signal in the wavelet domain results in two curves which ideally fulfill $|UL(x, w)| = |LU(x, w)|$. Under real conditions both curves will not be identical but describe the same overall trend while being obscured by inherent irregularities of the noise distribution and artifacts caused by sub-sampling the interpolation.

The so defined enveloping functions are based on assumptions about wavelet coefficients that are very similar to those made for the universal threshold but less demanding. The model condition that assumes noise to be composed of independent identically distributed $N(0, 1)$ random variables is dropped. The assumption that noise contaminates all wavelet coefficients still holds but with the addition of an unknown modulating function it is no longer a postulated that the influence of noise is equal for all coefficients. The assumption that signal influences are concentrated in a small subset of wavelet coefficients remains unchanged. Without a modeling assumption describing the noise distribution the threshold function is directly dependent on the observed coefficients values rather than on estimated parameters of a modeling function. Since both $|UL(x, w)|$ and $|LU(x, w)|$ describe enveloping functions which ignore some outliers a suitable threshold function $\lambda(x, w)$ can be obtained by scaling either $|UL(x, w)|$ or $|LU(x, w|$ with a factor that corresponding to the minimum signal to noise ratio. For illustrations shown here variable thresholds were calculated using the difference of UL and LU.

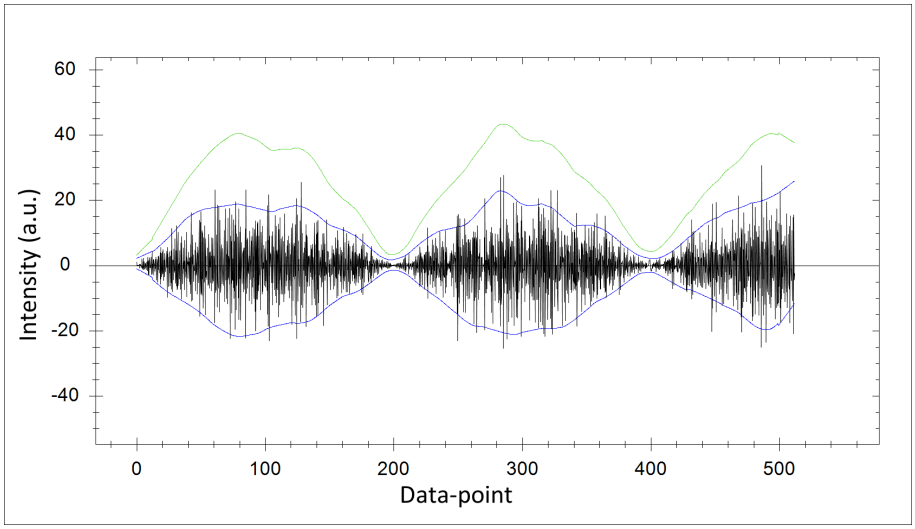

Figure 6.9: Illustration of the enveloping fucntions $UL(x, w)$ and $LU(x, w)$ for $w = 50$ and the resulting threshold function $\lambda(x)$. The signal is identical to figure 6.8 and contains 2048 data-points with $\Delta x = 0.25$

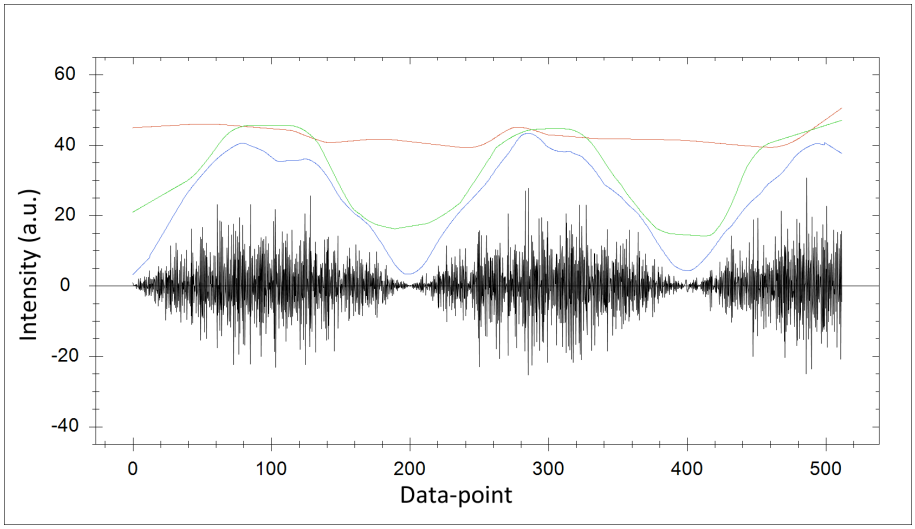

Figure 6.10: Noise threshold estimation using enveloping functions of window sizes $w = 50$ data-points (blue), $w = 200$ data-points (green) and $w = 500$ data-points (red). The shown signal contains 2048 data-points with $\Delta x = 0.25$

6.7.3 Results

For testing normal distributed noise $N[0, 1]$ was modulated using the functions given in Figure 6.11. Note that the functions were shifted for better visual representation; for the actual tests the minimum value of each modulating function was set to 1 and a maximum value set to 5. Noise suppression results presented in table 6.5 and table 6.6 show that performances of windowed median and enveloping filter are similar. The

Median Threshold Estimation MSE Values								
window-size	**ZigZag1**	**Rect1**	**Sine1**	**ZigZag2**	**Rect2**	**Sine2**	**line1**	**line2**
-	14,472	19,328	15,613	14,359	19,009	15,413	13,983	6,610
10	1,566	1,688	1,589	1,760	2,214	1,878	1,798	0,889
20	1,393	1,627	1,412	1,647	2,106	1,733	1,668	0,820
50	1,297	2,092	1,292	1,551	2,278	1,632	1,595	0,769
100	1,259	3,167	1,339	1,517	2,841	1,590	1,573	0,757
300	2,121	8,902	3,129	1,448	3,900	1,657	1,530	0,757

Table 6.5: Noise suppression mean squared error results of synthetic samples containing noise modulated by selected functions. The threshold function was estimated using the windowed median with different windowsizes. The first row of the table gives the MSE values of the unfiltered noisy data as a reference.

Envelop Threshold Estimation MSE Values								
window-size	**ZigZag1**	**Rect1**	**Sine1**	**ZigZag2**	**Rect2**	**Sine2**	**line1**	**line2**
-	14,472	19,328	15,613	14,359	19,009	15,413	13,983	6,610
10	1,552	1,987	1,600	1,731	2,137	1,802	1,650	0,850
20	1,248	1,540	1,256	1,449	1,773	1,500	1,415	0,732
50	1,303	1,872	1,310	1,550	2,312	1,624	1,537	0,768
100	1,546	2,366	1,674	1,721	2,903	1,823	1,812	0,878
300	2,016	3,409	2,193	1,868	3,286	2,024	2,254	1,072

Table 6.6: Noise suppression mean squared error results of synthetic samples containing noise modulated by selected functions. The threshold function was estimated using the envelop method with different windowsizes. The first row of the table gives the MSE values of the unfiltered noisy data as a reference.

main differences between both methods are distinct behaviors using relatively small or large window size values. While the windowed median converges towards the global median for large widow sizes the enveloping methods converges towards a threshold capable of suppressing the highest noise intensity. This means that given a modulated noise environment and large widow size parameter the windowed median will fail to suppress higher intensity noise distortions while the enveloping method will suppress all noise but on the other hand is more likely to cause signal degradation. Figure 6.12 gives an example of the threshold estimation using both methods when noise intensities have been modulated by a rectangle function. While the sharp edges of the rectangle are flattened in bot estimation processes the general shape of the modulating function is clearly visible in both cases. Furthermore both threshold estimations manage to capture the actual noise intensities visible in the noisy synthetic data reasonably well in high and low intensity noise regions.

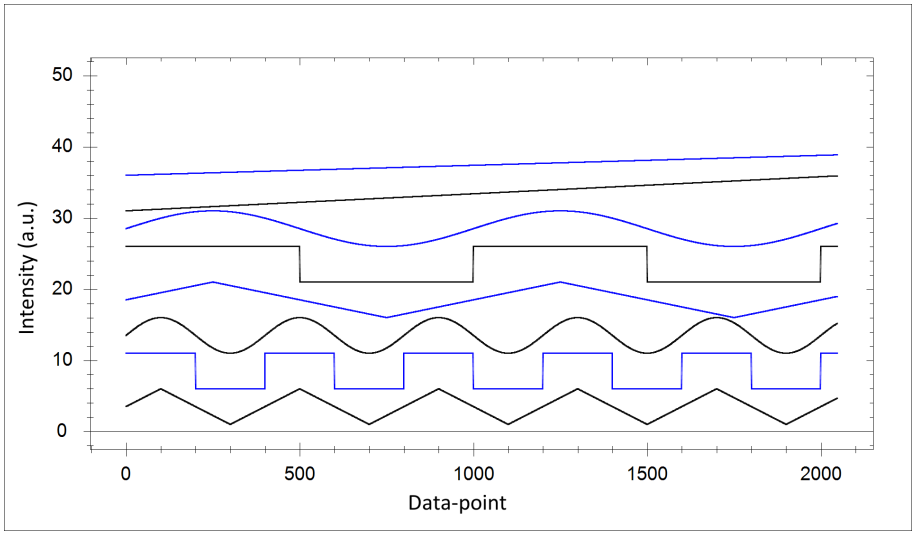

Figure 6.11: Illustration of the modulation functions used for the results described in table 6.5 and table 6.6. Function labels from bottom to top : *ZigZag*1, *Rect*1, *Sine*1, *ZigZag*2, *Rect*2, *Sine*2, *Line*1 and *Line*2.

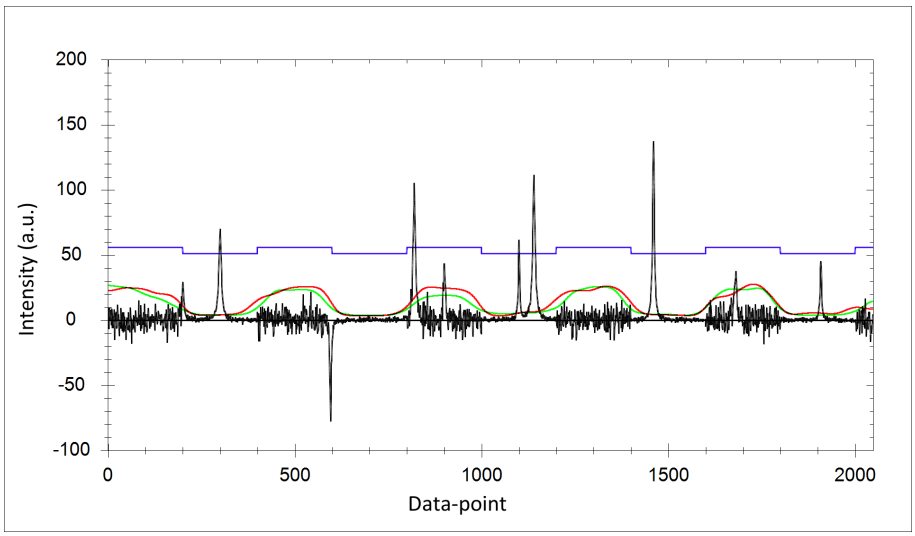

Figure 6.12: Illustration an exemplary synthetic spectrum afflicted with modulated noise. The noise modulating function is labeled *Rect*1. The noise suppression threshold function estimated using the windowed median is given in green while the envelop version is given in red. Also a shifted version of the modulation function (blue) is given as reference.

6.7.4 Discussion

Positive aspects of the threshold estimation via enveloping functions are slightly enhanced calculation speed since no ordered list of intensity values needs to maintained, independence of model assumptions concerning the noise distribution function and a more desirable asymptotic behavior for large window-sizes. For window-size parameters

that do not allow the approximation to follow the modulating function close enough the positive aspects of a variable threshold are partially negated however other than the median approach which allows portions of modulated noise to pass unfiltered the asymptotical behavior of the envelop based threshold approximation ensures that the filter is suitable to suppress the upper bound of the contained noise variance without sacrificing the ability to follow the trend of a modulating function for suitably small window sizes. A weakness of the approach is the limited capability to correct for outliers. Clusters of outliers that are dense with regard to the window-size w cannot be corrected by chaining of upper and lower enveloping function or lower and upper respectively. Note that other than for the median based approach larger window-sizes generally result in poorer performance with regard to outlier correction. This behavior is based on the selective nature of the enveloping functions, for large windows the number of points that form the sets M_{high} and M_{low} is smaller than for smaller window sizes. Window-size is an important parameter for both the envelop approach and the windowed-median approach. Ideally window size could be determined automatically however it is easy to see that estimating changing variance values using sample values is difficult as it cannot be assured that any set of samples corresponds to the same variance value. In fact given a simple linear increasing modulating function no two data-points correspond to the same variance. Generally smaller window sizes approximate modulated variances better than larger ones but are more prone to errors introduced by outliers of other artifacts which can be caused by residues of signal features that are visible even high detail scales of the wavelet transform. So while the use of a threshold function greatly enhances noise suppression characteristics in variable noise environments the function is more difficult to estimate than a singular value. However practical tests using standoff Raman and LIBS spectra have shown that in practical applications both methods produce good results given a suitable choice of the window-size parameter. For real spectra containing 2096 data-points which showed indications of variable noise parameters in the wavelet domain window sizes of between 30 and 100 data-points returned good results independent of the chosen estimation method.

Chapter 7

Baseline estimation by adaptive local regression

Capabilities and drawbacks of existing baseline estimation and correction techniques have been discussed in chapter 4. The aim of the new method introduced here is not to be more accurate than already existing methods in an ideal scenario but to reduce the requirements made by baseline correction techniques in order to make automatic baseline estimation more accessible to end users of spectroscopic data like physicists, chemists, biologists etc. Unlike most other baseline estimation techniques adaptive local regression (ALR) is not disrupted by multidirectional distortions or variable noise and is able to handle a wide variety of signal characteristics without the need for external parameters. The method is illustrated by first discussing robust locally weighted regression including its characteristics, possible problems and solutions. After that the concepts of adaptive local regression are explained in detail, limitations examined and results are discussed by comparisons with other state of the art baseline estimation techniques.

7.1 Robust locally weighted regression

Robust locally weighted regression was introduced by Cleveland [Cle79] to enhance the visual information of scatter-plots. The method by Cleveland extended the already known local regression by an iterative fitting schema that added robustness against outliers to the regression which otherwise were known to present a considerable problem and lead to misleading regression results. The idea behind using the algorithm for baseline correction in spectral data is the observation that signal features can be interpreted as outliers within the sampled approximation of the baseline curve.

7.1.1 Algorithm by Cleveland

Because the algorithm for robust locally weighted regression is a key aspect of ALR baseline correction it is summarized here: Given a dataset $d = (x_i, y_i)$ with $i = 1, \ldots, n$ and an unimodal, symmetric non-negative weight function W, the robust locally weighted regression algorithm is defined by four major steps.

Robust locally weighted regression: Step one

For each point i calculate the initial local polynomial regression of degree d : $(x_i, \hat{y}_i)$ with

$$\hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(x_i) x_i^j, \tag{7.1}$$

where $j = 0, \ldots, d$ and the $\hat{\beta}_j(x_i)$ as the values of β_j that minimize

$$\sum_{k=1}^{n} (W(\frac{x_k - x_i}{h_i}) \cdot (y_k - \beta_0 - \beta_1 x_k^1 - \ldots - \beta_d x_k^d)^2), \tag{7.2}$$

where $\beta_0, \ldots, \beta_d$ are the coefficients of the chosen polynomial and h_i the bandwidth of the neighborhood around x_i, which defines the number of data-points that are used for the local regression of x_i.

Robust locally weighted regression: Step two

Calculate the robustness weights $\delta(k)$ using the bisquare weight function,

$$\delta(k) = \begin{cases} \left(1 - \left(\frac{e_k}{6 \cdot med(|e_i|)}\right)^2\right)^2 & \text{if } e_k \geq med(|e_i|) \\ 0 & \text{else,} \end{cases} \tag{7.3}$$

with $e_i = y_i - \hat{y}_i$ defining the residuals between the original data y_i and the regression $\hat{y}_i$.

Robust locally weighted regression: Step three

For each point i calculate the local polynomial regression of degree d using modified weights in the calculation of $\hat{\beta}_j(x_i)$, changing it to the values of β_j that minimize

$$\sum_{k=1}^{n}(\delta_k \cdot W(\frac{x_k - x_i}{h_i}) \cdot (y_k - \beta_0 - \beta_1 x_k^1 - \ldots - \beta_d x_k^d)^2). \tag{7.4}$$

Robust locally weighted regression: Step four

Repeat steps two and three t times with t being the number of user determined iterations. In the original article Cleveland states that based on tests with real and artificial data $t = 2$ is adequate for almost all situations. Note that these tests were aimed to test the algorithms properties in scenarios where data is only corrupted by noise and occasional uncorrelated outliers. These assumptions are not true if the algorithm is used to estimate baseline behavior where signal contributions are interpreted as outliers.

7.1.2 Choice of parameters

Weight function

The precise choice of the weight function has no large impact on results obtained by the algorithm [Sim96], [FG96] as long as weights are decreasing with increasing distance from the current data-point. Often the weight function is realized by the tricube weight function given in equation 7.5 which was proposed by Cleveland for its favorable characteristics - unlike e.g. the Gaussian function it does not have long tails and is truly zero outside the neighborhood defined by h - and its comparably simple calculation.

$$W(x) = \begin{cases} (1 - |x|^3)^3 & \text{if } |x| < 1 \\ 0 & \text{else.} \end{cases} \tag{7.5}$$

Polynomial degree

Cleveland's recommendation for the polynomial degree of the regression polynomial is $d = 1$ which translates to the assumption of local linearity. For $d = 0$ one would assume a locally constant function which is unrealistic for the purpose of baseline estimation. For higher polynomial degrees computational complexity begins to become a factor. Furthermore the assumption of linearity can be interpreted as the most neutral regression model for the purpose of baseline estimation and higher polynomial degrees lead to regressions easily following the shape of signal peaks instead of eliminating them in the iterative process. Note that the linearity assumption in the regression does not necessarily result in a piecewise linear function since each points is regressed separately and neighborhoods change for every data-point.

Number of iterations

The number of iterations t controls the robustness of the regression. Since the target or optimal regression is unknown it is possible to define criteria that evaluate the changes in the regression curve from one robust iteration to the next and define a threshold at which to stop thus determining t automatically. Cleveland assesses this as "needlessly complicated" and proposes a fixed number of $t = 2$ based on a large number of experiment with real and artificial data sets. As mentioned before the number of iterations is more important when using the algorithm for the purpose of baseline correction as signals are generally more frequent than random outliers in scatter plots. Additionally signal influences are correlated and usually affect a larger number of neighboring data-points requiring a higher number of iteration steps.

Neighbourbood size

The neighborhood h is a sensitive factor and often changing the neighborhood size has the most visible impact on the final result compared to other parameters. Larger neighborhoods generally result in smoother curves while smaller neighborhoods approximate local changes more accurately. For the purpose of baseline correction it would seem that larger neighborhoods are consequently more suitable than smaller ones. This is generally true as very small neighborhoods are clearly suboptimal for the purpose of eliminating distortions, however determining a neighborhood size that is optimal for a variety of baselines and more importantly a variety of signals which raise the complexity of the estimation task is not trivial. Neighborhood-size usually needs to be fit to the data characteristics but often a single global neighborhood-size will not produce optimal results.

Strengths and weaknesses

Robust locally weighted regression is able to automatically adapt to a given measurement and identify points that do not follow the overall trend. Outliers are determined by their distance to the regression function and their influence on the next iteration is reduced by automatically adapting their relevance or weight in the calculation. With regard to scatter plots the criterion which determines the relevance of a data point is the density. Points in close neighborhood to other points are more relevant than isolated ones. This allows the simultaneous correction of positive and negative errors. The neighborhood function is constant in Cleveland's original work but can be easily adjusted to a variable function without compromising the algorithm, however determining a suitable neighborhood for each data-point is a challenging task. Furthermore influences on the estimated curve are defined locally meaning that changes in the intensity of data-points only affect a local region around those data-points which is defined

by the neighborhood function. In contrast to this, algorithms using polynomial fitting cannot guarantee locality of changes because the fitting is done for the entire domain. Generally a piecewise or local fit strategy mimics the human workflow more closely making locally contained influences easier to understand and predict. Weaknesses of the method with regard to its application as a means to estimate baseline distortions are the underlying, implicit assumptions about singular outliers and the overall prevalence of the target function. The prevalence of the target function, in this case the distorted baseline curve sans signal features, is a condition that is implicitly assumed by all baseline approximation methods. If signal features are too dense baseline correction (without additional external information) is impossible. On the other hand while true outliers are commonly isolated the key idea behind using robust locally weighted regression to estimate baseline distortions is that signals can be interpreted as outliers. However clusters of relatively dense signal features are a common occurrence and thus must be treated accordingly. This can be done by adjusting the neighborhood dynamically allowing the algorithm the bridge across larger gaps of irrelevant data. Furthermore while a fixed number of iterations $t = 2$ might be sufficient for the purpose of smoothing scatter plots the distribution of data-points in measurements containing signal features could benefit from more robust steps and if possible a dynamic number of iterations. The assumption of local linearity can be interpreted as the most neutral regression model for the purpose of baseline estimation as even low grad polynomials can easily follow singular peaks and thus are not suitable for the purpose of baseline correction. This idea is similar to simple linear interpolation proposed by Prakash et. al. although the realization is different [PW11]. Robust locally weighted regression replaces data-points separately and uses several neighboring data-points for each calculation while the principal of linear interpolation in [PW11] replaces cluster of data-points by an interpolation between the start and end point of the cluster. So while robust locally weighted regression uses a group of data-points to linearly estimate a single value, linear interpolation uses two data-points to estimate a group of values. As a result the curves estimated using robust locally weighted regression generally do not contain truly linear segments.

7.2 Baseline correction using robust locally weighted regression

Given the original parameters defined by Cleveland for which robust locally weighted regression was designed, some of the disadvantages when using it for the purpose of baseline estimation come as no surprise. Specifically the algorithm was developed to accommodate data that satisfies

$$y_i = g(x_i) + \varepsilon_i, \tag{7.6}$$

where g is a smooth function and ε_i are independent random variables with $E(\varepsilon) = 0$ and $\sigma^2 = const.$ However, with regard to the problem of baseline correction the data can be more accurately modeled as

$$y_i = b(x_i) + s(x_i) + \varepsilon(x_i), \tag{7.7}$$

where $b(x)$ is the smooth baseline function, $s(x)$ the signal and ε_i a random variable representing the noise distortions. The idea behind using robust locally weighted regression to approximate a baseline from a measurement containing signal features as well as noise and baseline distortions is to interpret signal features as outliers. The weights of outliers are then automatically set to zero during the iteration process so that they no longer influence the calculation of the local regression.

While the contributions of signal features are assumed to be sparse compared to the entire domain of y and influencing singular data-points, signal features are often not evenly distributed and are characterized by a correlated influence and a number of adjacent data-pints while signal free areas are influenced only by $b(x)$ and $\varepsilon(x_i)$.

7.2.1 Local weight adaptation

The adaptation of weights is a key aspect of Cleveland's method as it represents the means which allows the algorithm to automatically adapt to the data at hand, identify outliers and remove their influence from the calculation of a smooth approximation. While the tricube function given by equation 7.5 defines weights based on the position in the neighborhood the bisquare weight function given in equation 7.3 is used to determine the robustness weights. In the original definition robustness weights are based on the undirected distance between fitted regression values $\hat{y}$ and original data values y. In other words, the distance between original data and regression curve is an indicator if the current point is an outlier. The greater the distance between $\hat{y}(i)$ and $y(i)$ the less weight is assigned to $y(i)$. The threshold determining the distance at which a weight of zero is assigned to a point, thus essentially removing it from the calculation of the regression curve, is defined by equation 7.3 in Cleveland's original article and can be rewritten to

$$t = 6 \cdot med(|e_i|). \tag{7.8}$$

While not explicitly stated by Cleveland the threshold choice can be related to the estimated standard deviation of the expected error which in the case of measurements is equal to noise. The threshold given by Cleveland can be rewritten to

$$t = b \cdot \sigma, \tag{7.9}$$

which based on the median absolute deviation, see section 3.5, gives a factor of $b = 4.05$ based on the estimated noise standard deviation, see also [RJFD01]. Lower factors of b describe stricter demands to data-points in terms of variability of intensity which generally result in smoother regression results while higher factors represent more robust estimators which often stand for curves that follow the trend of the original data more closely. Rewriting equation 7.3 to be consistent with the interpretation of threshold based on estimated noise standard deviation renders

$$\delta(k) = \begin{cases} \left(1 - \left(\frac{e_k}{4.05 \cdot \sigma)}\right)^2\right)^2 & \text{if } e_k \geq 1.48 \cdot \sigma \\ 0 & \text{else.} \end{cases} \tag{7.10}$$

Note that equation 7.10 is no longer dependent on e_i but instead on the noise standard deviation σ. While this change is merely technical with regard to the original data in Cleveland's studies that consisted only of noise and a slow moving signal it becomes very interesting when facing data that is defined by equation 7.7. While the estimated error e_i in equation 7.8 is clearly influenced by signal features $s(x)$, σ in equation 7.9 and equation 7.10 is not.

7.2.2 Gaps in the estimation

Assuming a normal distribution of residuals a lower factor b in equation 7.9 results in a larger number of data points classified as outliers. For a function of pure zero-centered, normal distributed noise a factor of $b = 3$ is equal to 99.7% of all data-points used for the regression, following the relation of standard deviation and distribution function. Thus the choice of the threshold $b = 4.05$ for separating outliers from relevant points is theoretically well founded. However to understand the method and its behavior regarding real measurements it is important to be aware of the assumptions made. The residuals between the regression result and the original measurement are interpreted as noise while the residuals actually represent a mixture of noise and partial signal features. With regard to baseline correction the ideal final residuals ideally consist of noise and all signal features, while during the robust iterations the residuals cannot be as sharply defined since the estimation of baseline distortions is per definition imperfect during this

stage of the algorithm. While the estimation of the standard deviation using the global median is robust against outliers, measurements containing a high number of signal features can influence the results. Furthermore even disregarding signal influences the assumption of constant noise intensity, which is inherent in the usage of a global median, is often not true for real measurements. Facing variable noise conditions and/or wide areas of signal features can result in situations in which entire regions of the original domain cannot contribute to the estimation of a suitable baseline. Considering the original static neighborhood definition it is possible that neighborhoods only contain data-points that are classified as irrelevant either because of signal features or because of noise intensities in that particular area of the measurement that are higher than the median estimation based on the entire domain. This leads to obvious errors in the estimation as no regression values can be calculated of single points or intervals of points resulting in a curve containing gaps.

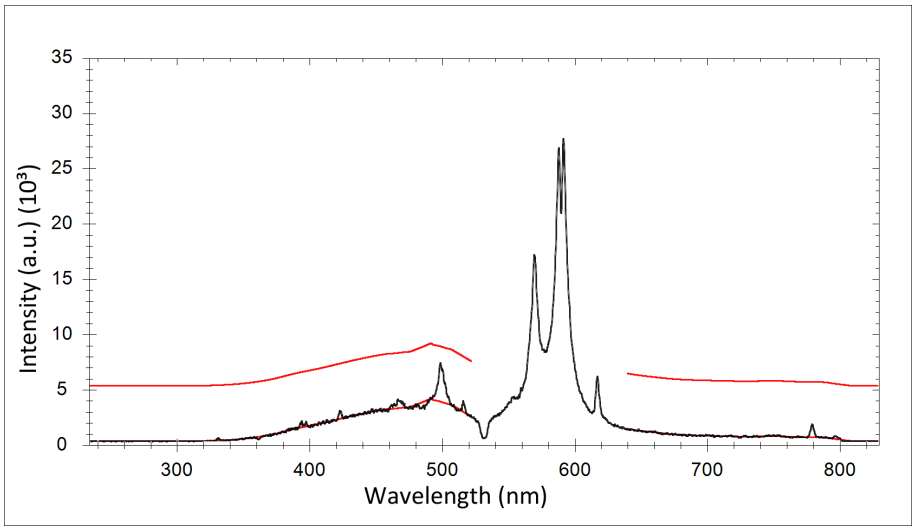

Figure 7.1: Illustration of possible gaps in the regression curve due to fixed window sizes and wide areas that do not contain data-points suitable to contribute to the estimation. The resulting local regression curve (red) is illustrated twice, once in its original position and once shifted upwards for better visibility.

Local threshold estimation

One possible option to prevent gaps in the regression curve is to change the threshold from a global value to a locally defined value. An example is the Matlab implementation of Loess which uses the median absolute deviation of neighborhoods to determine data-point weights. This guarantees that at least half of the points within a neighborhood contribute to the local regression. However this purely local estimation cannot bridge regions of dense signal features well simply because of this. If half of the neighbor-

hood points contribute to the regression, the maximum deviation from original curve is somewhat fixed and in some regions might be too restrictive for the purpose of baseline estimation. Since only a limited number of data-points are used to estimate the correct error threshold the estimation becomes less robust. Additionally in regions containing a high signal density or wide signals it cannot be guaranteed that at least half of the neighborhood points are not affected by signal features. Thus the risk that the data points used for the regression contain signal features is significantly higher and as a result the influence of the signal on the estimation is higher than in global approaches. In other words, gaps in the estimation curve are avoided at the cost of freedom in the estimated curve. Figure 7.2 gives an example of Loess regression using local error estimations.

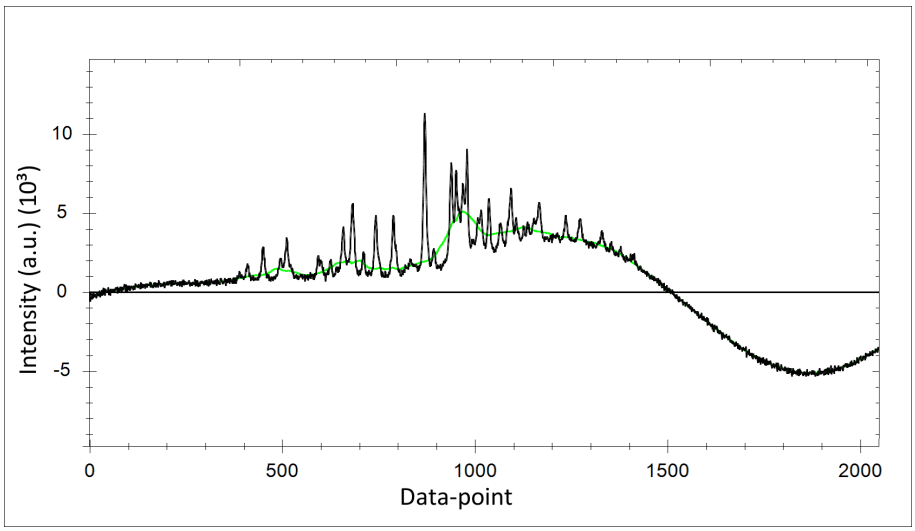

Figure 7.2: Illustration of locally weighted regression that uses local error criteria to avoid gaps in the regression.

Local neighborhood adaptation

Another way to avoid gaps in the regression curve is an adaptive neighborhood definition which prevents empty neighborhoods by expanding the boundaries adaptively. Adjustment of the neighborhood size is primarily a tool that allows the algorithm to bridge local clusters of outliers without the need for a global increase in neighborhood size that would increase the risk of over-smoothing. In local neighborhoods a high number of data-points can be irrelevant for baseline approximation and in extreme cases a local neighborhood could even contain no relevant data-points, making a sensible regression impossible. Choosing larger local neighborhoods can solve this problem but can also create difficulties when dealing with baselines of relatively strong curvature. In that case a large neighborhood will result in a misclassification of baseline areas as

outliers which effectively results in over-smoothing. To address the issue of sparsely populated or even empty neighborhoods which lead to poor regression results it is possible to change the definition of neighborhoods from a purely distance based approach to one that takes the weight or relevance of data points into account. The key idea, to postulate a fixed number of data-points in a neighborhood, but only count those which actually contribute to the regression, was proposed by Ruckstuhl et al. [RJFD01]. This implicitly creates two types of neighborhoods, the effective neighborhoods,

$$\mathcal{N}_k^{eff} = \left\{ (x_i, y_i) | (|x_k - x_i|) \frac{1}{h_k} \leq 1 \right\}, \tag{7.11}$$

which are of constant size but only count those data-points that possess weights greater than zero and the full neighborhoods,

$$\mathcal{N}_k^{full} = \left\{ (x_i, y_i) | (|x_k - x_i|) \frac{1}{h_k^{eff}} \leq 1 \right\}, \tag{7.12}$$

which vary in size depending on the number of outliers contained. In the original paper by Cleveland h_k is defined as the r^{th} smallest number among $|x_k - x_i|$ for $i = 1, \ldots, n$. To characterize the effective neighborhood, h_k^{eff} is defined as the r^{th} smallest number along $|x_k - x_i|$ for $i = 1, \ldots, n$ with $\delta_k > 0$. The definition of h_k^{eff} ensures that data-points with weights $\delta_k > 0$ do not contribute to the regression. During each iteration step of the robust locally weighted regression algorithm the weights of data-points are calculated depending on their relative distance to the active data-point and the assigned robustness weights, analog to the original algorithm. In addition the neighborhoods are adjusted to meet the condition of constant effective neighborhood size. While this guarantees a gap free curve just as the local estimated threshold does, the proposal also possesses its own drawbacks. In the original algorithm relative distance represents a constant because the neighborhood size is constant; the distance based part of the total weight assigned to each data-point does not change during the iteration process while the robustness weights provides a sense of adaptation to the data. Using effective neighborhoods relative distance is no longer static.

When using effective neighborhoods robustness weights that are set to zero implicitly also change the distance based portion of the weight calculation of all other data-points in every affected neighborhood. So locally adaptive neighborhoods not only introduce an additional variable to the adaptation process but also influence the weight calculation process. The risk inherent to the automatic extension of neighborhoods is a reciprocal effect with the calculation of the robustness weights. Neighborhoods grow but if the regression fails to approximate the effective points to an acceptable degree,

this in turn causes more data-points to be marked as irrelevant and creates a circular dependency of cause and effect. The resulting regression curve cannot contain gaps but the quality of the regression as a baseline estimate can suffer severely. While regressions using locally estimated thresholds often follow the original data too closely regressions using effective neighborhood display opposite characteristics and sometimes cannot follow the original curve close enough, especially in scenarios containing curved baselines. To illustrate a drawback of variable neighborhoods, figure 7.3, shows a real Raman spectrum and the result of baseline approximation using the modified locally weighted regression based on effective neighborhoods.

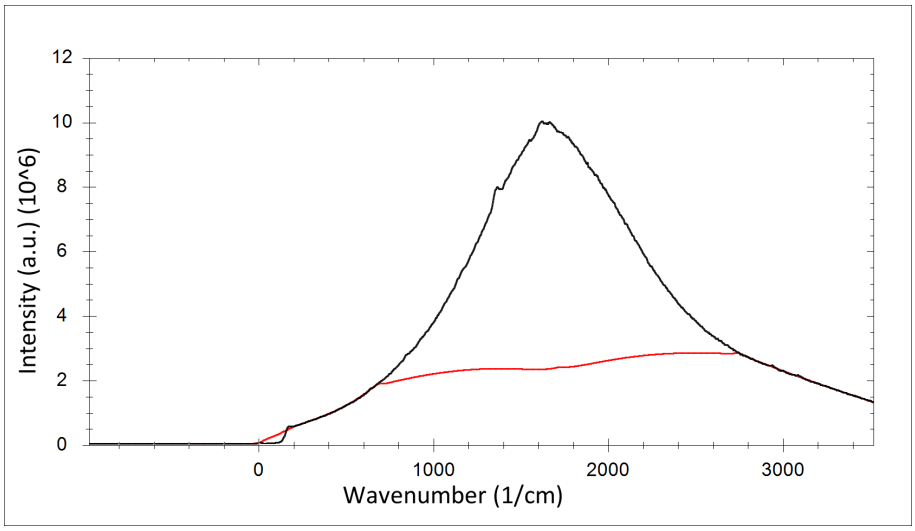

Figure 7.3: Example of baseline correction using variable neighborhoods on a Raman measurement showing a high fluorescence background. Shown are the real Raman measurement (black) and the estimated baseline using the baseline estimation method proposed by Ruckstuhl et.al. using adaptive neighborhoods

The measurement is dominated by strong fluorescence showing only a few signal features however since no artificial data is used the specific signal, noise and baseline components of the measurement are unknown. The cause for the poor performance of the method using effective neighborhoods in the example is caused by two characteristics. The first is the strong baseline modulation - the baseline in the measurement shown in figure 7.3 is, even though it is not accurately known, obviously not flat. The second reason is the modulation of noise. The difference in scale of noise and background fluorescence makes it difficult to judge the noise levels directly therefore figure 7.4 shows the highest detail scale of the stationary wavelet transformed measurement using the Daubechies 4 wavelet. It can be observed that coefficient intensities appear to be significantly higher towards the center of the measurement and lower towards the left and right edges. Using a global tolerance threshold based on the median deviation data-points thus causes too many data-points located in middle region to be labeled as

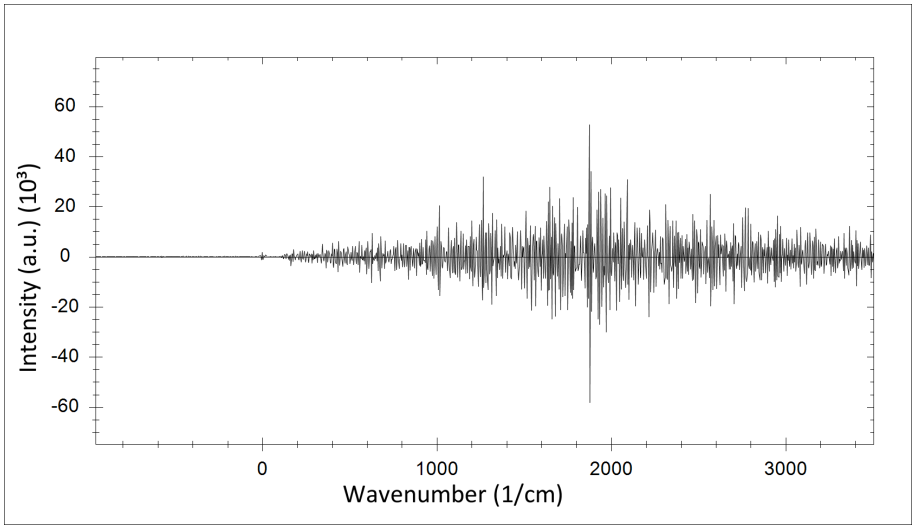

Figure 7.4: Illustration of the shift invariant wavelet transform fine detail coefficients of the measurement show in 7.3 to illustrate probable variable noise conditions found in real measurements.

outliers during the robust iteration process. The effective neighborhood compensates for this by iteratively adding data-points that are located further to the left and right and consequently yield a baseline estimation which a human observer easily - and with a high probability correctly - classifies as inaccurate. The problem that causes the erratic behavior in the example case can be described as follows: The initial fit marks several points in the middle of the measurement - the region of highest noise intensities - as outliers because their absolute residuals are too high. Since those points are now invalid the neighborhood region for points near the middle increases adding more data from the fast descending slopes of the spectrum. The next iteration yields a new approximation which due to the added data-points located on the slopes generates larger residuals for data-points that were still valid during the first iteration. The residuals increase faster than the median global residual because the flanks of the spectrum are still approximated relatively well leading to an effect that comes to a stop when approximately half of the spectrum is affected since by then the median increases very fast. The resulting baseline estimation does not agree with expert expectations and can be classified as bad or insufficient.

7.2.3 Adaptation conflicts

Aside from the above described difficulties effective neighborhoods add another challenging aspect to the adaptation process. A variable neighborhood size represents a second variable in the iteration process that needs to be adjusted depending on the distance between current regression and the original data. Experiments showed that

changing both, robustness weights and neighborhood size at the same time can have unwanted and unfavorable effects on the regression. Particularly in the first robust iteration steps correct for numerous outliers - which in this case also include signal features - neighborhood adjustment impedes on the robust regressions ability to adapt to the remaining data. The precise effects are hard to describe as they cannot be linked to particular measurement features like wide signals or strong baseline curvature. However in some cases the fact that neighborhood size and robustness weights are adapted at the same time results in strangely shaped regressions that do not reflect the original data. These effects appear to be caused by random relations between original curve, the regression curve and the existing influence between neighborhood size and robustness weights. If adaptive neighborhoods are used the adaptation of weights and neighborhoods should not be calculated in the same adaptation step. It is advisable to allow several iterations for the weight adaptation process without changing neighborhood parameters and adapt neighborhoods only when a stable state for robustness weights is reached. This does not solve the problem of over-smoothing curved baselines in general, but reduces the erratic reciprocal effects of simultaneous weight and neighborhood adjustment.

7.3 Adaptive local regression technique

While robust locally weighted regression was not developed to serve as a tool for baseline estimation, concepts like local adaptation and identification of outliers are very useful properties for exactly that purpose. Ruckstuhl et. al. [RJFD01] proposed a slightly altered version of Cleveland's algorithm to estimate baselines. However, as shown in section 7.2.2, the introduced alterations are prone to cause errors in realistic scenarios and sacrifice the ability to perform well in environments that contain multidirectional distortions. Adaptive local regression (ALR) which is proposed here, is a new baseline correction technique designed to perform well in scenarios containing severe distortions, by automatically analyzing an interpreting the results of several robust regression with varying parameters.

7.3.1 Enhancements to the original algorithm

The original approach of the local regression algorithms uses the difference or error values between a changing estimation and the input data as a measure to determine whether deviations are caused by an outlier or noise variations. The assumption that the median difference is related to the noise standard deviation works well when faced with data that follows definition 7.6 but is doubtful for data that follows definition 7.7. Furthermore changing characteristics of noise pose a serious problem. To enhance the performance of the algorithm, the estimation of noise parameters and therefore

also the threshold which is used to separate outliers from relevant data is done using shift invariant wavelet transform. Estimating noise parameters via wavelet transform minimizes the effect of signal components and also offers the ability to estimate changing noise characteristics.

7.3.2 Robust iteration count

The robust iteration process creates approximations and removes the influence of those data-points that are classified as outliers based on the error threshold. Given a fixed window size the iterative process usually converges towards a stable curve that represents an approximation of the source data. However in some cases parts of the approximation are not stable but exhibits oscillating characteristics, meaning one or more regions constantly change between two - or sometimes more - semi stable shapes. This behavior is caused by unfavorable relations between a fixed neighborhood size and the distribution of data-points. The possibility that broad signal features or dense, overlapping features can cause the robust approximation method to create regions of zero influence has already been discussed in section 7.2.2. Similar problems can cause an oscillating effect when the weights of singular data-points in larger regions of outliers change from zero influence to non-zero influence or vice versa.

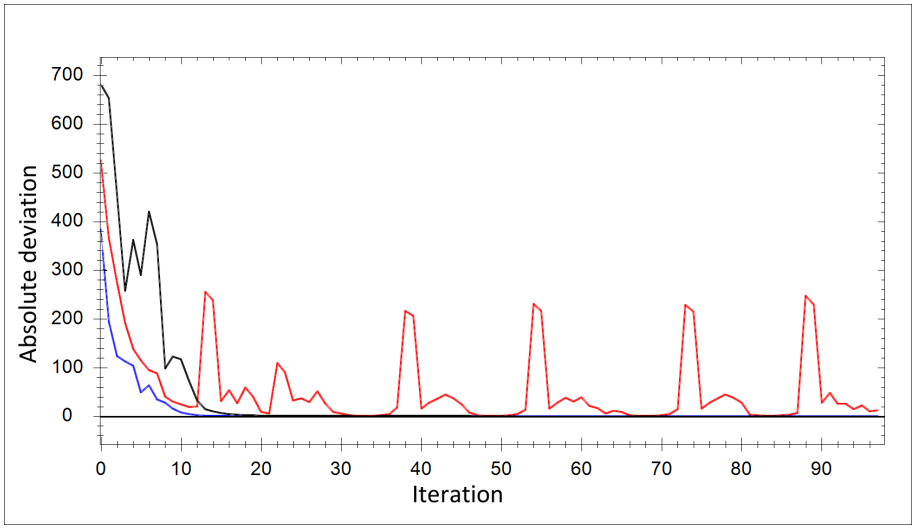

Figure 7.5: Example illustrating that global convergence of local regression curves in not guaranteed. The three curves shown belong to mean absolute intensity deviations corresponding to regression using fixed bandwidths of 0.02 (blue) 0.07 (red) and 0.09 (black) percent of the total measurements bandwidth.

In the original paper by Cleveland a convergence criterion is mention that could be defined so that the iterative process continues until the criterion is satisfied. However in the case of oscillating robust regression curves there is no global convergence and no

general distance criterion can be formulated unless the number of total iterations is a factor, see figure 7.5. A simple example could be a tolerable deviation that increases over time (iterations) making the initial value and function that defines the increase in tolerance the most important factors. The downside of such definitions is the introduction of an unpredictable runtime. In addition the problem of oscillating regression values is not solved; no matter what type of termination criterion is defined the state of the curve in its oscillating region when the criterion is reach is almost random. Without reliable means of determining a meaningful number of automatic iterations a fixed number of robustness iterations present the solely sensible alternative. It was already noted that in the context of baseline estimation due to the correlated nature of signal features more robust iterations are needed than in Cleveland's original proposal. While a fixed number of iterations appears to be a drawback, the fact that some regions of a regression curve do not achieve a stable state is not only an inconvenience in terms of automatic iteration, but also offers additional information that can be exploited for the goal of finding a baseline estimate. The excepted value E and standard deviation σ can be used to describe the behavior of data-point values during the robust iteration process. In order to obtain meaningful results for expected values and standard deviations the first robust iterations should be avoided. The values for a large number data-points change rapidly during these iterations due to the quickly changing influence of outliers. To separate stable from oscillating data-points it is therefore sensible to concentrate on robust iterations where stable points are expected to change only marginally. Let k be the number of robust iterations, then the expected values and standard deviations of data-points during the robust iteration process can be calculated according to

$$E_{robust}(x) = \frac{1}{k-s} \sum_{i=n-(k-s)}^{n} y_{x,i}, \tag{7.13}$$

$$\sigma_{robust}(x) = \frac{1}{k-s} \sum_{i=n-(k-s)}^{n} |y_{x,i} - E_{robust}(x)|, \tag{7.14}$$

with s denoting the number of iterations that are dropped from the evaluation in order to reduce the influence of initial changed to the overall expected and deviation values. In tests a number of $k = 30$ robust iterations and $s = 20$ were sufficient to achieve good results; fewer robust iterations carry the risk of using to many 'startup' values, while more robust iterations can improve results at the cost of increased runtime, however the exact values of k and s were not critical. Studying the characteristics of regressions that display oscillating characteristics in the global error criterion it becomes apparent that the part of the series that actually oscillates is small. In other words, it is not always possible to estimate a fully stable data series, but when examining the behavior of each data-point individually it is possible to differentiate between data-points that achieve

a relatively stable state during the fixed iteration process and a usually small amount of points that do not. However while not achieving a stable state do not allow for a sensible regression under the given parameters the information gained is still valuable. Sample points that do not achieve a stable state in the regression curve are not suitable for baseline estimation and their data values can be ignored.

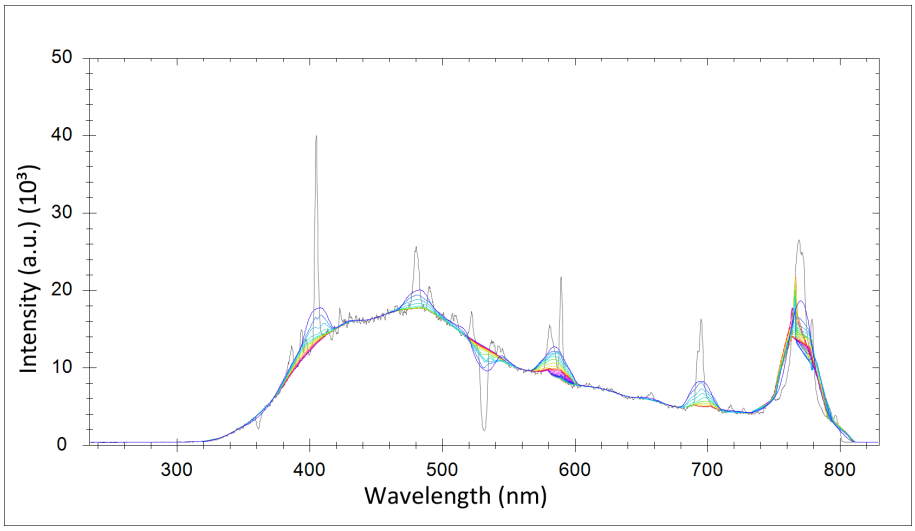

Figure 7.6: Robust locally weighted regression results using constant neighborhood size and iteration counts from $k = 1$ to $k = 30$

Identifying stable points

To discriminate stable from non-stable data-points a threshold is needed that limits the deviation of data-point values. The definition of a fitting threshold that defines allowed or tolerated deviations is challenging as characteristics of measurements can be vastly different. Absolute numbers naturally only work for very small subsets of possible measurements. Therefore thresholds should be derived from the measurement itself. Since the noise levels describes a measure of uncertainty related to a specific data-series a standard deviation that is below the noise threshold are marked as stable,

$$s(x) = \begin{cases} 1 & if \quad \sigma_{robust}(x) < f_{threshold}(x) \\ 0 & else. \end{cases} \tag{7.15}$$

The noise criterion may prove problematic for measurements containing no noise contributions. However this is usually not the case in real measurements and the threshold based on estimated noise characteristics has proven effective in a large number of tests with real measurements of varying characteristics. Furthermore given a noise threshold

that is not constant for the entire spectral domain but incorporates locally changing noise intensities the outlier rejection also inherits adaptive qualities. Given a non-constant threshold function accepted deviations are dependent on the position in the measurements domain. In other words the definition of outliers is dependent on the spectral region instead of a global criterion allowing better adaptation to neighborhoods of higher and lower overall noise characteristics similar to a local threshold calculation without losing the ability to identify larger groups of overlapping, or simply very broad signal features.

Summary

Possible oscillations can prevent a convergence towards fully stable regression curves. An analysis on different iterations of the robust regression that identifies so called stable and unstable data-points offers an alternative by identifying regions in globally unstable curves. Each curve or partial curve corresponds to a fixed neighborhood size and a fixed number of robust iterations. So, given a measurement and a bandwidth the convergence analysis results in a regression curve that may contain gaps in regions of the spectral domain that do not converge within a predetermined number of robust iterations. Figuratively speaking the uncertainty in runtime is eliminated at the cost of uncertainty in the resulting regression curve. However uncertain regions in the regression do contain the valuable information that sample points within those regions are irrelevant for the estimation of baseline distortions.

7.3.3 Variable bandwidth

The bandwidth or neighborhood size parameter is a very important factor determining the result of the regression curve. While the polynomial degree, weight function and iteration count also influence the regression result those parameters can be fixed and still achieve good results for a vast array of different data-series. The bandwidth on the other hand is the only variable that needs to be adapted to the data in order to achieve sensible results in terms of baseline estimation. Therefore it makes sense to say that given a sufficient number of robustness iterations and disregarding possible oscillation effects the smoothness and shape of a regression curve is determined by the underlying data - in this case the measurement - and the chosen bandwidth. The difficulties arising when attempting to determine the correct bandwidth automatically during the iterative process have already been discussed in section 7.2.2. Comparing the bandwidth parameter to Eilers' [EB05] smoothing method larger bandwidths could be interpreted as shifting the balance parameter towards the curvature criterion as larger bandwidths usually result in smoother curves while results using smaller bandwidths show similarities with shifting the balance parameter towards smaller distances to the

original data.

This does not mean that broad bandwidths are automatically more suitable in terms of baseline estimation than smaller ones. In fact regions that are void of signal features are often approximated most accurately using relatively small bandwidths since there are no outliers and the regression is only needed to compensate for noise influences. Without knowledge of signal distribution and baseline shape an a priori selection of the most fitting bandwidth is impossible. Other methods like e.g. airPLS which is based on Eilers' idea of using the Whittaker smoother to approximate baseline behavior, see chapter 4 and [ZCL10], avoid this problem by using a fixed global smoothness criterion that has to be selected by the user. Given a fixed bandwidth or smoothness parameter the calculation of the correct estimation becomes trivial. However even if it were possible to reliably select one global parameter to describe smoothness or local regression bandwidth in practice there might not be one parameter that is optimal for the entire domain. In fact small and large bandwidths both perform well under circumstances that are commonly found in most measurements without being exclusive.

Most real measurements include regions in which baseline distortions are best approximated using small bandwidths and also other regions which contain wider signal features thus making larger bandwidths necessary for a suitable estimation. The challenge that is essential for sensible baseline approximations is to find the right balance between the usage of wide bandwidths that can compensate for regions in which the baseline shape is masked by signals and small bandwidths that can accurately follow relatively strong curved baseline behavior.

7.3.4 Multiple bandwidths

An adaptation process that changes robustness weights and bandwidth at the same time is not advisable as both parameters also influence each other. Instead changes in the bandwidth parameter can be seen as adding an additional dimension to the estimation process and should be treated separately. Since real measurements have a natural maximal bandwidth defined by the measuring instruments the number of possible applicable bandwidths in the regression is strictly limited. In theory it is possible to try every possible bandwidth and analyze the results to estimate a suitable baseline. In practice the number of bandwidths can be reduced dramatically by incorporating knowledge about baseline behavior and observations about the regression results using varying bandwidth parameters. To limit the number of bandwidths the first step is to define lower and upper bounds that are more meaningful than the explicit limits given by the data. Assuming a measurement given as a series of equidistant data-points the absolute lower limit for the bandwidth parameter is given by the distance between two sample points $\delta_\epsilon x$. Any regression using this minimal bandwidth simply returns the original series. A more reasonable lower bound can be found combining knowledge

about the maximum signal frequency and assumptions about baseline frequencies. The maximum retrievable frequency of a sampled series following the sampling theorem by Nyquist and Shannon is given by equation 7.16,

$$f_{max} < \frac{1}{2} f_{sampling}, \tag{7.16}$$

with $f_{sampling} = (\delta_\epsilon x)^{-1}$, where $\delta_\epsilon x$ is the constant distance between two neighboring data-points. Assumptions about baseline behavior are unfortunately soft and not easily formulated as equations. It is generally assumed that the maximum baseline line frequency is significantly lower than the maximum signal frequency and that the baseline is smooth with respect to the measurement domain. To accommodate both requirements the minimal bandwidth in terms of baseline estimation can be defined by

$$w_{min} = \frac{x_{max} - x_{min}}{100}. \tag{7.17}$$

The value of w_{min} is somewhat arbitrary but captures the loose definition of a slowly changing baseline and is aimed to describe the smoothness quality with regards to be entire domain. With a reasonable minimum bandwidth defined the second step to refine the proposed approach is to select an upper bound bandwidth that is theoretically feasible and practically applicable. Analog to the absolute minimum bandwidth the upper bound bandwidth in practical applications is given by the data itself. In this instance the maximum bandwidth that is equal to the entire domain. When dealing with sloped or wave shaped baselines large bandwidths have serious detrimental effects on the estimation as even relatively low frequency changes are classified as outliers. On the other hand large bandwidths are arguably a reasonable choice for linear or almost linear baselines since eliminating the influence of outliers - in this case signal features - is the main target and large bandwidths have no detrimental effect. Since there is no minimum frequency for possible baseline the simplest case is linear baseline, either flat rising or falling is as feasible as a curved baseline and lowering the upper bound bandwidth estimation cannot be justified in the general case leading to the definition of w_{max} as

$$w_{max} = x_{max} - x_{min}. \tag{7.18}$$

A hard upper bound for the number of different bandwidths in $[w_{min}, w_{max}]$ is N, the number of data-points in the series. The exact number is lower than N as several smaller bandwidths do not need to be considered as they are below w_{min}, but since a flat baseline is a possible and in fact favorable realistic case an approximation that uses the maximum bandwidth, on first sight, remains a suitable approximation.

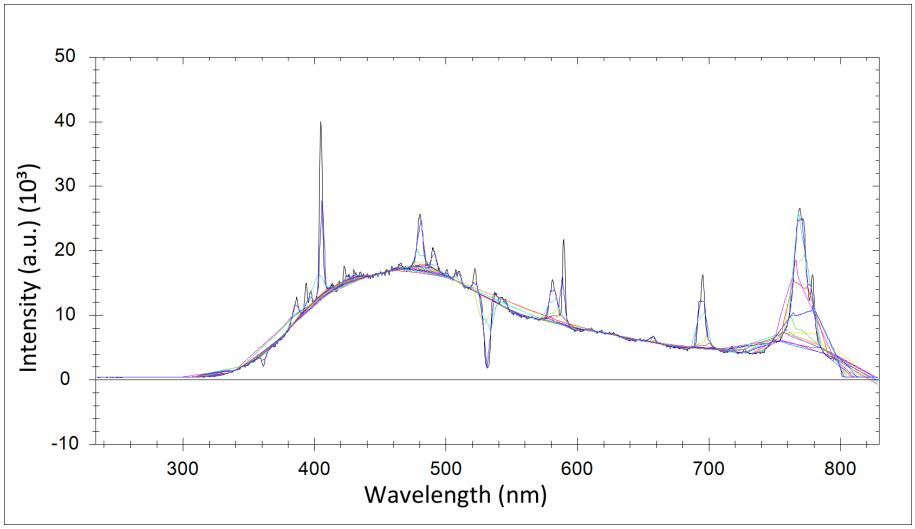

Figure 7.7: Robust locally weighted regression results using different neighborhood sizes and constant iteration count $I = 30$

7.3.5 Calculating suitable baseline points

Calculating a number of local robust regression curves using different bandwidth parameter is the first step towards a suitable baseline estimate. Each curve then corresponds to a fixed bandwidth and as such may not optimal for the entire domain containing signals of various frequencies. To combine the results achieved using multiple robust regressions of several different bandwidths the first task is the elimination of those estimated values in each curve that are not suitable for baseline estimation.

The regression curves corresponding to various fixed bandwidth generally contain regions that represent a suitable baseline approximation while other regions do not represent sensible baseline estimates. Reasons for poor regression results in terms of baseline estimation are directly related to the corresponding bandwidth. If a regression curve offers a poor baseline estimation in a particular region of the data-series the corresponding bandwidth is either too small, resulting in an estimation that follows the original data to closely or too wide thus not following the baseline distortions close enough.

Distance criterion

To eliminate those data-points that do not follow the data close enough a simple distance based criterion can be used. In other words the distance between original data-series and estimated value must be within a range that is proportional to the noise level. Following the arguments in 7.3.2 the noise level is chosen to specify the accepted distance because it can be derived from the measurement and because it can include

local variations. A fixed distance value can obviously only be only optimal for measurements with very specific features and even a constant factor could cause problems when encountering variable noise levels that are common within real measurement. In regions that are dominated by noise and baseline influences and contain no signal the robust regression estimate essentially represents a smoothed version of the original data-series while the iteration process and the adaptation of the robust weights eliminate influences of spectral features that are sharp in relation to the fixed bandwidth. The distance criterion obviously favors small bandwidths and cannot be used to correct for errors in regions where the regression curve follows signal features too closely. The distance calculation is given by

$$d_y(i) = \sqrt{(y_{source}(i) - y_{regression}(i))^2}, \tag{7.19}$$

while the threshold criterion describing the threshold at which the influence of sample points is set to zero given by

$$d_y(i) < y_{noise}(i). \tag{7.20}$$

Stability criterion

Since is it not possible to identify regression values that follow the original curve to closely by the distance between original and regression data another kind of criterion needs to be found. Given the limited knowledge about baselines and the goal to make a viable approach for a large variety of possible baseline shapes the criterion needs to come from the data rather than from an additional external source. Observing local robust regression curves of different bandwidths it becomes apparent that data-points more suitable for baseline estimation are usually more stable when considering the values at the same position in consecutive robust regressions of a monotone changing bandwidth. This new data dimension (along increasing bandwidths) is referred to as z-axis. The expected value of a data-point i in the z dimension is defined as

$$E_\zeta(i) = \frac{1}{Z} \sum_z y_z(i), \tag{7.21}$$

and the standard deviation σ_ζ accordingly as

$$\sigma_\zeta(i) = \frac{1}{Z} \sum_z \sqrt{(E_\zeta(i) - y_z(i))^2}. \tag{7.22}$$

While the first stability criterion determined stability in the dimension described by the iteration count t, the second stability criterion judges stability in the z dimension by comparing results achieved with varying bandwidth parameter. It is not possible to compare all possible bandwidths at once since huge differences in bandwidths will

always results in very different regression given a sufficiently curved baseline. There for stability is measured in subgroups and for each subgroup a set of stable data-points is created. To understand the reasons for this a few general considerations need to be made. Given a measurement that possesses a noticeable non-monotone baseline and disregarding any signal influences the value of almost every point will change considerably along the z axis, only baselines that are largely linear produce relatively constant regression results for changing bandwidths. Therefore it is very difficult to determine a suitable threshold for the allowed absolute deviation or allowed standard deviation over all instances of z that result in the desired filtering effect. However analyzing smaller subgroups adds robustness to the filtering process and because of the smaller expected deviations within the groups the noise intensity can be used as a sensible and locally adaptive threshold. Additionally while the criterion based on distance between original and regression data naturally favors smaller bandwidths, a constant windowed filter in the z dimension is - given a constant Δ_z - more restrictive towards smaller bandwidths than to larger ones simply because the relative change in bandwidth size decreases for larger bandwidths and thus the expected changes in the regression result are likewise expected to be smaller. The size of subgroups used to determine stable points corresponding to multiple bandwidths Z is an important parameter that has a major influence on the algorithm's results in areas underneath wide signal features. To determine if a data-point i is classified as stable the estimated baseline values at i are compared and evaluated according to the greatest absolute distance. If the overall greatest distance within a group is equal or smaller than the allowed error at i the point is stable in the examined group of baseline estimates and the expected value is used as representative value for the group estimate, if not it is unstable and its value is discarded of the examined subgroup. The parameter Z defines the group size of the described subgroups. Since the absolute error in smaller subgroups of consecutive estimates is generally smaller than in larger groups that contain baseline estimated with greater differences in their underlying bandwidth parameters a smaller parameter of Z corresponds to more lenient filtering while larger values of Z can be interpreted as more aggressive filtering, requiring data-points to be stable for a larger subset of estimations to be eligible for the final estimation.

Figures 7.8 and 7.9 show real examples of the effect different values of Z have on the final estimated baseline. In both examples three baseline estimates are shown which correspond to values of $Z = 5$, $Z = 8$ and $Z = 12$. A value of $Z = 5$ achieves a good result figure 7.8 but the estimated baseline using the same parameter does not fully match human intuition in figure 7.9. The same argument can be made for $Z = 12$ which leads to and intuitively good estimation in figure 7.9 but in turn the result in figure 7.8 leaves room for improvement. While the example measurements shown are extreme examples it shows that finding an optimal value for Z in every possible scenarios is a

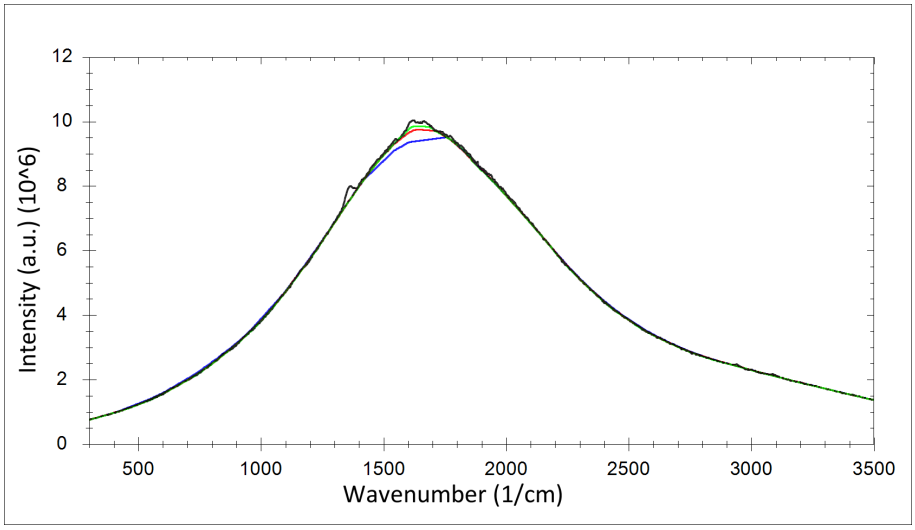

Figure 7.8: Illustration of baseline estimates using different stability parameters for a real measurement showing strong baseline distortions. Baseline estimations are shown for $Z = 5$ (green), $Z = 8$ (red) and $Z = 12$ (blue).

challenging task. The main difference to other baseline estimation techniques requiring a parameter that controls the tolerance of the estimation is that Z does not relate to a single estimation or to a specific value but to constant values between different estimates. In practice $Z = 8$ has returned reliably good results even in the extreme examples shown in both figures. For most real measurements containing less intense baseline distortions and/or narrower peak signals the Z value is not critical and $Z = 8$ returned overall good results for a wide array of different baselines. Due to the local nature of the estimation changing the value of Z within reasonable bounds has almost no effect on regions in which many data-points are available for the regression.

Optimizing the number of needed bandwidths

While the theoretical upper bound bandwidth cannot be reduced, practical applications show that two main reasons raise the sensibility of using very large bandwidths to question. The first very practical argument is simple runtime; the more possible estimations for different bandwidths need to be calculated the longer the cumulative effort for the final baseline estimation becomes. This alone is course no argument that allows for an omission of possible valuable information, however a closer look at the calculations reveals that very large bandwidths offer little additional information and can even skew the estimation results. Data-points that are within bandwidth range of the edges of the domain need to be treated differently than other data-points because no standard neighborhood around those data-points can be defined. With growing bandwidth that number of data-points which need special treatment grows so that for very large band-

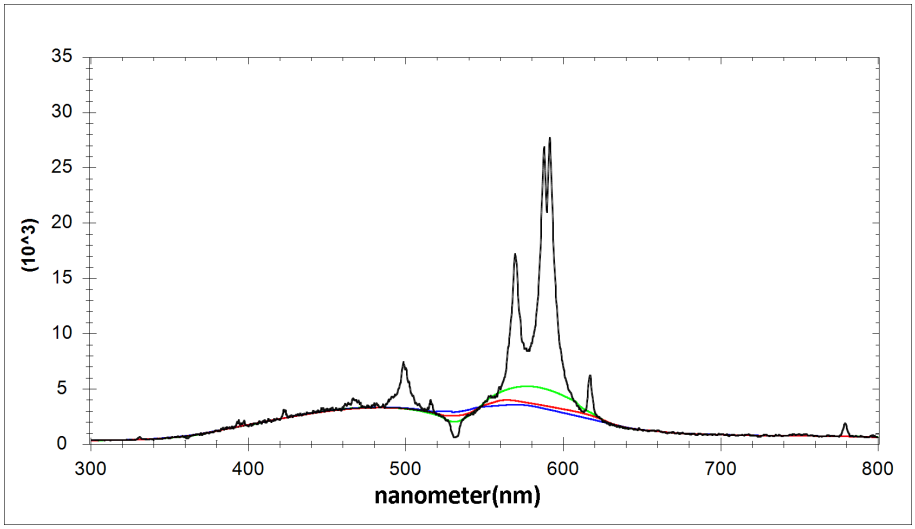

Figure 7.9: Illustration of baseline estimates using different stability parameters for a real measurement containing moderate baseline distortions and extremely wide signal features. Baseline estimations are shown for $Z = 5$ (green), $Z = 8$ (red) and $Z = 12$ (blue).

widths only a small fraction of data-points is submitted to the intended calculation while the larger fraction requires some form of exception regulation. This means that the pro arguments of very large bandwidths, namely being able to bridge large areas that offer no information about baseline shape and producing smoother results, only truly holds for a minority of data-points. On the other hand a bandwidth which covers e.g. twenty-five or thirty percent of the total domain can be considered large enough to compensate for regional changes without the need for bandwidths that grow beyond the fifty percent mark. Furthermore if a sizeable area of the domain offers no information about baseline behavior and thus requires very large bandwidths to allow a reasonable interpolation then the sensibility of such estimation could be called to question. Tests with real and simulated data have shown that for practical purposes bandwidths that span more than twenty percent of the entire domain are usually irrelevant.

This observation is also supported by reversing the argumentation. Generally speaking large bandwidth produce better estimations in cases where smaller bandwidths follow the data series too closely and thus result in baseline approximations that include strong influences of actual signals. In other words bandwidths greater than a given value x are required when a series contains signals that extend over a region greater than x. This can either be a single, wide signal or a very dense cluster of multiple signal features that are partially overlapped. Single, wide signals are usually not problematic as they possess large gradients leading to a gap in the estimation rather than a curve that follows too closely. Problematic constellations can arise in situations of low signal gradients and false plateaus. Low gradients cause a generally unavoidable

problem; if signals overlap to create a "mountain side" that is characterized by a low gradient a reasonable estimation of baseline characteristics becomes impossible without additional information. The same can be said for false plateaus as both constellations become locally indistinguishable from baseline distortions. The question is then, when a constellation is 'local'. Since there is no clear definition a local constellation which spans a region that is larger than twenty percent of the full bandwidth than can be assumed to actually be a baseline characteristic rather than an unfavorable signal cluster.

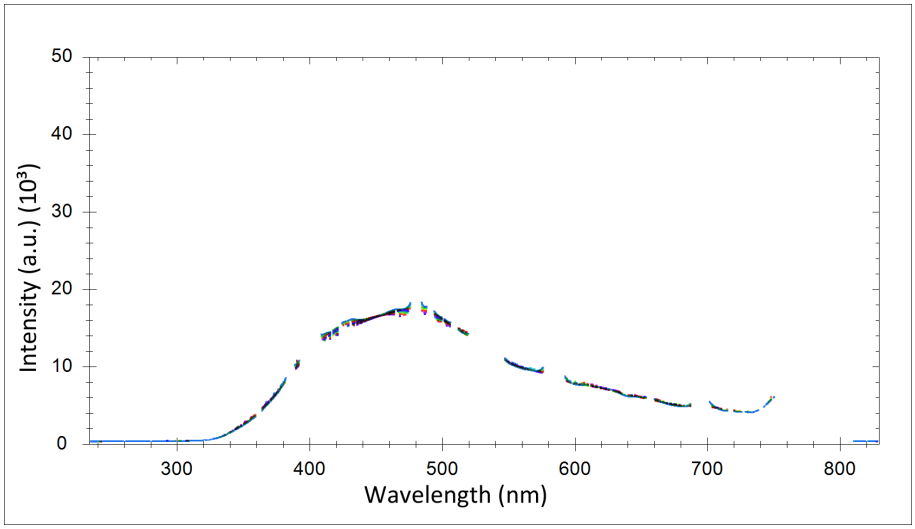

Figure 7.10: Filtered data-points using distance and stability criterions.

7.3.6 Gap closing

Analyzing the data series by filtering stable points in the two dimensions of robustness iterations and neighborhood size renders a set of points that are based on both stability criteria suitable for baseline estimation. In most scenarios some data-points will be over-defined meaning that more than one possible value exists that it suitable while other data-points will be under-defined, meaning that no possible value meets both stability criteria. To deal with the gaps created by under-defined points a three stage technique is employed. In the first step only points that correspond to the same stable neighborhoods are analyzed. The gaps in each set which are smaller than the mean neighborhood size of the set are linearly interpolated. The interpolation can also be done using low order polynomials but in practice linear interpolations have proven to be more robust as they are less prone to errors introduced by oscillations. This first stage closes small gaps in estimation set but leaves larger regions in which no suitable estimation sample point could be found unchanged and under-defined. Step two deals with over-defined data points. These can be handled in a variety of ways including averaging, median calculation, or center of gravity estimation however since

most outliers have been eliminated at this stage a minimum approach renders the best practical results. Step three then starts with a single set of data-points in which each point either possesses a single value or is undefined. Undefined points only appear in relatively large clusters as smaller clusters of singular undefined values have been eliminated in step one. These clusters represent areas in which no suitable baseline point could be derived from the original dataset. If the set of data-points does not contain undefined values the estimation is complete and the algorithms stops returning the dataset as the estimated baseline. If the set contains undefined values these values are linearly interpolated a synthetic noise signal is added according to the estimated noise parameters of the original data and the resulting dataset is submitted to the baseline estimation routine again.

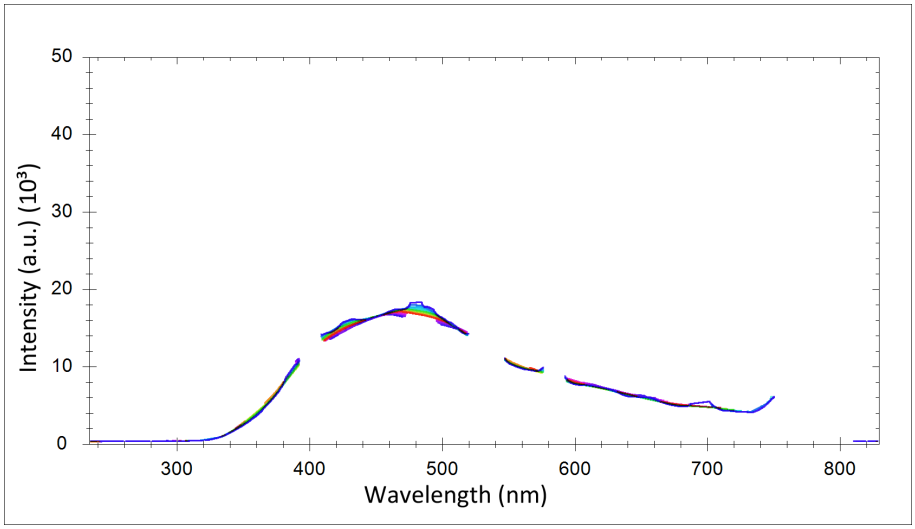

Figure 7.11: Filtered datapoint series after using the gap-close operation.

7.4 General characteristics and behavior

Since ALR is based on partial linear regression it can be shown that baseline estimation via ALR is equivariant against additive changes (shifting) and multiplicative changes of measured spectra. The linear regression of a series S consisting of data-points (x_i, y_i) can be described by

$$\hat{y} = a_S + b_S x, \tag{7.23}$$

where a_S and b_S are defined as

$$b_S = \frac{\sum_i (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}, \tag{7.24}$$

and

$$a_S = \overline{y} - b_S \cdot \overline{x}. \tag{7.25}$$

with $\overline{x}$ and $\overline{y}$ denoting the mean values of x_i and y_i in S. Under the assumption that the x_i remain unchanged, the behavior of the regression to additive and multiplicative changes of the source signal can be examined by varying the y_i. We define $g_i = y_i + k$ with the corresponding data series G consisting of data-points (x_i, g_i) and $h_i = y_i \cdot k$ with H consisting of data-points (x_i, h_i). For additive changes which represent simple shifts of the signal in vertical direction a constant value $k \in \mathbb{R}$ is added to each y_i:

$$\begin{aligned}
\sum_i (x_i - \overline{x}) \cdot (y_i + k - \overline{g}) &= \sum_i (x_i - \overline{x}) \cdot (y_i + k - (\overline{y} + k)) \\
&= \sum_i (x_i - \overline{x}) \cdot (y_i + k - \overline{y} - k) \\
&= \sum_i (x_i - \overline{x}) \cdot (y_i - \overline{y}),
\end{aligned}$$

which leads to $b_S = b_G$, meaning the shift operation does not effect b_S, and

$$\begin{aligned}
a_G &= \overline{g} - b_G \cdot \overline{x} \\
&= k + \overline{y} - b_S \cdot \overline{x} \\
&= k + a_S.
\end{aligned}$$

Let $\hat{y}^+$ be the linear regression of the data-series G then

$$\begin{aligned}
\hat{y}^+ &= a_G + b_G x \\
&= k + a_S + b_S x \\
&= k + \hat{y}.
\end{aligned}$$

Similar behavior can be shown for multiplicative changes to the source signal. Again assuming the sampling frequency is not affected changes to b_S are limited to the right half of the numerator in Equation 7.24:

$$\begin{aligned}\sum_i (x_i - \overline{x}) \cdot (y_i \cdot k - \overline{h}) &= \sum_i (x_i - \overline{x}) \cdot (y_i \cdot k - (\overline{y} \cdot k)) \\ &= \sum_i (x_i - \overline{x}) \cdot (k \cdot (y_i - \overline{y})) \\ &= k \cdot \sum_i (x_i - \overline{x}) \cdot (y_i - \overline{y},\end{aligned}$$

which leads to $b_H = k \cdot b_S$. Also it can be shown that $a_H = k \cdot a_S$:

$$\begin{aligned}a_H &= \overline{h} - b_H \cdot \overline{x} \\ &= k \cdot \overline{y} - k \cdot b_S \cdot \overline{x} \\ &= k \cdot (\overline{y} - \cdot b_S \cdot \overline{x}) \\ &= k \cdot a_S.\end{aligned}$$

Let $\hat{y}^*$ be the linear regression of the data-series H then

$$\begin{aligned}\hat{y}^* &= a_H + b_H x \\ &= k \cdot a_S + k \cdot b_S x \\ &= k \cdot (a_S + b_S x) \\ &= k \cdot \hat{y}.\end{aligned}$$

Let $B(M)$ be the estimated baseline of measurement M then $B(M) + k = B(M + k)$ representing the case of additive distortions and $B(M) \cdot x = B(M \cdot k)$ representing multiplicative distortions. Figure 7.12 and Figure 7.13 illustrate additive and multiplicative measurement transformation visually. Note that all components of the synthetic measurements - signal, noise and baseline distortions - were subjected to the same additive or multiplicative parameter respectively. If the components are treated differently in the transformation the resulting baselines are not equivariant as ALR works on the composite data.

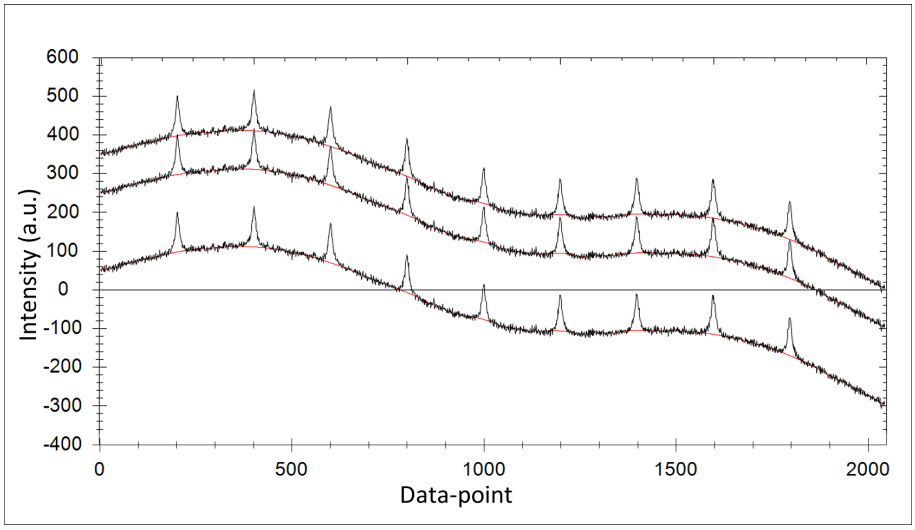

Figure 7.12: Illustration of additive changes on synthetic measurements and the corresponding baseline estimation. Additive factors are $M = 100$, $M = 0$ and $M = -200$.

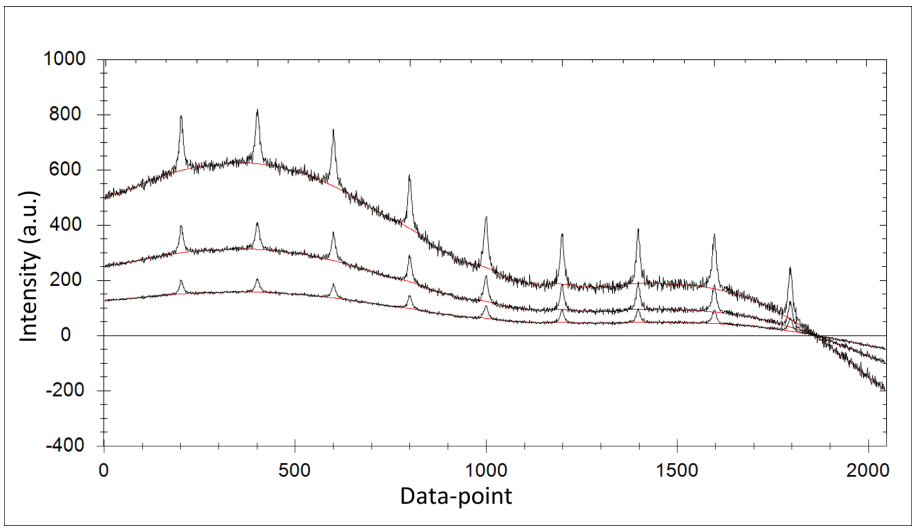

Figure 7.13: Illustration of multiplicative changes on synthetic measurements and the corresponding baseline estimation. Multiplicative factors are $M = 2$, $M = 1$ and $M = 0.5$.

7.4.1 Results of baseline estimation

For real data the result of a baseline correction method can be presented in two ways. One is the corrected data-series which allows a visual comparison of user expectation and the calculated result, while the other is an illustration of the estimated baseline itself. Since the real baseline is unknown no measure of error can be given. On the other hand synthetic data for which the correct baseline is known allows a simple error calculation. The mean squared error

$$E_{square} = \sum_i (y_i - \hat{y}_i)^2 \tag{7.26}$$

or mean absolute error

$$E_{abs} = \sum_i |y_i - \hat{y}_i| \tag{7.27}$$

are typical measures which can be used to quantify errors between a known correct baseline and an estimation derived from a data-series that represents a superposition of signal information, baseline distortions and noise interferences. Both measured possess averaging characteristics however the mean squared error also emphasizes larger singular errors which often models a human perspective in which many small errors are less severe than a single large error while the mean absolute error provides a more neutral evaluation. For the purpose of measureing an error in connection with spectroscopic data-series one has to keep in mind that all real measurements are affected by noise and that this noise introduces an uncertainty that is generally accepted. Various measures are taken to reduce these uncertainties but in the end it is generally accepted that no two measurements will ever be completely identical and small derivations are to be expected. With this in mind the mean squared error is the most suitable simple error measurement that can be given to describe the quality of baseline estimations in the context of synthetic data and known correct baseline. However this measure only takes the original and estimated baseline into account and ignores the signal. The signal's relation to the baseline estimation error plays a significant role in real scenarios and the relation between signal and baseline estimation error is similar to that between signal and noise. For large signals a baseline estimations error may be inconsequential, but if one imagines a signal becoming smaller while the baseline estimation error stays constant it is apparent that estimation errors are gain in significance with decreasing signal intensity. Two scenarios have to be considered. In the first case a signal could be completely eliminated by the baseline correction process because the estimation error is greater than the signal intensity. Some baseline correction methods prevent this by postulating a baseline estimation that is always below the values of the original data-series. This of course diminishes the estimation robustness towards outliers and noise. The other scenario is signal creation. An estimate can be off by being either too high or too low. An estimate that is too high may lead to signal loss, while and estimate that is too low leads to signals that are exaggerated in the corrected data-series. Taking a closer look at the relation between signal and estimation error reveals the same behavior for negative errors as for positive errors. The smaller the signal the more significant the estimation errors, as the relative change to the signal increases. Small signals can be changed to a significant degree and if the estimation is too low in an area that does

not contain signal then corrected data-series may contains bulges where it should be relatively flat. Assuming that the corrected data-series only contains signal and noise then leads to the misinterpretation of said bulges as relevant signal. Let $e_B(i)$ be the squared error between estimated baseline and correct baseline at i then the relation between baseline estimation error and signal intensity can be described as

$$C = \sum_i \frac{e_B(i)}{q(i)}, \tag{7.28}$$

with

$$q(i) = max(s(i), \sigma_{noise}(i)). \tag{7.29}$$

To give an impression of the effectiveness of ALR, errors measures between the estimated and correct baseline in synthetic measurements are calculated and compared to results achieved using airPLS. AirPLS was chosen as a reference because it is a general baseline correction technique that is not limited to a single type of measurement and performs well in a wide variety of synthetic and real scenarios. AirPLS is not adaptive which means that the results shown here are using standard parameter settings unless otherwise stated. Adaptive local regression possesses local characteristics while airPLS optimizes globally. Both methods present their own strengths and weaknesses which will be addressed and discussed here. Figure 7.14 shows a synthetic data set consisting of four signal peaks, normal distributed noise $\mu = 0$, $\sigma = 1$ and comparably low frequency baseline distortion. The function line represents the baseline estimation via airPLS the red function the baseline estimation via adaptive Loess. Both methods achieve good baseline approximations in regions without signal contributions and at the three smaller peak signals; the interesting region is the large signal at position index 200. The airPLS estimate follows the peak as much as the global curvature parameter allows forming a bulge beneath the peak that reaches farther up than the actual baseline does. The adaptive loess method is not able to identify data-points that are suitable for baseline estimation in the region beneath the peak and thus offers a partial linear interpolation, thus resulting in a flatter estimation than the actual curve. The weighted error functions reveal that the estimation achieved by Adaptive local regression offers the better approximation of the real baseline. This behavior stands for a typical case in which relatively wide peak signals and a curved baseline cause the global curvature criterion of the airPLS estimation to create suboptimal results. Peak signals often possess large tails which are characterized by a relatively low but steadily increasing curvature, without a reliable mathematical model these tails create problems for every baseline estimation method as there is no mathematical parameter that allows distinguishing them from baseline distortions which also possess low curvature characteristics.

Figure 7.15 gives another example of the problems caused by wide peaks. Other

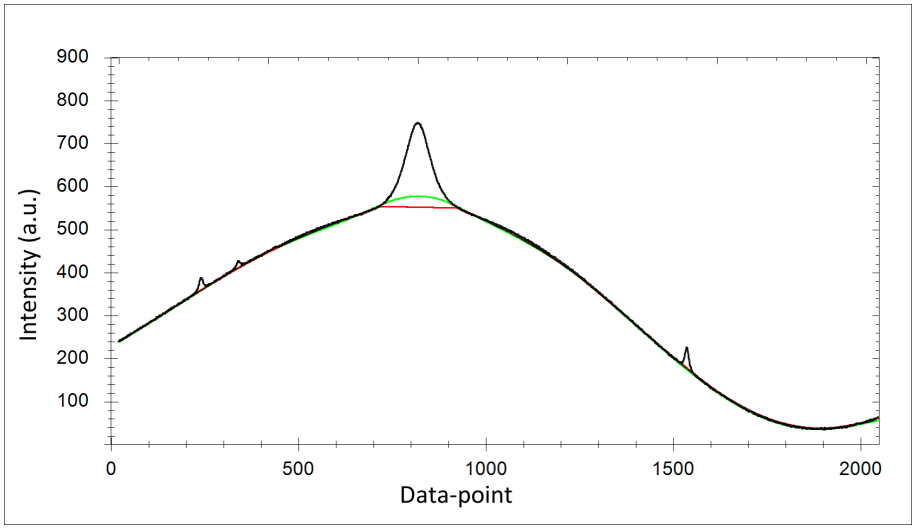

Figure 7.14: Signal consisting of three lorenztian peaks on a synthetic baseline created as a sum of randomized sine waves (Scheme 1) and afflicted with normal distributed noise N[0;1]. Mean squared weighted Error : airPLS = 7.56 ; ALR= 2.76

than the data in figure 7.14 the synthetic spectrum shown in figure 7.15 is a manual reproduction of a real spectrum created with the aim to create more realistic test scenarios, for more information see chapter 8.3. The original data contains a negative distortion caused by a notch filter in the measuring hardware visible as a negative dip at position index 530. Since airPLS by definition is not robust against negative distortions the dip has been partially removed similar to the fashion it would be removed in a real spectrum. Figure 7.15 contains two airPLS estimates, the green function represents the estimate achieved by using the standard parameters while the blue function uses relaxed curvature parameters. Using the standard parameters results in an estimated baseline function that is overall too flat and significantly too low for a large portion of the left half of the synthetic spectrum. Using more relaxed curvature demands creates an estimate that erroneously forms a bulge underneath the most prominent signal but also follows the baseline in the left portion of the spectrum more closely. However due to the global nature of the airPLS technique the estimation using relaxed curvature demands also demonstrates the risk of introducing undershoots. In approximate region between position index 450 and 520 the estimated baseline visibly undershoots the actual baseline. This behavior is indirectly caused by the wide peak; the estimation optimizes the global distance and global curvature and in minimizing the distance to the main peak signal by creating a bulge beneath it - thus reducing the distance to the peak top - the global curvature prevents the estimate from following the spectral shape more closely. Relaxing the curvature parameter even further would eventually remove the undershot but only at the cost of a severely diminished main peak signal. The

estimate achieved by Adaptive local regression is given by the red function. Since ALR is designed to produce reliable baseline estimations even in the presence of negative distortions the negative dip in the synthetic data was not removed. The complete synthetic spectrum and the baseline estimate are shown in the upper half of figure 7.15, note that spectrum and baseline estimate are shifted for illustration purposes only. The estimate calculated via ALR contains a bulge beneath the main peak similar to the one that can be observed in the airPLS estimate, the reason for this however is different. While airPLS optimizes the global distance between spectrum and baseline estimate ALR identified the main peak as not suitable for baseline estimation and thus ignores it almost completely. The bulge is caused by the wide tails, symbolically speaking Adaptive local regression follows the tails too far upwards before data-points are identified as not suitable. As a result the bulge is flatter and the remains of the wide peak resemble a tree stump. On the other hand the local character of ALR does not produce any undershoots and thus the baseline estimate in left region of the spectrum is very close to the original. The weighted error values confirm that the estimation via Adaptive local regression achieves a good approximation of the original baseline. Furthermore the relaxed curvature parameters for airPLS offer a better approximation in this scenario than the standard parameter values.

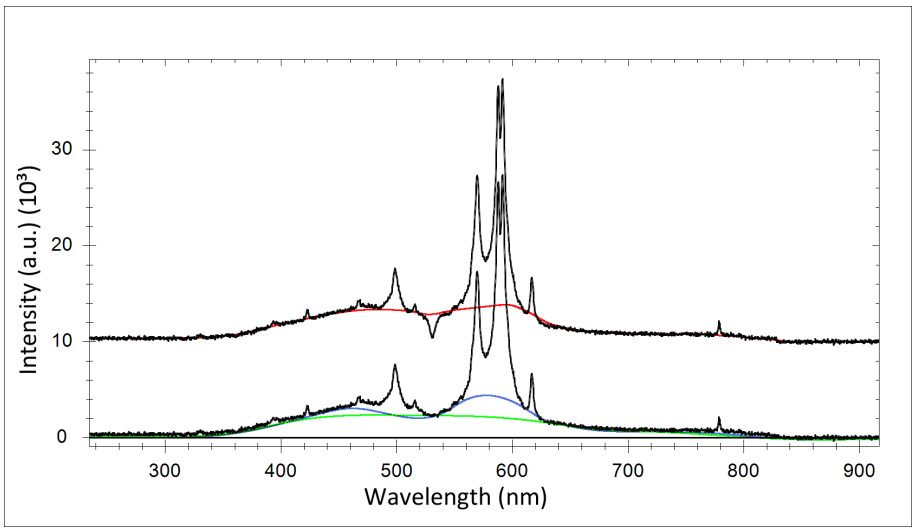

Figure 7.15: Illustration of baseline estimates by airPLS and ALR. Bottom graph shows the synthetic spectrum and two estimates by airPLS using parameters of 10^5 (green) and 10^4 (blue). The estimation by adaptive local regression (red) is shown in the upper graph which has been shifted for visualization purposes. Mean squared weighted Error: airPLS($\lambda = 10^5$) = 1401.4; airPLS($\lambda = 10^4$) = 950.9; ALR = 109.9

Up to this point two examples were discussed in which adaptive local regression offered more favorable results. However there are also circumstances in which airPLS is clearly superior due to its global characteristics. ALR relies on local data without

global controlling parameter, this offers advantages in cases one has to deal with very variable peak widths and baseline curvatures but it also hampers the methods ability to approximate baseline shapes in regions that are characterized by dense signal features. Figure 7.16 shows a synthetic spectrum that contains a relative large cluster of dense signal peaks. The red function represents the baseline estimation obtained using ALR the green function represents the estimate via airPLS. Visual inspection shows that the estimation via Adaptive local regression is more accurate in the areas that are dominated by noise as it is designed to run through the estimated mean of noise while the airPLS estimation always keeps below the noise. The same observation can also be made for the region which is dominated by dense signal features, the ALR estimation runs through the middle of those points that are identified as suitable for baseline estimation while the airPLS estimate again stays below. The main difference is that due to peak tails that add up in regions of dense signal features the baseline is actually below the lowest data-point value. This means that data-points that are identified as suitable for baseline correction are in reality located on the tails of peaks. Consequently ALR estimates the baseline using a local selection of data-points that are well above the actual baseline resulting in an estimate that is too close to the data-points. AirPLS is generally faced with the same problem however two characteristics of the method work in favor of the estimation in regions of dense signal features; one is the fact that airPLS generally stays below the original data values and two is global curvature criterion. Minimizing the curvature and staying below the original data-points results in an effective way to bridge across regions of dense signal clusters, given that all signals and distortions are positive.

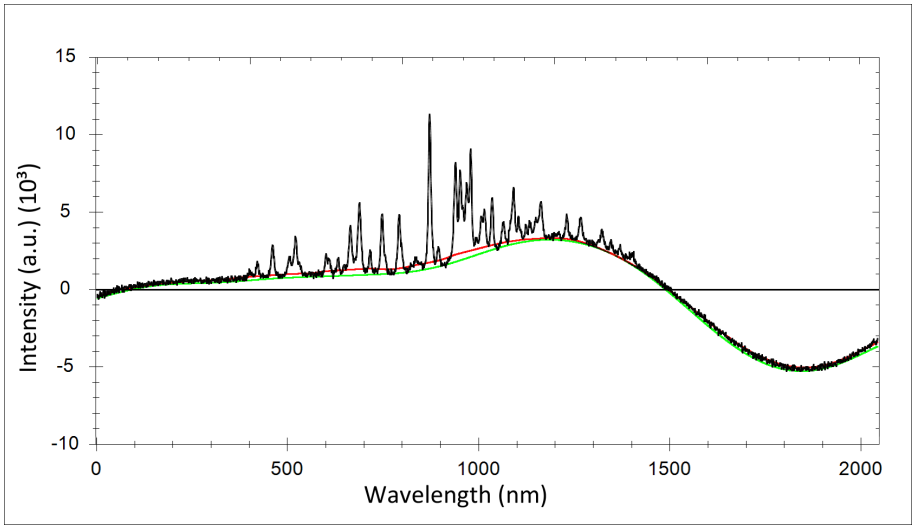

Figure 7.16: Example of a synthetic spectrum with dense clusters of peak signals. Mean squared weighted Error: airPLS(green) = 642 ; ALR(red) = 19278

It is interesting that the just described characteristics of airPLS and ALR which enable airPLS to estimate a more accurate baseline than ALR in the scenario shown figure 7.16 can have the opposite effect in a different scenario. Considering an environment that contains modulated noise ALR generally produced favorable results because the baseline is estimated within the noise while airPLS estimates a lower bound estimate that consequently estimates that baseline as too low in regions of high noise intensity. Figure 7.17 shows on example of high and low intensity noise in a single measurement and its effect on baseline estimation. Note that the negative distortion in the synthetic data at $x = 530$ was removed before applying airPLS to avoid additional negative influences on the baseline estimation result. The distortion did not have to be removed in case of ALR underlining the robustness of the method against distortions and variable noise parameters.

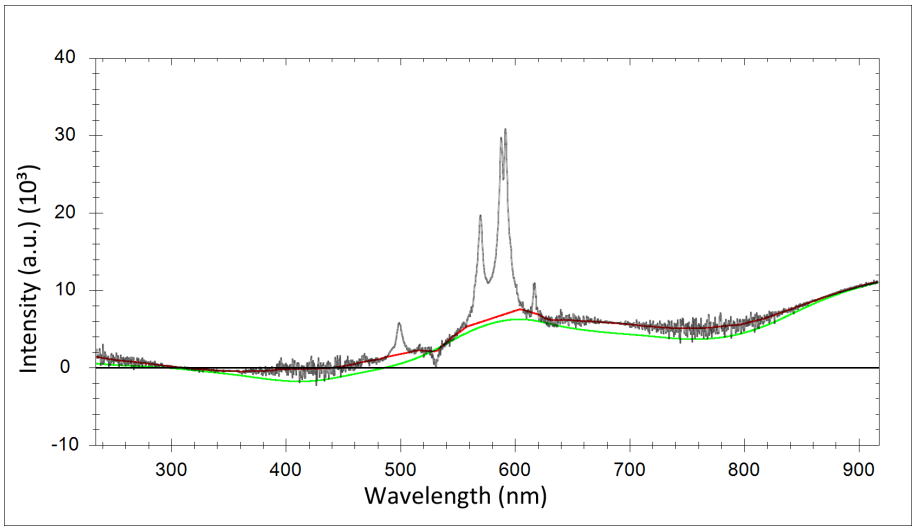

Figure 7.17: Example of a synthetic spectrum containing modulated noise intensities. Baseline weighted error : airPLS(green) = 7768.88 ; ALR(red) = 200.07.

It must be noted that the performance of baseline estimation techniques in general is always limited in relatively dense signal clusters. If signals are dense enough that the baseline is completely masked - meaning that no data-point has a value that is close to the actual baseline due to overlapping signals - over a significant region then no baseline estimation technique will be able to calculate a reliable approximation without relying on additional information.

To illustrate the effect of negative distortions on ALR and airPLS figure 7.18 shows a real Libs spectrum. As with all real data the correct baseline is unknown so no mathematical measure of error or distance can be given. The spectrum contains two negative distortions one at approximately 360nm of unknown origin and another one at 530nm which is caused by a notch filter in the measuring instruments to remove

the wavelength of the laser pulse from the measured data. The negative distortions have almost no effect on the estimation via ALR resulting in a baseline that contains no serious deviations from the overall baseline shape that a human observer would expect. The baseline estimate achieved by using airPLS on the other hand contains several obvious errors that can be described even without knowing the exact trend of the actual baseline. Both negative distortions cause severe deviations from the expected baseline. This behavior is to be expected since airPLS is designed under the assumption that negative distortions do not occur. Since the data violates the design prerequisites of airPLS this behavior is not truly faulty but instead an example of an unsuitable combination of data and estimation technique. The other flaw in the baseline estimation by airPLS mirrors the observations made in figure 7.14 and figure 7.15 of bulging and undershooting. The suspected superposition of peak signals at approximately 770nm is 'too wide' witch causes the airPLS estimation to follow the original data upwards creating a bulge. Simultaneously the curvature introduced by the bulge is compensated by an undershoot that ranged from approximately 640nm to 740nm exaggerating the signal features in that area.

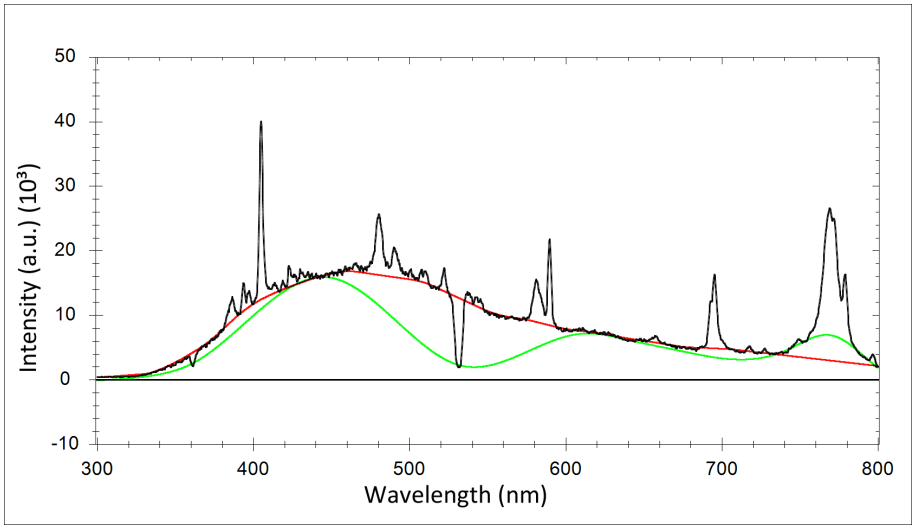

Figure 7.18: Illustration of a real Libs measurement and baseline approximations by airPLS (green) and ALR (red). Mathematical error measures are not given since the correct baseline is unknown in real measurements.

Summarizing it can be noted that both airPLS and ALR are powerful tools for the purpose of baseline estimation. AirPLS is well suited to estimate baseline shapes in environments that can guarantee that distortions and signal are directed in the same direction and signal peaks are not too wide. ALR is able to produce viable results in environments that cannot guarantee that distortions are unidirectional. ALR also able to handle a greater variation in signal widths and variable noise intensities, but in

return is less ideal when dealing with extended clusters of dense signals and distortions.

Chapter 8

Signal recognition results

After discussing and comparing the characteristics of adaptive local regression in chapter 7 this chapter illustrates the capabilities and limitations of baseline correction via ALR using a number of synthetic and real examples.

8.1 Signal comparison

To judge the effectiveness or accuracy of baseline estimation techniques the first solution that comes to mind could be a measure of error between the estimated baseline and the real, correct baseline. This of course implied that the real baseline is known and thus is not applicable in scenarios utilizing real data. In most real scenarios baseline correction is not the goal but simple a means to an end and the truly important information is the reduction in signal variability. Very often the real and simple question is: "Is an unknown measurement X similar to a measurement Y that is known?" Baseline distortions obviously make the answer to this question more difficult and consequently correction techniques are employed as part of data processing prior to actual comparisons. The quality of baseline correction methods can thus indirectly be evaluated by the calculating the similarity of measurements that contain the same relevant information but different baseline distortions. The similarity calculation should be robust against scaling since the absolute intensities of real measurements can be influenced by a variety of factors. For synthetic measurements this indirect method appears unnecessary since the correct baseline is known, but in order to have a consistent method to describe estimation quality it makes sense to use a method that can also be used for real measurements.

Correlation

The correlation between two functions describes the degree to which these function resemble each other. Correlation is expressed by the correlation coefficient c with

$c \in [-1, 1]$. A correlation coefficient of zero indicates that no relation between data exists, the datasets are independent. A coefficient of $c = 1$ indicates that the datasets describe the same behavior while a value of $c = -1$ indicates a countermotion, if function X rises function Y falls and vice versa. Mathematically the correlation between two random variables X and Y is given as the covariance of X and Y divided by the product of variances. Since spectroscopic signal are usually given as finite sets of data the correlation coefficient is described by

$$c = \frac{\sum_{i=1}^{n}(x_i - mean(X))(y_i - mean(Y))}{\sqrt{\sum_{i=1}^{n}(x_i - mean(X)^2)\sum_{i=1}^{n}(y_i - mean(Y))^2}}. \tag{8.1}$$

The correlation of two signals can also be described by the cross correlation which adds a translation parameter t to the definition above. However since it is assumed that signals are, except for minor errors, translation invariant only the cross-correlation coefficient at $t = 0$ is given for synthetic measurements. If for real measurements $t = 0$ does not represent a local maximum the cross correlation of the local maximum closest to zero and its corresponding t value are given instead.

8.2 Test conditions

To determine the effectiveness of distortion suppression techniques it is important to have knowledge of the undistorted values. Without such knowledge filtered values can only be compared to the original unfiltered measurement or a model of the pure signal. A model scenario is unrealistic in general since any sufficiently accurate signal model would implicitly solve the problems created by noise and baseline distortions. On the other hand comparisons with the noisy, distorted measurement are difficult because no measure of distance can be reliably linked to positive of negative filter influences. Since in practice measurements are never completely free of noise and baseline distortions the effectiveness of baseline estimation and noise reduction techniques is difficult to judge in an objective fashion. Synthetic spectra are a common way to eliminate the problem of an unknown optimization goal. Synthetic spectra use model functions like Gaussian or Lorentzian function to create ideal signals and add noise and/ or other distortions. Since the ideal spectrum which contains only the pure signal contributions is now known the results of any filter can be qualified by comparing them to the values of that ideal.

A general problem that arises when using synthetic spectra is the degree of realism or in other words how well good results obtained for artificial spectra translate to applications using real measurements. Synthetic spectra tend to unintentionally eliminate features that are common in real measurements such as unsymmetrical peak shapes,

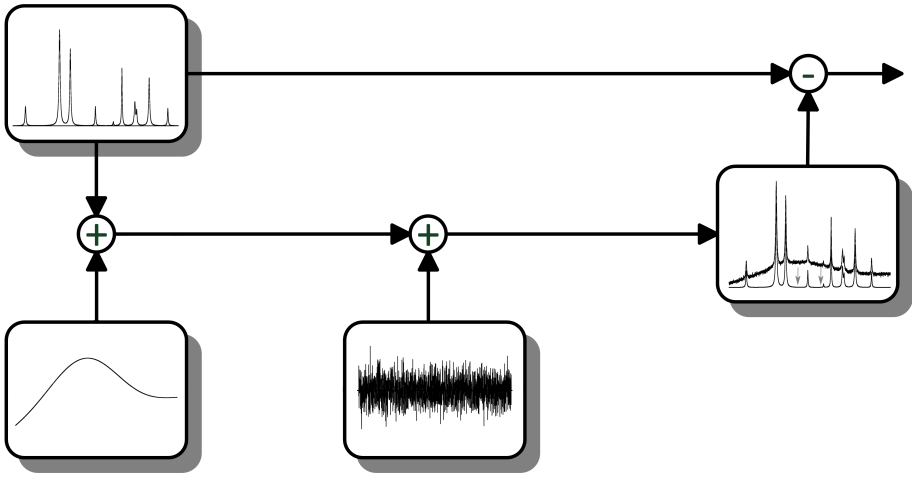

Figure 8.1: Block diagram illustrating the general test procedure using synthetic data. The + and − operators indicate signal addition and subtraction respectively. Each block represents a step in the test procedure. Signal generation, baseline generation, artificial noise generation and the baseline correction method. The difference between the corrected synthetic spectrum and the initially generated data can be used to quantify the quality of the correction method.

noise variations and irregular baseline frequencies etc. So while synthetic measurements usually consist of well separated singular peak signals, very low frequency baselines and constant noise levels real measurements often entail much less ideal environments. Signals and noise can vary greatly in intensity within a single measurement, baselines can spot distortions introduced by hardware filters and peaks can be distorted by several overlapping signals appearing as a singular, broad peak. While those idealizing effects of synthetic spectra cannot be completely eliminated in the general case as there is no model which is able to fully simulate real measurements, it is possible to make synthetic spectra more challenging by mimicking difficult to correct instances of real spectra. One must keep in mind that the mimicking synthetic version of a real spectrum does not guarantee a correct reproduction of the real conditions of signal or distortions but that it merely represents a way to test filter results against challenges that could occur in real measurements.

8.3 Synthetic data

Real spectra often present challenging environments for baseline estimation techniques as real data may contain superpositions of signal and several different interferences that are difficult to discern from a mathematical point of view. To test the quality of baseline estimates for real data series the underlying components of baseline signal and other interferences need to be known. This however is not the case in real scenarios as, generally speaking, real measurements always contain a degree of uncertainty that can-

not be modeled. Impurities, environmental conditions or minimal changes in measuring instruments are common causes for variations that cannot be predicted. Therefore to be able to mathematically measure the quality of estimates the estimation has to be tested on synthetic data. Synthetic data offers complete control of every component contained in a measurement, signal, noise, baseline distortions and if desired also other interferences. This chapter presents mathematical descriptions that can be used to create synthetic measurements which are relatively close to real measurements and offer total knowledge about all components contributing to the analyzed data.

8.3.1 Peaks

Signal peaks follow precise mathematical formulas which describe peak shapes based on the parameters of position x_0, height or intensity H and peak width w. Peak widths is either given by the full width at half maximum (FWHM) or half width at half maximum (HWHM). The most common peak formula is the Gaussian peak,

$$F_{GaussPeak}(x) = H \cdot e^{-\frac{(x-x_0)^2}{w^2}}. \tag{8.2}$$

The Lorentzian or Cauchy peak function describes a peak shape that is somewhat slimmer than the Gaussian peak and possesses long reaching tails,

$$F_{LorentzianPeak}(x) = H \cdot \frac{(\frac{w}{2})^2}{(\frac{w}{2})^2 + (x-x_0)^2}. \tag{8.3}$$

The Pearson VII function basically describes a Lorentzian function raised to the power of m. The additional parameter can be used to differentiate the exact peak shape,

$$F_{PearsonPeak}(x) = H \cdot \frac{w^{2m}}{(w^2 + (2^{1/m} - 1)(x-x_0)^2)^m}. \tag{8.4}$$

The pseudo Voigt function is an approximation of the Voigt profile using a combination of Gaussian and Lorentzian function. The true Voigt profile is created by convoluting Gaussian and Lorentzian function, however since this operation is computational costly the convolution is replaced by a linear combination for the pseudo Voigt peak function [SBC97]:

$$F_{PseudoVoigt}(x) = H \cdot (\eta \cdot F_{Lorentzian}(x) + (1-\eta) \cdot F_{Gaussian}(x)), \tag{8.5}$$

with

$$F_{Gaussian}(x) = e^{-(ln2)(x-x_0)^2/w^2}, \tag{8.6}$$

and

$$F_{Lorentzian}(x) = \frac{1}{(x - x_0)^2/w^2}. \tag{8.7}$$

8.3.2 Noise

Noise in synthetic data sets is described by statistical distributions. In the general case noise is simulated by normal distributed random variables with mean zero and standard deviation $\sigma = I_n$ where I_n depends on the simulated noise intensity,

$$F_{Noise}(x) \sim N(\mu, \sigma). \tag{8.8}$$

For each value of x the corresponding value of the noise functions is drawn independently from the distribution described by $N(\mu, \sigma)$. The quality or randomness of the simulated noise depends on the random generator used. Different random generators possess varying qualities in terms of unpredictability where the degree of unpredictability is generally correlated with computational effort. Simulation of randomness is a subject that will not be discussed here and it will merely be assumed that the created values are sufficiently random to be used as a simulation for the intended purpose.

8.3.3 Baseline shapes

In general there are no predefined formulas for simulating baselines in synthetic datasets. Specific spectroscopic methods may offer formulas for baseline phenomena that are caused by known physical events similar to peak functions. However in these cases a specialized form of correction also warrants better and more reliable results than a general approach to baseline estimation without additional information. Without robust mathematical descriptions of baseline shapes simulations can be made using low order polynomials and trigonometric formulas to guarantee that baselines are of low special frequency and overall smooth. In a sense the degree of freedom also shows a basic obstacle; when it comes to baselines, be it estimation or simulation, very little is known and almost nothing can be ruled out completely. A general polynomial description of a baseline function can be given by

$$F_{Polynom}(x) = a_n x^n + ... + a_2 x^2 + a_1 x + a_0. \tag{8.9}$$

To introduce a random element to synthetic baselines different shapes were created using sine waves of random amplitude, offset and frequency. Two schemes were chosen that calculated a random baseline as the sum of randomized sine waves. Sine waves are categorized in levels and each level L contributes exactly L sine curves to the final synthetic baseline. The maximum frequency is decreased for each level, meaning that there are statistically fewer sine waves of higher frequencies than lower frequencies con-

tributing to the synthetic baseline. In scheme one the maximum amplitude is constant in scheme two the maximum amplitude is also decreased with level. Offset is uniform randomized between 0 and 1 for both schemes and all levels. The achieved results are smooth and offer a good degree of variability however the two schemes are not exclusive and many other suitable methods to create randomized baselines exist. The synthetic baselines created using sine functions can be described by

$$F_{Trig}(x) = \sum_i a_i \cdot \sin(\pi(f_i \cdot x - b_i)), \tag{8.10}$$

with frequency f_i, offset b_i and amplitude a_i.

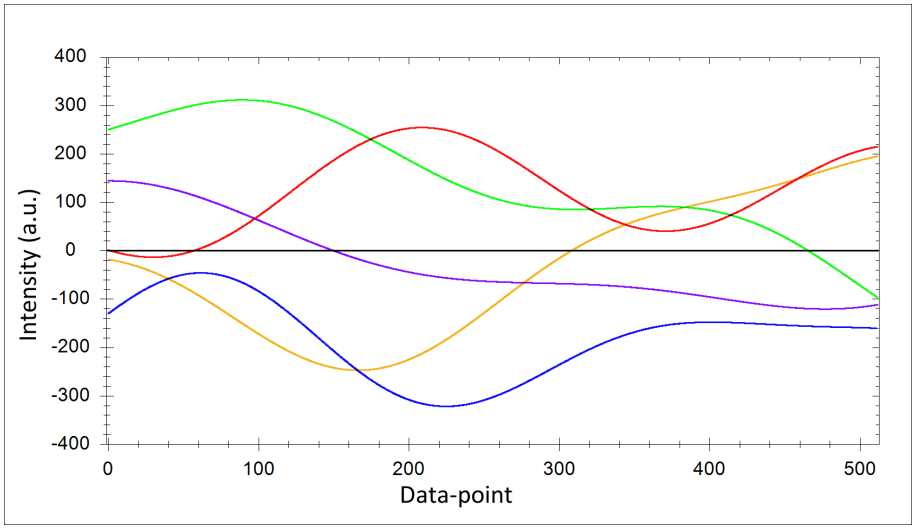

Figure 8.2: Examples of synthetic Baselines created with constant amplitudes (Scheme 1). Each data series consist of 2048 data-points with $\delta x = 0.25$. Shown baselines are created using four levels which corresponds to a total of ten randomized sine waves per synthetic data series. Maximum amplitude $A = 100$; maximum frequency $f_{max} = 0.0078125$, which corresponds to four oscillations over the domain.

Because the average amplitude and thereby the influence of curves from higher levels is smaller than in scheme one baselines created using scheme two generally possess fewer strong curvatures but more subtle changes. Both methods produce curves that are generally believable as possible baseline shapes and can be easily adapted to different domains and overall intensity values.

To create synthetic data sets that manage to capture the conditions in real measurements, another, manual extrapolation method was used to create what could be called a semi-synthetic measurement. The aim of this approach is to recreate selected measurements that exhibit baseline and signal characteristics which make an estimation challenging. Since the actual baseline these measurements is of course unknown the baseline shape is guessed by manually removing peak signals followed by manual

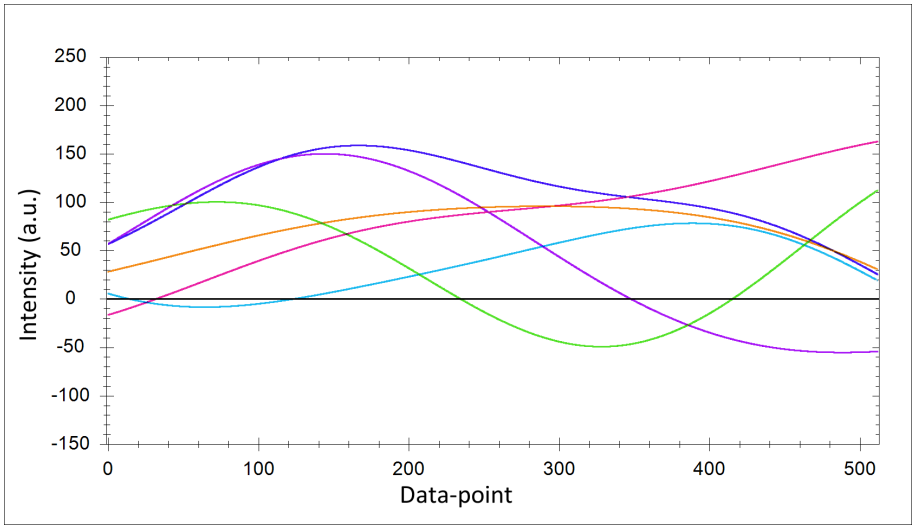

Figure 8.3: Synthetic Baselines created using decreasing amplitudes in increasing levels (Scheme 2). Each data series consists of 2048 data-points with $\delta x = 0.25$. Shown baselines are created using four levels which corresponds to a total of ten randomized sine waves per synthetic data series. Maximum amplitude $A = 100$; maximum frequency $f_{max} = 0.0078125$, which corresponds to four oscillations over the domain.

interpolation and smoothing of the remaining data. The resulting curve is based on the original data but also contains unspecific modeling assumptions that are reflected in the interpolations. This method is naturally biased as it relies heavily on the user's guesses to eliminate signal features and time consuming as each baseline must be created manually. Furthermore to approximate the signal features in the original measurement, mathematical peak descriptions have to be fitted to the data so that the combination of extrapolated baseline and fitted peak model resembles the measured data.

Figure 8.4 shows an example of the manual composition process. The basis is a real measurement that is to be recreated, and in a first step the baseline, as guesses by a user, is removed. Then the pure signal components are manually recreated by guessing and fitting peak modeling function to the remaining data until the approximation is deemed close enough to the original data to capture its main characteristics. If no sufficiently accurate synthetic spectrum can be created, the whole process has to be repeated starting with a different user defined baseline.

8.4 General characteristics and behavior

The quality of baseline estimation via ALR is related to the ratio of signal and baseline influences on the measurement. Dense signal features generally result in poorer quality in the estimation as illustrated in figure 8.5. The two signals shown in the figure contain identical baseline and noise ($\sigma = 5$) distortions, but differ in the density of artificial peak

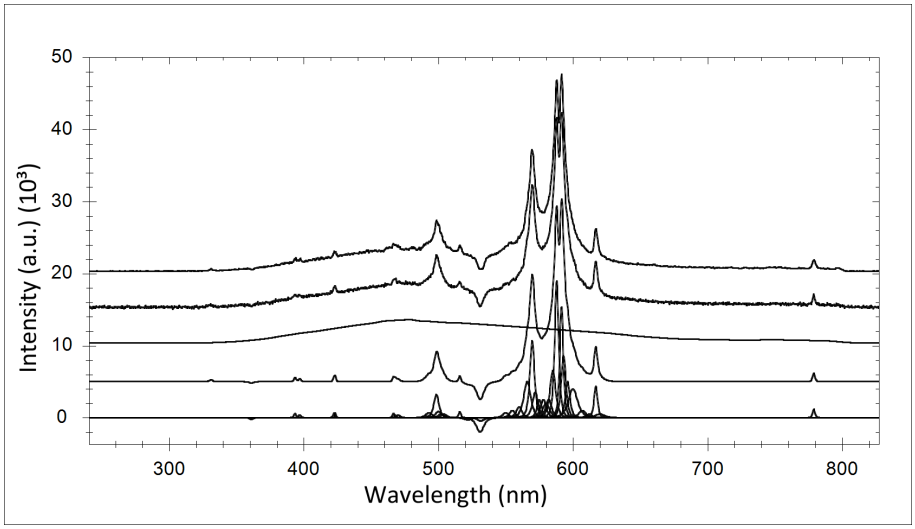

Figure 8.4: Illustration of the manual decomposition process necessary to create semi-synthetic spectra from real measurements. Shown are from top to bottom: Real measurement, semi-synthetic spectrum, manually extrapolated baseline, combined synthetic signals and the synthetic signals as manually fitted components.

signals. The synthetic curves have been shifted in the illustration for better visibility of the baseline estimations. The weighted error of the ALR estimation based on the low density is 0.54 while the weighted error corresponding estimation based on the high density measurement is 11.92.

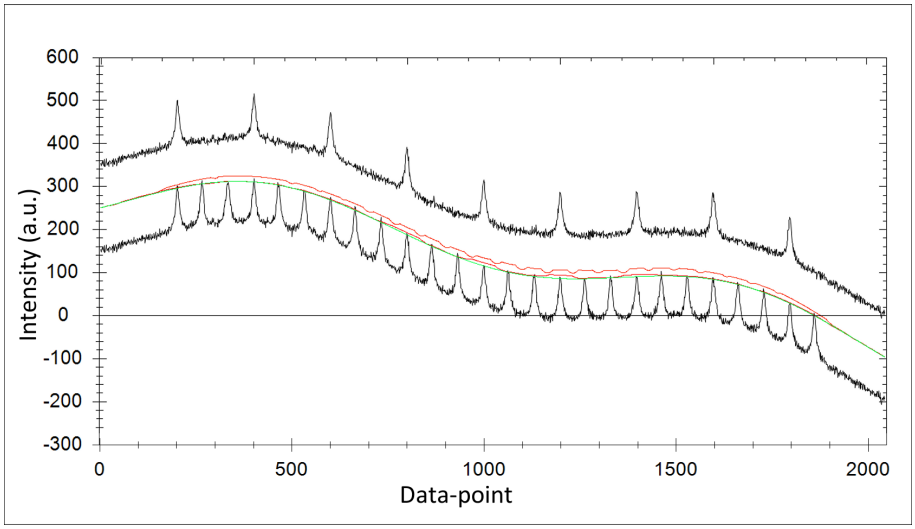

Figure 8.5: Illustration of the effect dense signal clusters have on the baseline estimation via ALR. Baseline estimations for the depicted synthetic spectra are given in red while the correct baseline is given in green. Spectra are shifted for better visual representation.

The results show that baseline estimation suffers when peak density increases; this

is to be expected as synthetic signals are modeled as Lorentzian peaks and as such possess wide ranging tails that influence large areas of the measurement. Dense clusters of peaks thus result in constellations that mask the original baseline completely and the estimation actually approximates the summation of peak-tail influences instead of the actual baseline. However peak density is not the only influencing factor. Figure 8.6 shows in example of identical signal density in terms of peaks per x-axis unit but varies the width parameter of the peaks. The effect on the baseline estimation is similar to the increase of signal density as the baseline estimation corresponding to the slim peak measurement is characterized by a weighted error of 4.90 while the wider peak signals result in an estimation that has a weighted error of 23.43. Peak width is usually given relative to the peak height and described by the full width at half maximum (FWHM) or half width at half maximum (HWHM). Changing the peak height parameter while keeping the width parameter constant changes the absolute width and therefore affects baseline estimation in a similar fashion as keeping the height constant and increasing the width parameter.

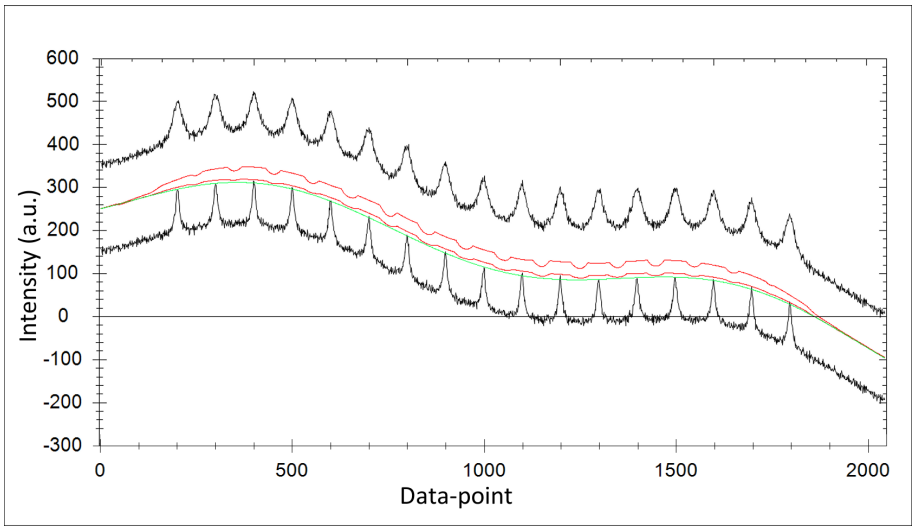

Figure 8.6: Illustration of the influence of peak width on the baseline estimation via ALR. Baseline estimations for the depicted synthetic spectra are given in red; the correct baseline is given in green. Spectra are shifted for better visual representation.

Quality or accuracy of the baseline estimation is consequently related to the signal density as well as the signal shape. Signal density and peak width both increase overall influence signal components have on the data series and thus reduce the accuracy of baseline estimation based solely on the measured data. However there is no breakdown point at which describes a set of parameters to which the estimation is accurate or beyond which it is no longer accurate. Increasing the influence of signal features gradually reduces the accuracy of the baseline estimation via ALR as less data-points in the mea-

surement represent pure baseline and noise distortions however in practice even a flawed correction can severely increase the overall quality, provided the errors introduced be baseline distortions are more intense than the inaccuracies of the correction.

8.5 Signal recognition on synthetic data

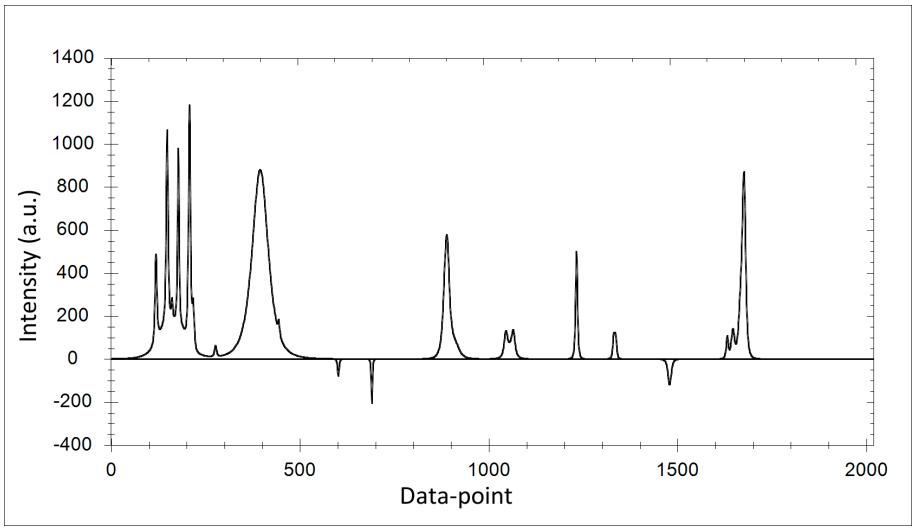

Figure 8.7: Synthetic signal containing a variety of Lorentz-peaks with different intensity and width parameters. Several peak functions are superimposed to mimic typical behavior found in real measurements.

Since synthetic measurements have no unit of measurements the x-axis unit of fully synthetic measurements is simply references by the sample point number. The total number of points was chosen as 2048 because it is a standard size for CCD based optic sensors. The synthetic spectrum used for testing baseline estimation contains several typical signal constellations that also appear in real measurements. The signal at $x = 1700$ is partially superimposed by two smaller signals at $x = 1670$ and $x = 1655$. Negative peaks at $x = 1500$, $x = 700$ and $x = 610$ represent possible notch filters, defective pixels, imperfect edge behavior that can appear within a measurement due to concatenation of two or more partial measurements, or unknown sources. Signals at $x = 1350$, $x = 1250$ and $x = 1070$ represent symmetric superimposed signals with different distances between singular peak signals. The signal at $x = 1250$ is a combination of three very close single peaks so that the result can also be interpreted as a representative of a pure single peak. An asymmetric superposition in which the component peaks no longer from separate local maxima is found at $x = 900$, this represents a typical signal found in real measurements where peaks signals are rarely pure and symmetrical. The most challenging signals in terms of baseline estimation are the wide peak at

$x = 400$ and the signal cluster reaching from $x = 100$ to $x = 230$. Wide peaks possess correspondingly wide tails that are extremely difficult to discern from baseline behavior without employing a peak model and differential considerations between a fitted model and the overall signal. As a result most baseline estimation techniques tend to remove portions of the peak base as it is challenging to determine where baseline curvature end and where a peak starts. The same is also true for the signal cluster with the additional difficulty of local minima. Assuming that only a small portion of the whole measurements is examined at a time, these minima can be mistaken as baseline points. Additional challenges are given by the positioning of the wide peak and the signal cluster. Both are relatively close to one edge of the measurement domain providing fewer possibilities to interpolate than for positions near the middle. Additionally cluster and wide peak are relatively close to each other thus providing few reliable hints to the actual baseline over a relatively wide region of the x-axis.

8.5.1 Full synthetic data

It is impossible to give an overview of all possible baselines but in this section a small portion of typical baselines with different characteristics is given. The synthetic data allows a mathematical description of the enhancement provided by the baseline correction. As the correct signal is known the correlation between the baseline corrected synthetic measurement and the signal and non-baseline corrected synthetic measurement and signal can be calculated and given as an indicator of enhancement. Generally it can be said that the higher the correlation between the baseline-corrected measurement and the original signal the better the baseline estimation.

Figure 8.8 shows an example of an asymmetric hill like baseline that is one typical form of Raman fluorescence. The baseline itself has a single local maximum but contains several changes in curvature. Baseline frequencies are overall low, but sloping generally becomes stronger towards the edges of the measurement. The result of the baseline estimation using ALR is shown in red while the correct baseline is marked in blue. As expected the largest difference between estimation and real baseline is found at the wide peak signal and the dense superposition. Baseline correction enhances the correlation between pure signal and synthetic measurement form 0.7725 to 0.9961.

Figure 8.9 shows a wave-like baseline that appears almost sine like but at closer inspections reveals asymmetric behavior and variable frequencies. Baseline behavior contains convex and concave elements that are approximated without problems. Increasing the overall frequencies used in the generation of the synthetic background signal, thus adding more "hills and valleys" can describe a natural threshold for baseline estimation techniques. As baseline distortions observed in real measurements usually only have one or two global maxima and minima these are more theoretical than practical assumptions. The correlation between signal and synthetic measurement is 0.6481

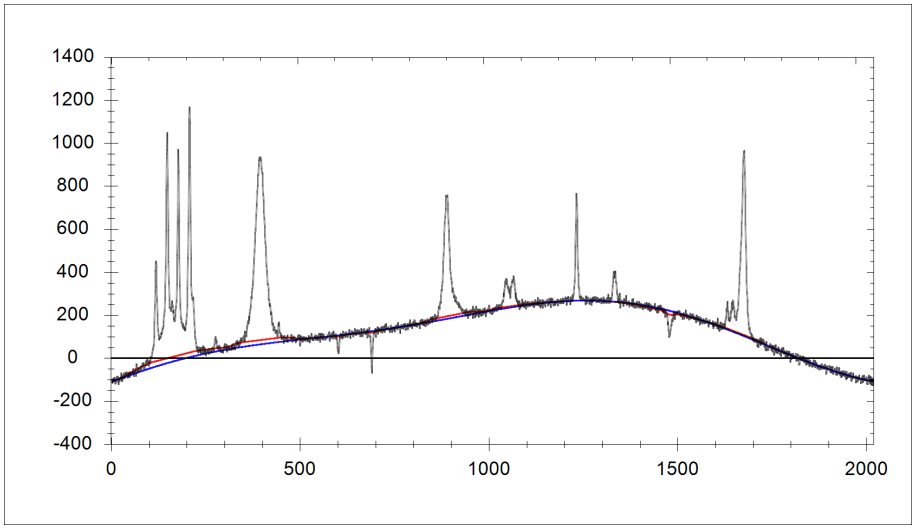

Figure 8.8: Synthetic measurement based on the signal shown in figure 8.7. Artificial baseline behavior is typical for fluorescence found in Raman spectra.

before baseline correction and 0.9955 after the correction.

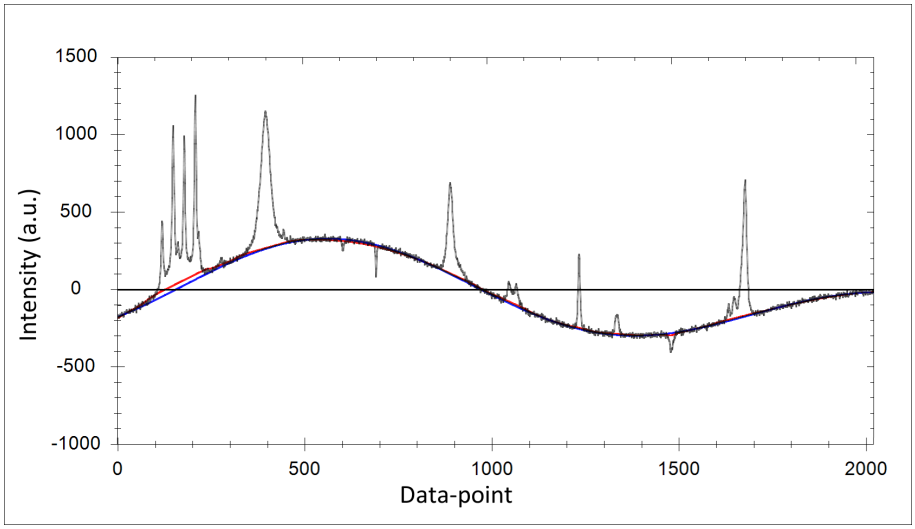

Figure 8.9: Synthetic measurement based on the signal shown in figure 8.7. Baseline behavior represents wave-like functions.

Figure 8.10 stands for a class of baselines that combine wave or sine like behavior with linear sloping behavior. Since the absolute baseline distortion in relation to the signal intensities is smaller than in the examples before the correlation pre-correction has a value of 0.8952. The correlation after the correction is 0.9967. Combinations of to different characteristics can often be found in measurements that are acquired in two or more different stages of on two separated sensors each measuring a part of the overall

domain. For the final measurements all partial measurements are combined sometimes introducing distortions at the transitions and mixing different baseline characteristics. In an ideal case each partial measurement would be corrected separately, however in some practical scenarios only the combined results is available to the correction software and the used tools have to be robust against errors introduced by the data fusion.

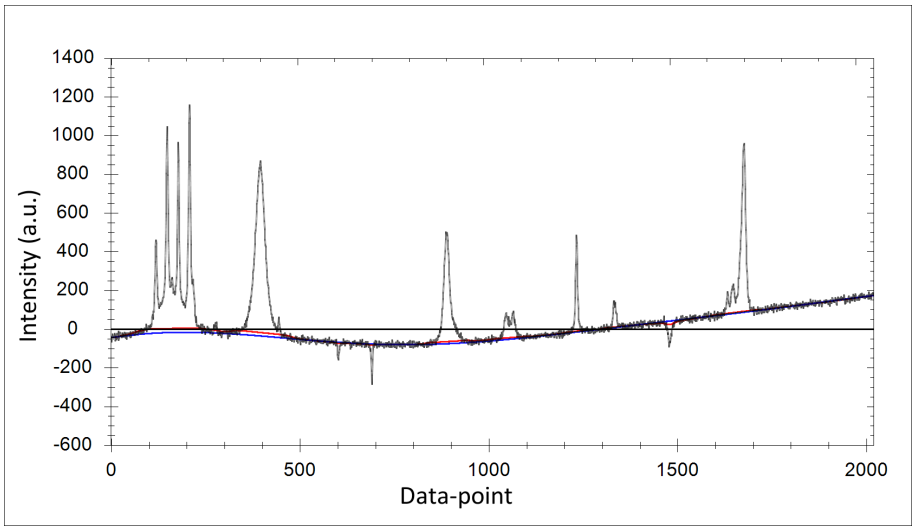

Figure 8.10: Synthetic measurement based on the signal shown in figure 8.7. The synthetic baseline represents a mixture of wave-like and linear behavior.

Figure 8.11 is an example of a valley-like baseline that consist of a single global minimum and asymmetric slopes with increasing intensity towards both edges. The baseline can be interpreted as the inverted counterpart of the baseline shown in 8.8 and shows that ALR is not dependent on a particular model but able to adapt to a wide variety of baseline shapes. The correlation of the uncorrected synthetic measurement with the pure signal is 0.5348 and is enhanced to 0.9939 after the correction.

Figure 8.12 gives an example of an almost linear increasing baseline with no inherent maximum or minimum. Examples of this type of baseline are numerous as it can be found in Raman, Libs, NMR and other real measurements. Linear or almost linear baseline distortions are usually the simplest case for any type of estimation. The baseline itself contains only extremely low frequencies and even lower frequencies typical for wide peaks are well separated from relevant baseline frequencies. Although the baseline is comparatively simple and thus easily compensated by a human observer the correlation between signal and uncorrected synthetic measurement is only 0.5348. The correlation value after baseline correction is 0.9941.

The baseline shown in figure 8.13 is closely related to the baseline shown in figure 8.9 but is overall of slightly higher frequency and while it does not represent a pure sine wave one can say that the predominant phase is shifted. The change is only

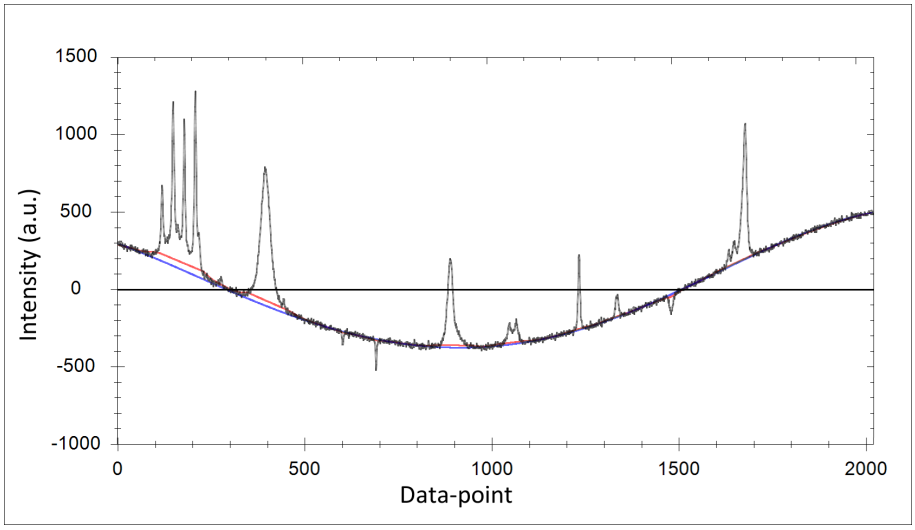

Figure 8.11: Synthetic measurement based on the signal shown in figure 8.7. The baseline represents concave functions with high intensity values at the edges of the measurement domain.

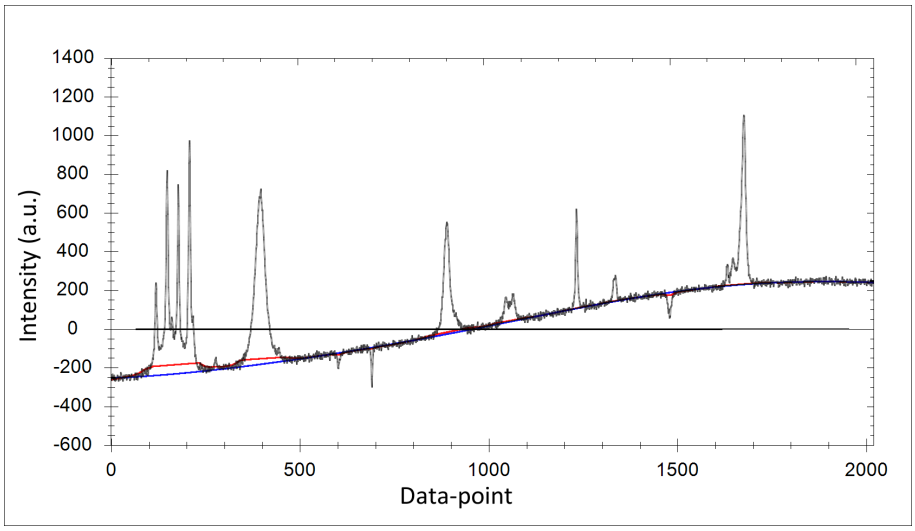

Figure 8.12: Synthetic measurement based on the signal shown in figure 8.7. Baseline behavior is monotonous, with increasing intensities from left to right.

relatively minor but becomes relevant in combination with the synthetic signal. Due to the increase in frequency and the phase shift the behavior at the left edge of the synthetic measurement is no longer linear but in fact the strongest curvature manifests itself directly underneath the signal cluster at $x = 100$ to $x = 230$. This curvature is not estimated correctly and consequently the error in the estimation increases. The correlation value before correction is 0.6636 and while the correlation after correction is 0.9892.

Figure 8.14 shows a different spectrum that was chosen to illustrate the weaknesses

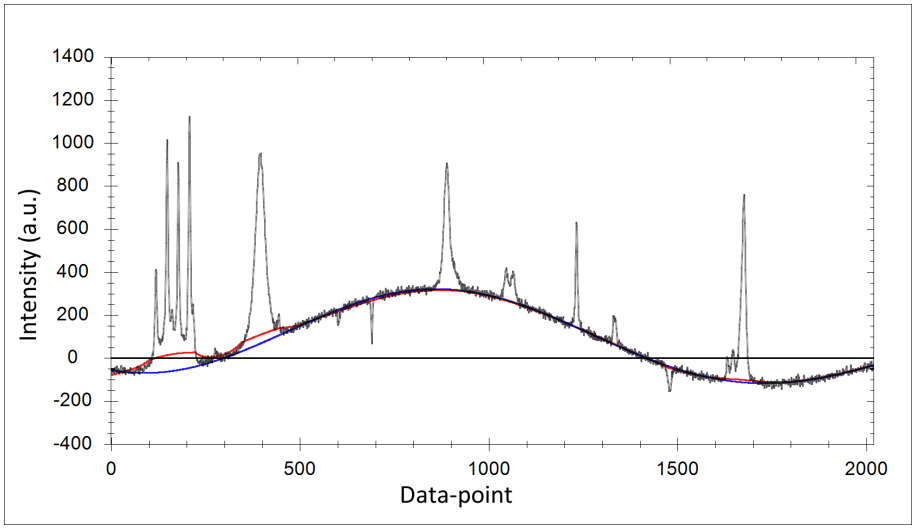

Figure 8.13: Synthetic measurement based on the signal shown in figure 8.7. Baseline behavior wave-like but curvature changes are unfavorably aligned with the signal.

of adaptive local regression. The baseline is relatively simple showing roughly the same behavior as the baseline illustrated in figure 8.9 and thus baseline behavior itself is not critical for the estimation. In this case however the synthetic spectrum was created deliberately to cause ALR to produce poor results by creating a wide ranging cluster of dense signal features that create a false baseline plateau at approximately $x = 800$ to $x = 1200$. Several local minima in the plateau are very similar in intensity which causes the stability criterion to identify interpolations through the plateau points to be classified as stable and consequently results in a baseline that follows the signal too closely. The behavior is also reflected by the correlation values of 0.9582 before the correction and 0.9646 after the correction.

An example of baseline distortions in combination with modulated noise is shown in figure 8.15. Since ALR estimates noise characteristics via shift invariant wavelet transform and incorporates the information into the iterative estimation process variable noise has no detrimental effect on the estimation as the estimation is centered in the noise regardless of intensity. High noise intensities naturally still increase the uncertainty of the estimation since error tolerances are derived from noise characteristics however this is true for all estimation algorithms. In the example correlation between pure signal and synthetic measurement is enhanced from 0.6058 to 0.9907 by ALR.

The effect of superior noise reduction are harder to judge by comparing the correlations between two signals as they are not nearly as noticeable are baseline corrective measures in realistic scenarios. Furthermore, noise estimation is already incorporated into the ALR baseline correction approach thus minimizing potentially remaining errors introduced by noise to a minimum. Figure 8.16 gives an example of a synthetic

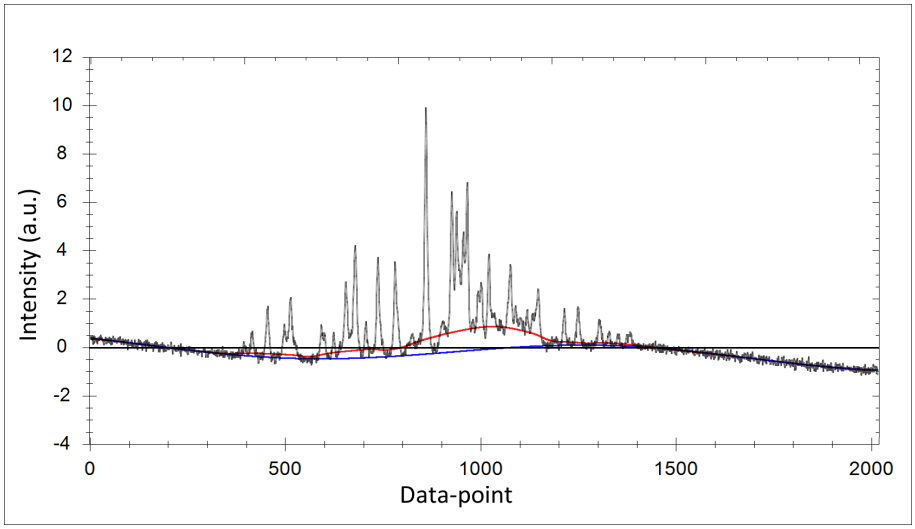

Figure 8.14: Synthetic measurement containing an extended dense signal cluster designed to induce a faulty baseline estimation by adaptive local regression.

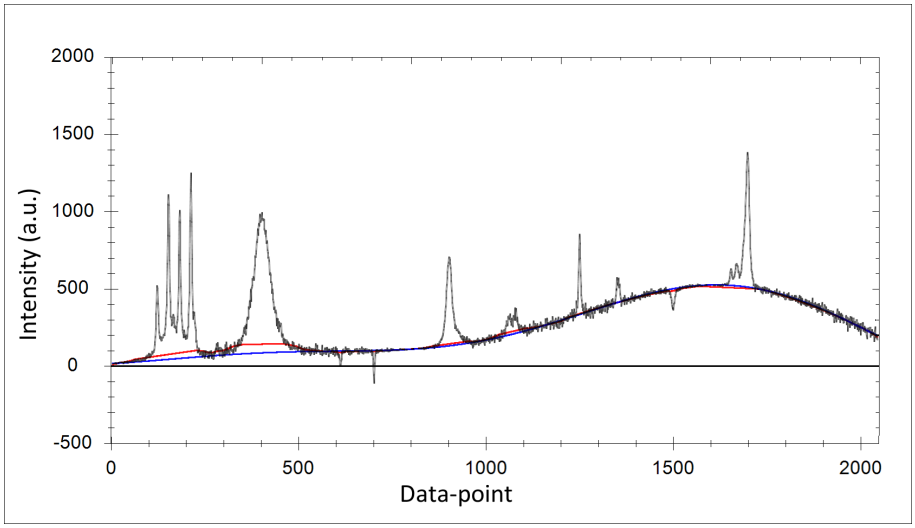

Figure 8.15: Example baseline estimation via ALR in a synthetic measurement containing baseline distortions and modulated noise.

measurement containing significantly lower signal to noise ratio that previous examples. The correlation before baseline correction is 0.7155 and 0.9263 after applying ALR. Reducing the noise influences by variable noise thresholding of the shift invariant wavelet coefficients further increases the correlation to 0.9760.

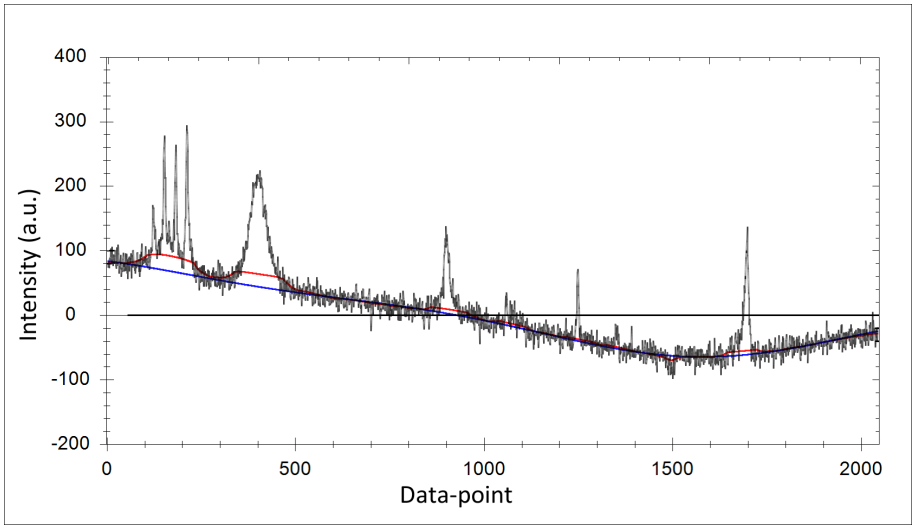

Figure 8.16: Baseline estimation in synthetic measurement exhibiting relatively low signal to noise characteristics.

8.5.2 Synthetic data based on real measurements

Synthetic data discussed up to this point had no direct connection to real measurements. Instead the synthetic data set used in figures 8.8 to 8.14 is an artificial assembly of typical singular signals. In contrast to that the synthetic data below is directly linked to real measurements as it represents synthetic rebuilds of real spectra. The rebuilds are not totally accurate but manage to capture the overall behavior of the real measurements and thus offer the possibility to qualify baseline estimation results under conditions that are very close to real measurements.

Figure 8.17 shows a manual rebuild of a real spectrum as described in section 8.3. Manual rebuilding of real spectra is an extremely time consuming task but it also offers a relatable method of testing baseline estimation techniques under real conditions. The two real rebuilds that are presented here were chosen due to their challenging characteristics. Rebuild one, seen in figure 8.17, possesses a relatively strong baseline curvature and several peak signals of significantly different widths and intensities as well as intense negative distortions. Rebuild two, presented in figure 8.20, possesses a relatively simple baseline but pushes the baseline detection to its limits due to extremely wide and intense signals while also containing smaller signals of less extreme parameters. In both cases the signal rebuilds were also tested against baseline distortions that do not mimic the ones of the original measurements. To give a simple and relatable mathematical measure of baseline estimation quality the correlations between the correlation between the pure synthetic signal and the baseline corrected noise signal is given. Since positioning does not play a role in this context only the correlation at a shift of zero is given.

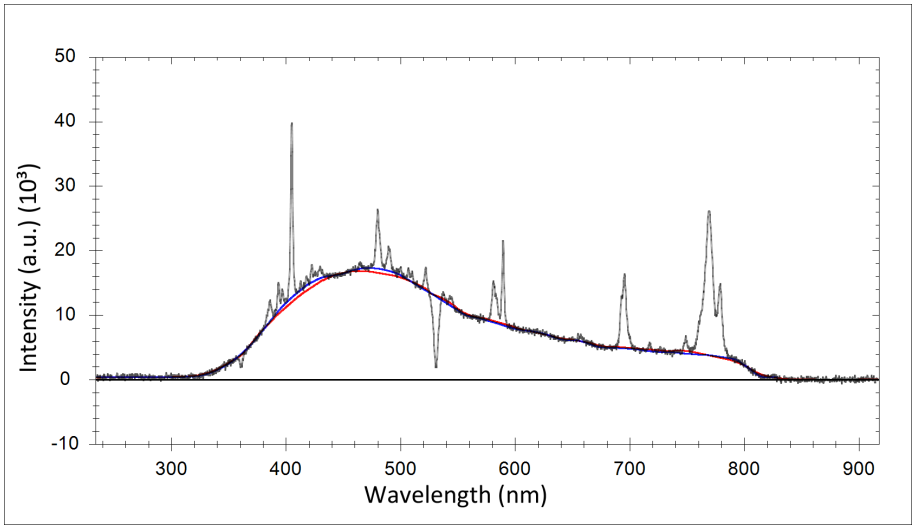

Figure 8.17: Baseline estimation via ALR in a synthetic spectrum based on real LIBS measurements. Signal and baseline features in the synthetic spectrum mimic the characteristics found in real measurements while allowing reliable comparison between correct baseline (blue) and estimated baseline (red) behavior.

Correcting the spectrum using the estimated baseline illustrated in figure 8.17 achieves a correlation of 0.9927 between corrected and pure signal compared to a correlation of 0.4951 between uncorrected measurement and pure signal. The estimated baseline is almost a perfect fit in regions that only contain few or no signals. Slight errors in the estimation are introduced in areas of relatively dense signal clusters due to the strong curvature in these areas. The peak cluster at approximately $x = 760$ to $x = 790$ introduces only a minimal error caused by the local minimum at the left flank of the large combined signal. The negative distortions at $x = 360$ and $x = 530$ have not deteriorative influence on the estimation.

Figure 8.18 shows the pure signal approximation extracted from the real spectrum shown in figure 8.17 in combination with a different synthetic baseline. The correlation results improve from 0.5713 in the uncorrected case to 0.9934 after baseline correction. Errors in the estimation are slightly smaller than in figure 8.17 due to the fact that dense signals now specially coincide with low curvature baseline behavior.

The third and final example based on the pure signal from figure 8.17 is illustrated in figure 8.19. The illustrated synthetic baseline distortions represent an irregular wave with approximately three fully oscillations in the spectral domain. The correlation of 0.7986 in the uncorrected case can be enhanced to 0.9919 by the baseline correction. The main source of the reduced performance compared to the estimation shown in figure 8.18, is again the relatively strong curvature in regions that include several dense signal features, in the example around $x = 400$ and around the negative distortions at $x = 530$. Despite some inaccuracies in the estimation the overall baseline correction

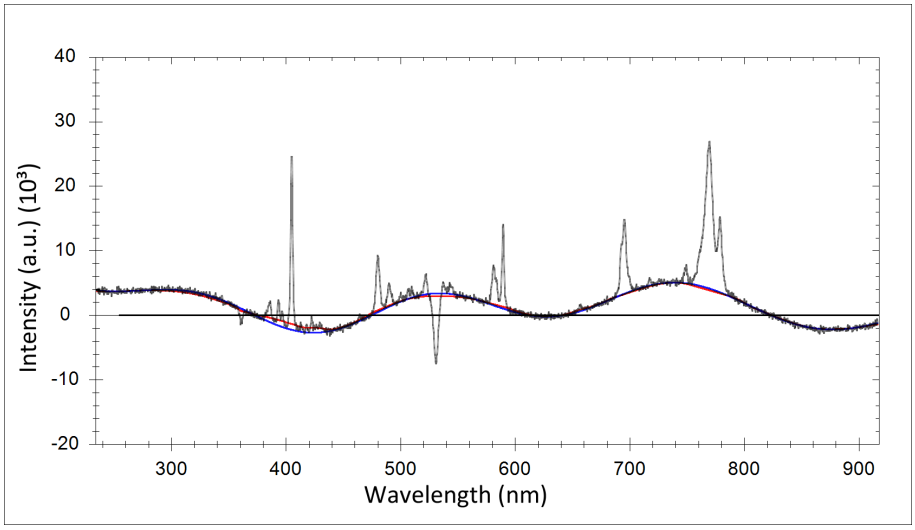

Figure 8.18: Baseline estimation via ALR in a synthetic spectrum. Signal components in the illustrated synthetic measurement are based on real data. The added synthetic baseline possesses significantly different characteristics from the baseline shown in 8.17 illustrating ALR's adaptive capabilities and independence from specific baseline models.

achieved with ALR is good as deviations from the correct baseline are minor and occur in regions where baseline behavior cannot be accurately determined from the data while the estimation is almost perfect in regions that contain only few signal features.

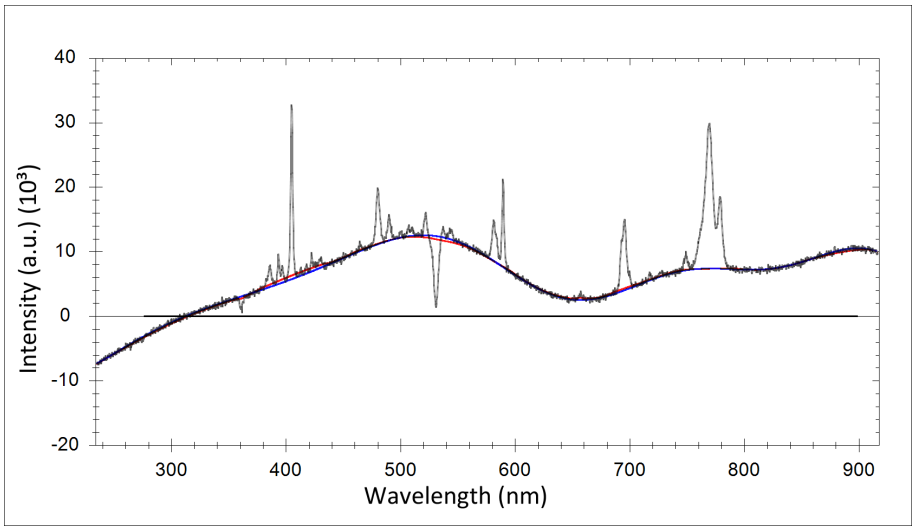

Figure 8.19: Example of baseline estimation via ALR using synthetic a measurement with signal features based on real data. Deviations between correct synthetic baseline (blue) and estimated baseline via ALR (red) are minimal.

Figure 8.20 illustrates the second synthetic measurement based on a real spectrum. The spectrum was selected as an example, because its characteristics in terms of variable

peak widths and intensities are extreme and therefore present a challenge for every general baseline correction technique. The figure illustrates the synthetic measurements with consists of the pure signal, random normal distributed noise and a low frequency baseline distortions function. The correlation of the uncorrected measurement is 0.9483 and can be enhanced to 0.9927 using ALR to estimation the baseline distortion. In contrast to the baseline estimations shown in figures 8.17 to 8.19 the estimation error shown in figure 8.20 is significant even though the baseline distortion function itself does not represent a challenge. The reason for this is simply the fact that peak tails of wide signals of signal clusters are extremely difficult to discern from low frequency baseline behavior and that an algorithm that works without additional information is not able to decide where a peak starts and where baseline distortions begin. Or in other words, the combination of baseline and signal makes the estimation challenging not the baseline or the signal alone. In the particular case discussed here the local minimum introduced by the small signal at the right slope of the main signal cluster is interpreted as a baseline points and thus results in an incorrect approximation. However while the estimation is clearly not optimal the overall enhancement of the measurement by baseline correction is obvious by studying the correlation results.

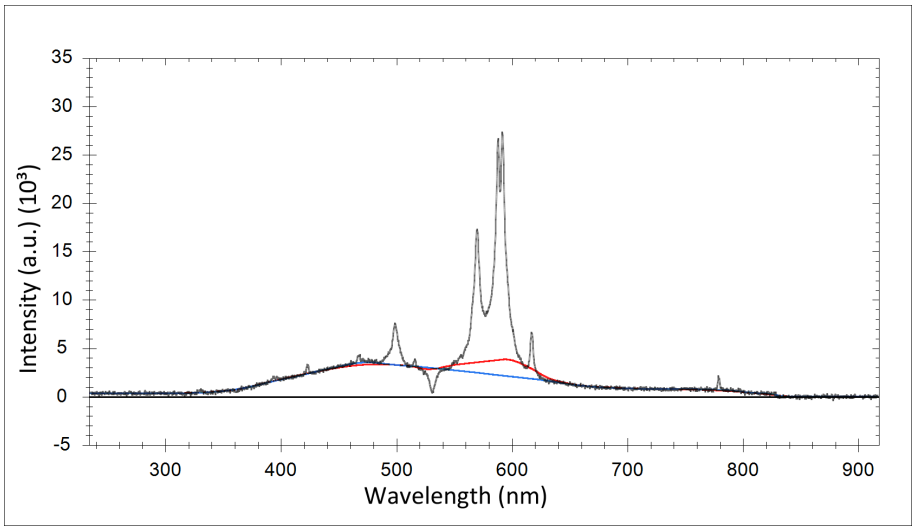

Figure 8.20: Illustration of a challenging environment for baseline estimation based on real measurements. Extremely wide signal components are difficult to discern from baseline variations leading to a clearly visible error in the estimation (red) when compared to the actual synthetic baseline (blue).

Analog to the first synthetic rebuild two other baselines are given using the same signal. Figure 8.21 shows the pure synthetic signal portions of the manual reconstruction seen in 8.20 in combination with a different baseline behavior. The baseline distortions is overall stronger than in the original reconstruction which leads to a correlation of only 0.1528 between the synthetic measurement and the pure signal. The corrected

measurement possesses a correlation of 0.9899 to the pure signal. It is interesting to see that the main error, the estimation of the baseline beneath the large signal cluster, returns almost the some characteristic than in figure 8.20. The additional errors in the estimation are minor and are introduced by relatively high frequency baseline behavior in the right portion of the synthetic measurement.

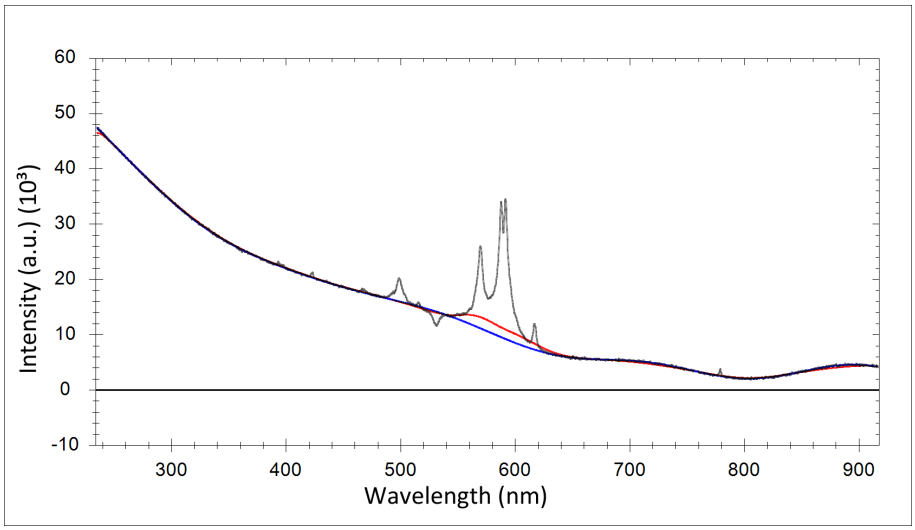

Figure 8.21: Illustration of ALR baseline correction using the same signal components as 8.20 in combination with a different baseline behavior. Estimated baseline (red), actual synthetic baseline distortion (blue).

The final synthetic baseline example shown in figure 8.22 stands representative for theoretical baselines that contain high frequencies. Baselines like this might exist however they were never encountered in real measurements available during the development of adaptive local regression nor did spectroscopy experts deem them typical. So, while more of a theoretical curiosity than a realistic simulation the baseline distortions show in figure 8.22 can be seen as an example of the limits of the proposed baseline correction technique. The relatively high frequencies contained in the synthetic baseline function keep several baseline regions from reaching stable conditions during the two dimension adaptation process. As a result several local minima and maxima that are part of the baseline are interpreted as signal components and consequently not approximated correctly. Maxima in the baseline are undershot by the estimation while minimal are overshot by the estimation. Still, the overall correlation improves from 0.7445 in the uncorrected case to 0.9705 in the baseline corrected case. Furthermore the baseline correction does not introduce new errors to the signal. Some features of the baseline are not fully removed but other than e.g. airPLS which is by design forced to stay underneath the original curve ALR can produce a baseline that reduces the distortions at every point in the domain without being able to remove them fully.

When faced with high frequency baseline distortions an estimation that must remain under the original curve can inadvertently produce undershoots that amplify distortions instead of attenuate them.

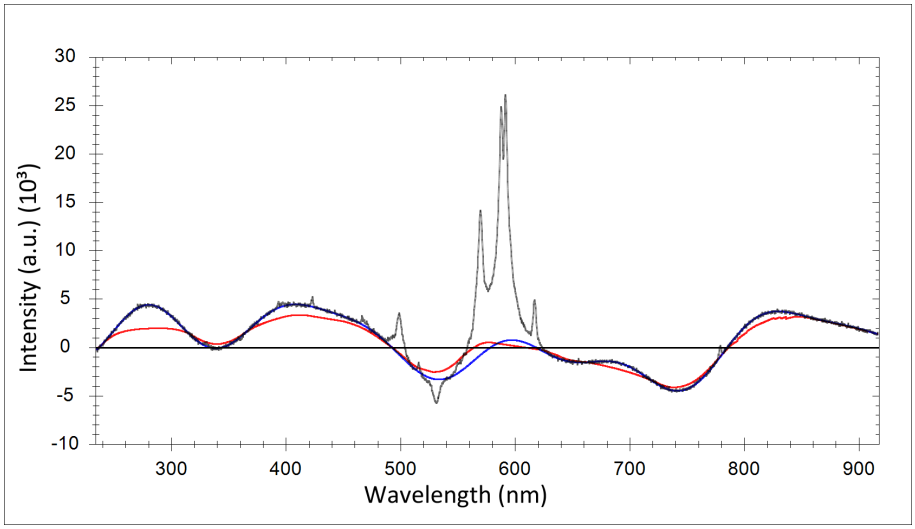

Figure 8.22: Example of ALR baseline correction displaying the limits of the method. Using the same signal components as in figure the synthetic baseline cannot be accurately estimated due to comparatively high frequencies. While unable to follow the baseline closely the estimation reveals the general trend of the distortion and is able to reduce them significantly.

8.6 Signal recognition on real data

Real measurements are by nature more challenging to analyze than synthetic ones simply because no two real measurements are one hundred percent identical which means that some measure of uncertainty always remain. As shown in section 8.3 the success of baseline estimation and correction is not only dependent on the baseline distortion itself or the signal features present in a measurement but specifically on the combination of both. A baseline that is easily estimated in one measurement can be impossible to determine in another. For this reason it is impossible to present an all-encompassing test or a complete analysis that determines the capabilities of baseline estimators. Instead several examples of real measurements are discussed below and the effects of baseline correction shown at these examples. Not all baseline distortions are particularly strong and often a whole series of measurements showing the same compound also contains very similar baseline distortions, simply thus making it difficult to motivate the need for a correction. The example measurements presented below were chosen because they represent cases in which a correction is absolutely necessary in order to achieve sensible results using automated techniques that determine similarity and show that even the correction of relatively subtle baseline distortions can improve

identification results.

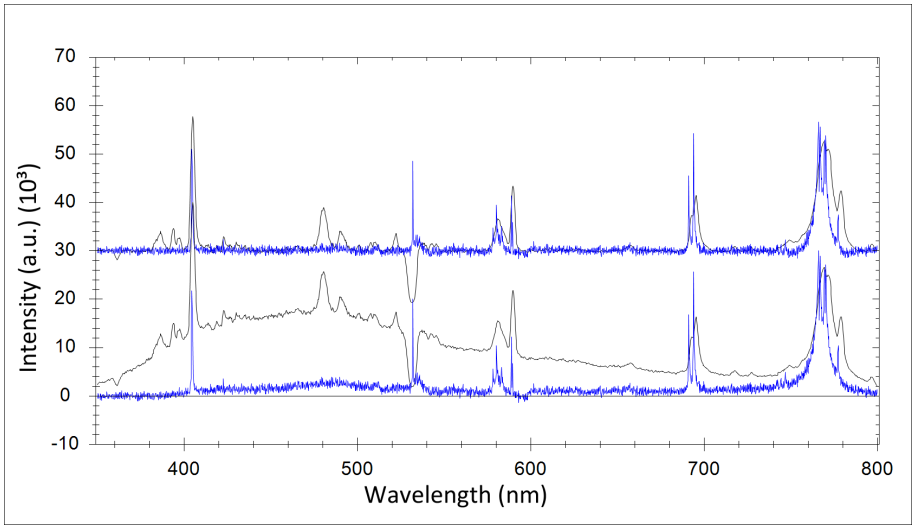

Figure 8.23: Visual comparison of real LIBS measurements showing the same explosive substance. Measurements were acquired using different instrumental setups. Shown are the uncorrected measurements (bottom) and the ALR baseline corrected measurements (top). Corrected measurements are shifted by a constant intensity for ease of inspection.

The first real example is illustrated in figure 8.23. Two versions of the same measured compound are shown, one in blue the other in black. The bottom versions in the figure are the uncorrected measurements as they were acquired while the top versions are baseline corrected using ALR. Note that the corrected data was shifted after the correction was applied for better visibility. Several peak signals appear in both measurements, most notably the single signal at $x = 400$, the signal groups at $x = 580$ and $x = 690$ and the large superposition at approximately $x = 770$. Measurement two (black) contains several additional signals that either represent impurities or a different composition of the measured substance or are hints towards higher quality of the signal or are caused by different parameters used during the measurement. The difference in visible signals is a challenge that cannot be solved by methods that are aimed to enhance signal quality. However since both measurements are examples of the same compound it is feasible to assume that the intense low frequency background visible in measurement two is in fact a form of baseline distortion and that eliminating the background dramatically enhances the automatic identification in both cases. This assumption is supported by the correlation values with is 0.4549 for the non-corrected measurements and 0.7429 for the corrected versions. As mentioned before a reason for the relatively low correlation of 0.7429 in the corrected case are the signals that appear in measurement two but are not visible in measurement one. In both cases, uncorrected and corrected, the correlation coefficient reached its maximum at a shift of $t = 7$ hinting at a small offset between the measured signals.

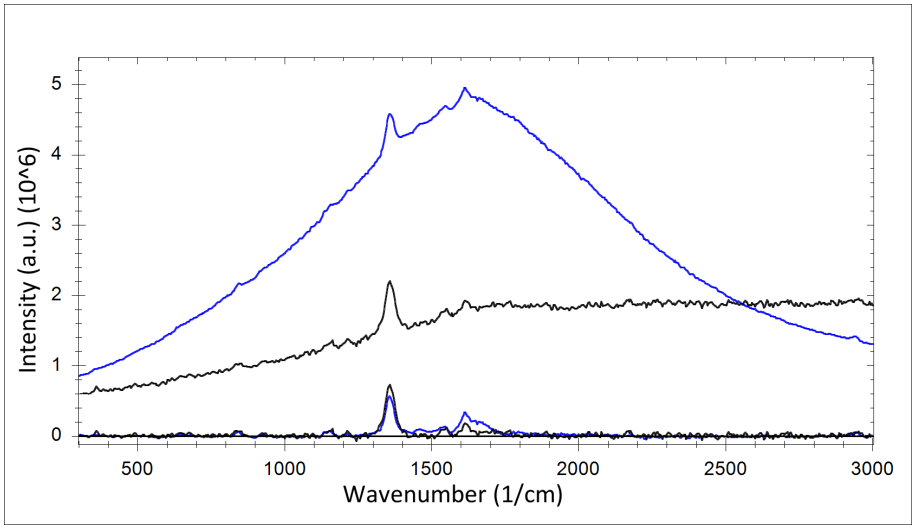

Figure 8.24: Example of a effective fluorescence suppression in Raman measurements via ALR. Shown are a reference measurement of a pure explosive substance (black) and the same substance in a mixture causing a high fluorescence background signal (blue). The corrected versions of both spectra are visible near the x-axis.

Another real example of measured substances distorted by significantly different background signals is shown in Figure 8.24. The figure shows two Raman measurements before and after baseline correction. The first spectrum shows the measured compound in its pure form, visualized in blue, while the second spectrum shows the same compounds in a mixture with another substance that adds severe fluorescence to the observed result. The correlation of the non-corrected spectra at a shift of $t = 0$ is 0.6709, however the cross correlation does form a local maximum near zero. Instead the maximum correlation is achieved by shifting the data series so that the global maxima of both signals coincide. Taking the pure measurement as fixed and shifting the spectrum of the mixture accordingly this would mean shifting to the left until the tip of the background "mountain" which is at approximately $x = 1600$ in the original coincides with the peak signal at $x = 1357$ in the measurement of the pure substance. The correlation coefficient of the baseline corrected versions is 0.8273 which also describes the global maximum of the cross correlation. Note that for visualization purposes the pure spectrum was multiplied with a factor of twenty.

Figure 8.25 illustrates another example of Raman spectroscopy measurements. Again a measurement of a pure compound and a mixture containing the compound which adds severe baseline distortions are given before and after the correction. The pure measurement spectrum contains relatively low intensity baseline distortions while the spectrum of the mixture is distorted by a baseline that increases rapidly in the left portion of the spectrum and slowly in the right portion before it drops off near the right edge of the spectral domain. The correlation between the uncorrected measurements is 0.3905

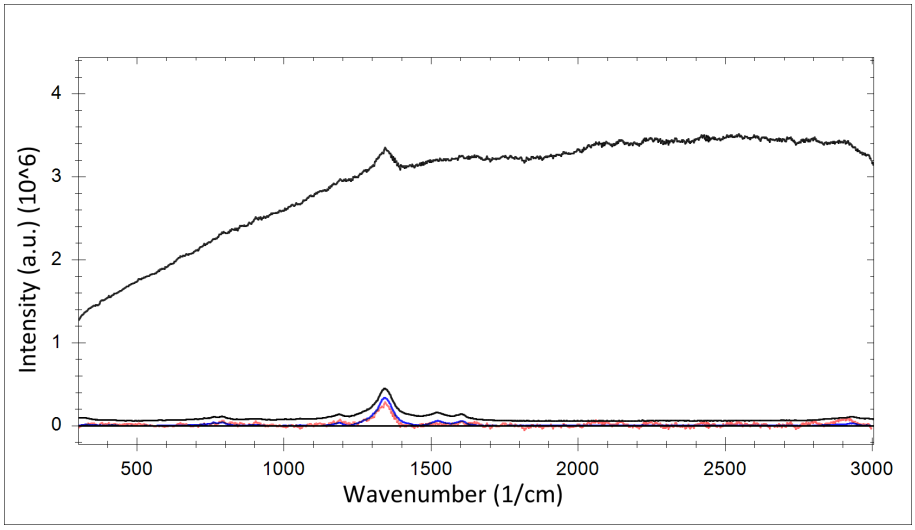

Figure 8.25: Real Raman measurements of an explosive substance. Non corrected measurements are given in black, pure laboratory measurement and standoff mixture measurement. Red: ALR corrected mixture measurement. Blue: ALR corrected pure laboratory measurement.

while that between the baseline corrected versions significantly enhanced to 0.8273.

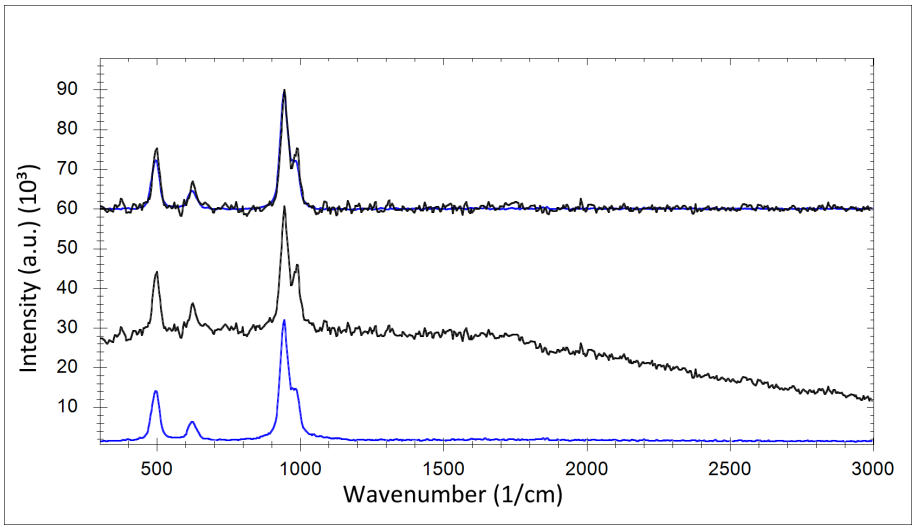

Figure 8.26: ALR baseline correction illustrated using real Raman measurements. Uncorrected Reference and standoff measurement are given in the lower half of the figure while the corrected spectra have been shifted upwards for better visibility. Note that in this example the reference measurements acquired under laboratory conditions contains almost not baseline distortions.

Figure 8.26 gives another example of typical baseline distortions encountered in Raman spectra. The example spectra were chosen due to the fact that the baseline distortions are relatively strong but other than in previous spectra does the mixture

does not introduce any visible signals that interfere with the correlation and reduce the achieved coefficient after the correction has taken place. The corrected spectra are again shifted for better visual representation. The correlation between the uncorrected pure spectrum and the uncorrected mixture spectrum of 0.6677 is enhanced to a coefficient of 0.9705 between the baseline corrected versions. Because the pure spectrum shows almost not baseline distortions it is also possible to calculate the correlation between the uncorrected pure spectrum and the baseline corrected mixture spectrum yielding a very good correlation coefficient of 0.9666.

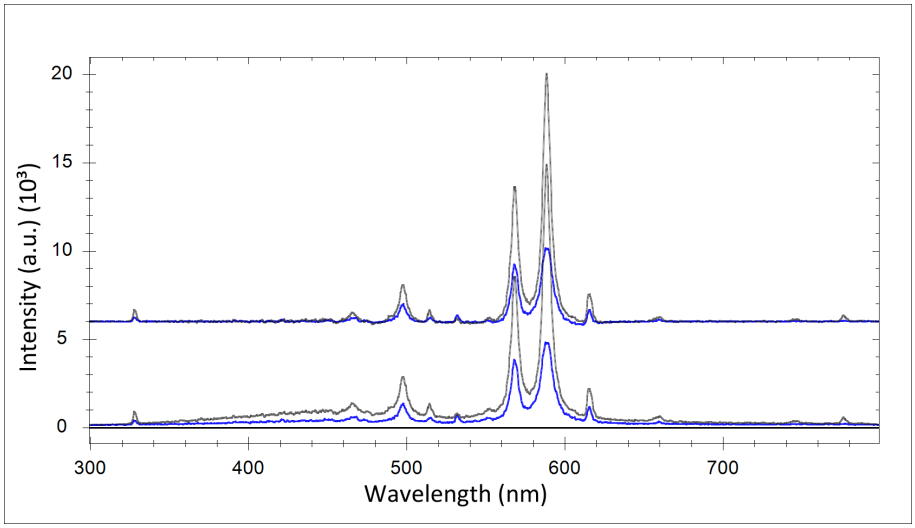

Figure 8.27: Example of potential enhancements to real LIBS spectra contain only minimal differences via ALR baseline correction. Given are uncorrected measurements of the same explosive substance belonging to the same measurement series and the corrected spectra, again shifted upwards for better visibility.

The last example of real spectra takes a closer look on spectra that are already similar to show the benefit of baseline correction in scenarios where it is not strictly necessary. The LIBS spectra shown in figure 8.27 are taken from the same measurement series and are consequently very similar. The examined compound is not only of the same general composition but is in fact identical. Additionally as the complete measurements series took place inside a lab in a timeframe of only a few minutes, the external and internal influences on the measurements are as close to identical as possible in real live scenarios. Consequently the correlation between both uncorrected signals is already very high yielding a correlation coefficient of 0.9794. Using ALR to correct for the slightly different baseline distortions in both examples manages to enhance the coefficient to 0.9814. The absolute enhancement achieved by the correction is comparatively small, however considering that maximum possible enhancement is 0.0206 and that baseline correction via ALR manages about 10% of that theoretical maximum, the results are impressive considering that a perfect match is practically unobtainable.

8.7 Discussion

The achieved results demonstrate that baseline correction is a valuable tool for a wide array of spectroscopic techniques that enhances the performance of automatic detection and identification techniques. Adaptive local regression is a baseline correction technique developed with the goal to offer a high amount of robustness in order to work well in a fully automated data processing routine and data obtained under uncontrolled conditions. The effectiveness of ALR has been shown using fully synthetic data sets, semi synthetic measurements that are based on real data and real measurements. A key feature of ALR compared to other modern baseline correction methods is its robustness against bidirectional errors in the measured data. While the common assumption of a baseline that lies underneath the data is usually correct ALR shows that good correction results are possible without a hard constraint that causes the correction to fail if even a single data-point contains a negative outlier. Furthermore due to its local characteristics and using a variable error threshold function ALR is able to return a baseline estimate that is centered in noisy regions of the spectrum instead of lying beneath the measured data amplifying the effect of changing noise intensities. A downside of the local characteristics of ALR is the behavior extended dense clusters of signal features in which the baseline estimate is usually closer to the original data than necessary. This characteristic is due to the fact that it is not possible to extract information about baseline behavior from dense signal clusters. Other baseline estimation techniques compensate for this lack of information by introducing global parameters to control the estimated baseline curve at the however these global parameters can also introduce errors and need to be carefully controlled and adjusted by technology experts. ALR possesses no global control parameter that determines baseline behavior but extracts the necessary information from changes of the estimation itself by determining constant and changing areas in estimations under changing conditions to extrapolate the final baseline. Since baseline characteristics are generally not clearly defined the soft assumption about baseline behavior have to be translated into firmer constraints in order to derive a baseline estimate. In the case of ALR the general assumptions about baseline behavior are contained in the bandwidths used for the local regressions that are the basis for the final baseline estimation. It is unavoidable for a baseline estimation technique to possess some sort of parameter that is that basis for determining the difference between signal features and baseline curvature simply because signals in various techniques can possess significantly different characteristics and no single parameter could work of all techniques. Bandwidths between one percent and twenty percent of the observed domain have proven sensible and reliable in all analyzed real spectra. However not all spectroscopic techniques could be examined and it is likely that other techniques will require other parameters simply because signals

possess vastly different characteristics or the ratio between signal size and measurement domain is not comparable with LIBS, Raman, Infra-red or NMR measurements. Considering the wide array of tested baseline characteristics the fact that ALR can be used for more than one technique using the same parameter set demonstrates its wide applicability and robustness. The achieved results are often very close to the optimal solution and even in cases where the baseline estimation did not return close to optimal results the estimated baseline still managed to significantly reduce distortions and corrected spectra were considerably closer to the non-distorted versions without the introduction of new errors.

Chapter 9

Optix hardware and software

9.1 Hardware

The Optix prototype allows the standoff detection of explosive materials outside of laboratory conditions. Figure 9.1 shows the test firing of the laser operating at 532 nm (green light) during heavy snowfall at the FOI facilities near Grindsjön lake in Sweden.

Figure 9.1: Optix standoff detection at the FOI facilities in Sweden.

Figure 9.2 gives a schematic view of the prototype build in the course of the Optix project. The laser block includes two different lasers types, one yttrium-aluminium-garnet (YAG) laser operating at optional wavelengths of 532 nm or 1064 nm and one quantum-cascade laser (QCL) operating at 5-6 μm. The articulated optical arm allows high precision aiming without realigning the entire prototype to the target. The weight of the YAG laser unit is approximately 45 kg the cooling unit weights about 22 kg and keeps the laser within the operating temperature of 10-30°C.

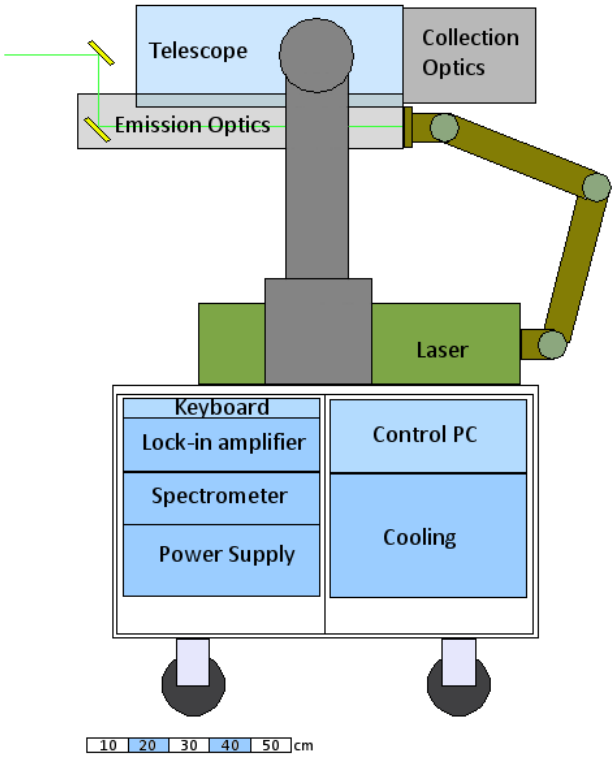

Figure 9.2: Schematic illustration of the Optix prototype and its components

The spectrometer unit includes a 3-channel LIBS spectrometer and a water cooled 1-channel Raman ICCD detector with a newly developed optical bench. The ranges of the LIBS spectrometer channels are given as follows: channel 1: 350-591 nm , channel 2: 600-807 nm and channel 3: 800-974 nm. The range of the Raman spectrometer is 534-620 nm which translates to wavenumbers (200-3300 cm^{-1}).

Figure 9.3 shows a more detailed view of the optical layout including laser unit, optical arm, beam expander, telescope as well as mirrors and filters.

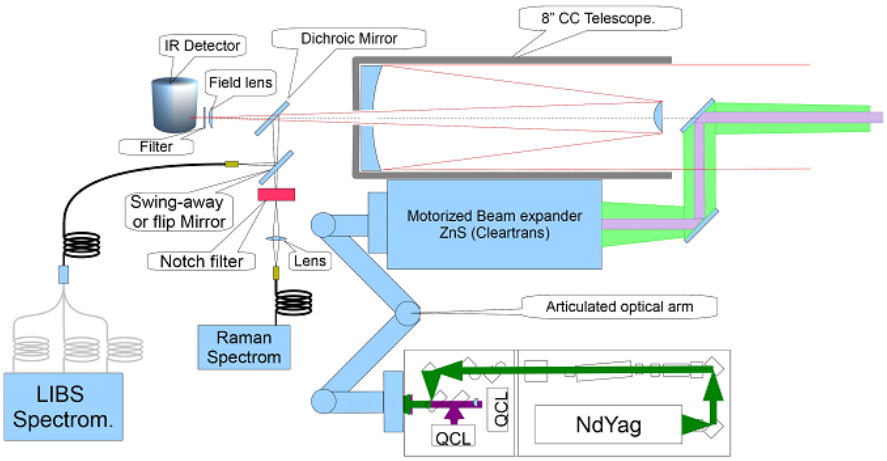

Figure 9.3: Optical layout of the Optix prototype

9.2 Software

The adaptive noise reduction and baseline correction methods proposed here are part of a full signal processing library consisting of various filters like mean, Gaussian and Fourier filters, analysis methods like regression and principle component analysis and an intuitive graphical user interface. Figure 9.5 gives an overview of the most important classes of the chemometrics software developed for the Optix prototype as well as the graphical user interface implemented to use the signal processing library and additional functionality independent of the prototype hardware. The software is written in C# and based on the .NET framework.

The key pre-processing components enabling automatic analysis of standoff real world measurements have been embedded in a stand-alone (dynamic link library) dll providing a wide range of chemometrics tools for the Optix prototype. Since one goal of the prototype is a high degree of atomization functionalities have been encapsulated into pipelines that use different sets of parameters fitted to accommodate each of the three technologies and its specific hardware. To allow easy access to all techniques in a stand-alone application, which is not tied to the Optix prototype, a full graphical interface and chemometrics tools library was developed. The SpecViewer application uses the Optix chemometrics library as an integral part and extends its functionality with additional filters, hardware and software interfaces as well as functionalities aimed to process additional types of spectral information.

Central parts of the chemometrics library are the *IO* and *Datatypes* classes. The *IO* class contains all methods necessary to read and write data references and results in several different formats while the *Datatypes* class serves as container for all data types that are used to interchange data between subclasses of the software project. The

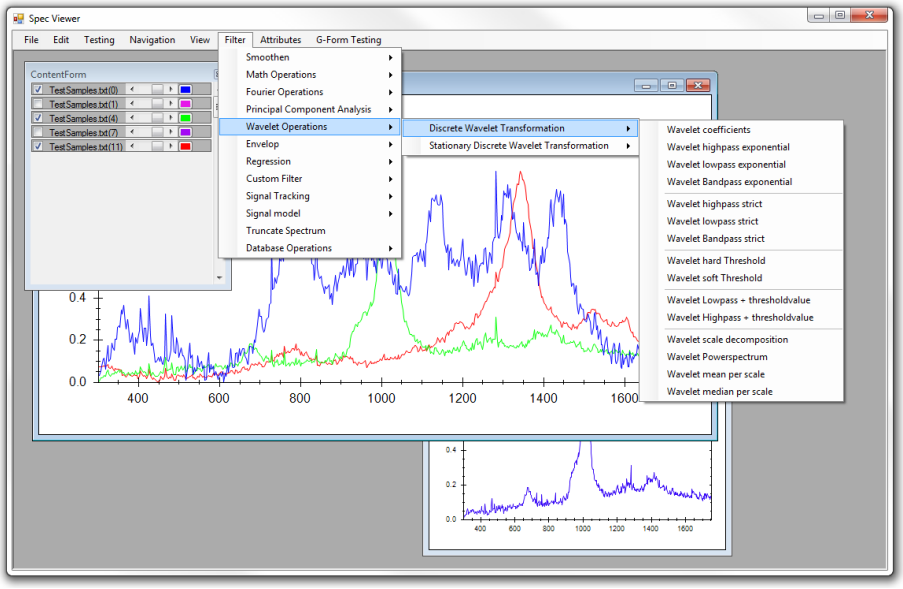

Figure 9.4: Screenshot showing the SpecViewer graphical user interface

Datatypes class also contains the spectrum data structure which serves as storage architecture for all spectrometric data, offers rudimentary search and control mechanisms and optional fields to store additional data about measuring instruments, domains and content identifiers. The technology module describes the abstract basis class for specialized processing steps aimed specifically at the three technologies and applications that are used in the Optix prototype. It uses the *R.Net* library, [Cod13], which offers access to the free statistical programming language *R* and powerful decision making tools like linear discriminant analysis, support vector machines and random tree forests. Each of the three Optix technologies is handled in a specialized technology module implementing the abstract technology class which parameters of pre-processing and identification methods for data acquired with the specific spectrometric technique. The *Filter* class serves as primary interface encapsulating internal processing methods for the technology modules or the graphical user interface. The *SignalOps* class contains a collection of simple operations that can be carried out on data represented in arrays of position and intensity and are not related to spectroscopic data processing like basic math operations, shifting, truncating, re-sampling etc. Analogous to the *Filter* class which is part of the chemometrics library the *Filter2* class serves as container and interface class with the graphical user interface for processing methods that are not part of the library environment. The *SignalGen* class serves as a collection for signal generating methods encompassing synthetic signals, different types of noise and artificial baselines. For other classes names generally reveal their main function. The user interface itself is realized as a multi-document-interface (MDI) that offers a frame application in which

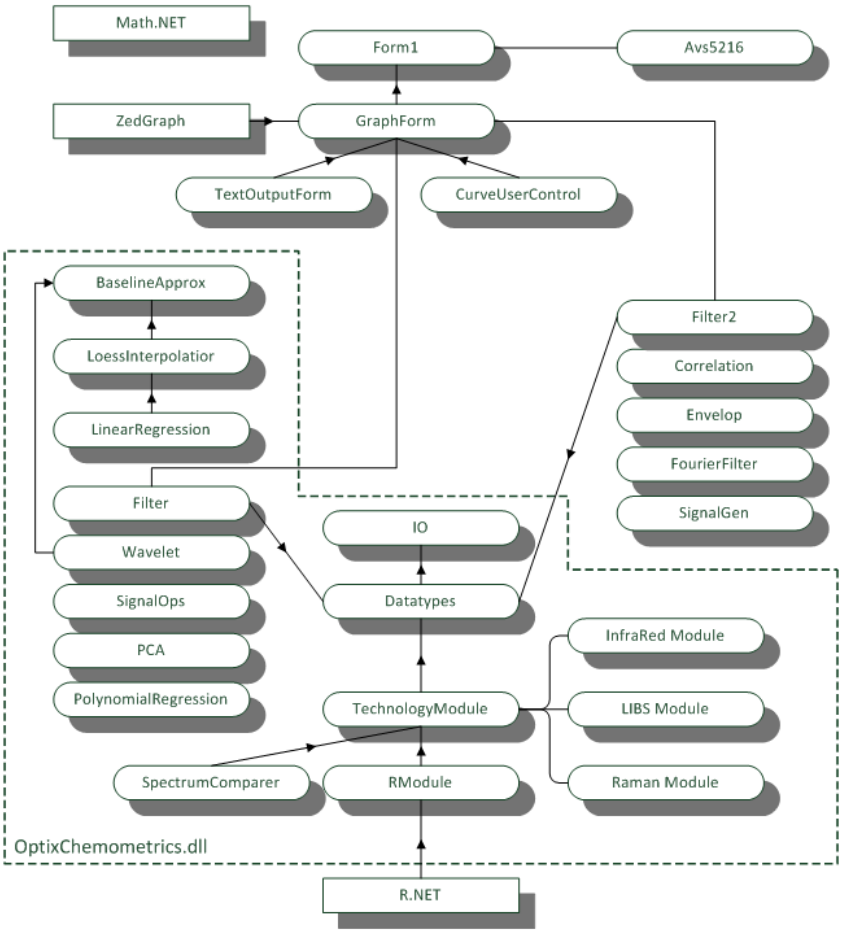

Figure 9.5: Schematic illustration of the specViewer software including the most important signal processing classes contained in the Optix Chemometrics library

multiple spectra can be viewed and manipulated at once. *GraphForm* which uses and extends the functionality supplied by the free *zedgraph* library to visualize data sets provides the functionality to display single or multiple sets of data in form of a sampled curve. The multi-document-interface allows the display of multiple graph forms at once and the software also supports the transfer of data from one graph to another. The *Avs5216* class implements a graphical interface for direct access to Avantes spectrometers, which are used in the Optix prototype, so the spectrometers can be used directly with the SpecViewer software. *Math.Net* is a free library containing a wide variety of mathematical applications that is used by almost all classes in the project.

9.3 Results

Detailed results of the Optix project contain security sensitive information about detection capabilities and thresholds of explosive materials and thus cannot be made available to the general public. However to convey a sense of typical results table 9.1 shows a test setup including 106 Raman measurements 91 of which contain different explosive substances in varying mixtures and concentrations as well as 15 samples that do not contain any explosives. True positive identification relate to cases in which the explosive was correctly identified while true negative cases relate to correct identification as non-explosive. False positive are false alarms where a non-explosive material was mistaken as an explosive and false negatives mark cases where the identification failed to find an explosive substance. The overall results are good given the challenging scenario of optical standoff detection of explosive materials.

Total samples	106	
Explosive mixtures	91	
Non explosives	15	
True positive	78	
True negative	14	
False positive	1	
False negative	13	
tp-rate	0.857	(recall)
tn-rate	0.933	(specificity)
fp-rate	0.067	(fallout)
fn-rate	0.143	(miss rate)
Correct-rate	0.868	
Error-rate	0.132	
F-measure	0.918	
Pos.pred.Value	0.987	(precision)
Neg.pred.Value	0.519	

Table 9.1: Exemplary identification results obtained using the Optix prototype in a standoff detection scenario.

The precision of the identification is high, 98.7% positive predictions are very satisfactory for a prototype instrument. On the other hand the negative prediction value of only about 50% appears problematic. The reason for this is that several measurements in which an explosive mixture was examined are not identified as explosive by the automatic classification. A closer look at the corresponding measurements reveals that the error does not lie with the pre-processing or identification algorithms, but in the acquired spectra that simply do not contain any relevant information. Figures 9.6 and 9.7 show samples based on the measurement series examined in table 9.1 that are classified as explosive but not identified as such by the automatic pre-processing analysis methods. The reference measurement of the contained pure explosive material is

given to show the discrepancy between expected and acquired signal. In mixtures the acquired spectrum usually does not match the reference precisely due to distortions and signals introduced by other components of the compound. However the most relevant signals of the reference must but visible in order for a positive identification.

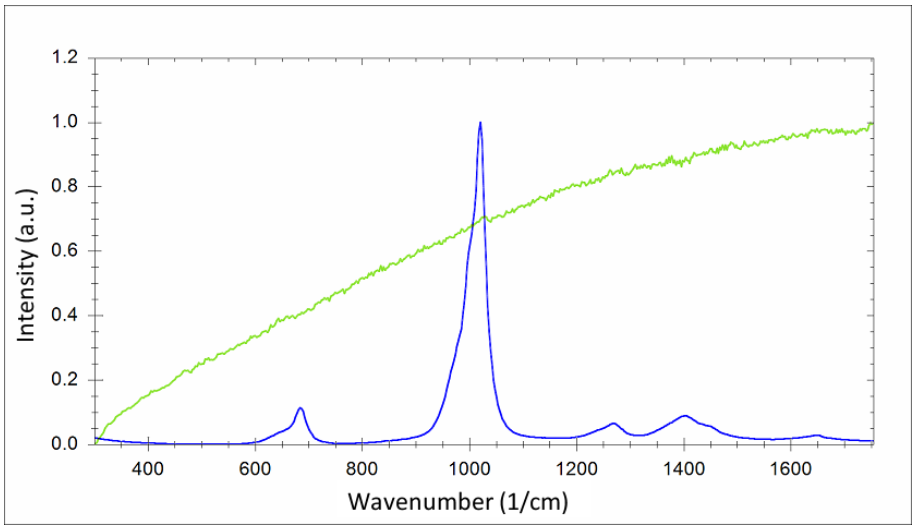

Figure 9.6: Illustration of false negative identification. Acquired standoff Raman measurement shown in green, corresponding pure explosive reference spectrum shown in blue.

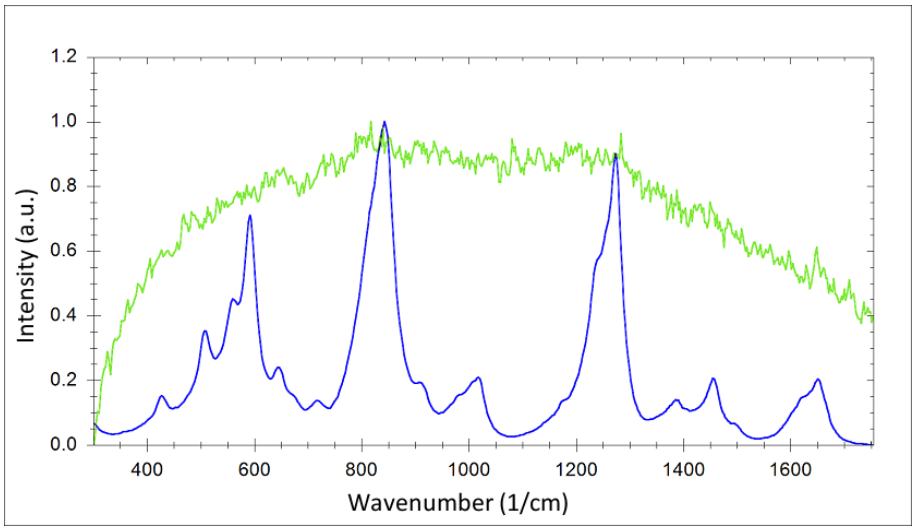

Figure 9.7: Illustration of false negative identification. Acquired standoff Raman measurement shown in green, corresponding pure explosive reference spectrum shown in blue.

The most probable reason for spectra not containing information is that the explosive traces in the examined mixture were not properly hit. In real scenarios this can be avoided by multiple sampling at slightly different positions of a suspicious object.

Chapter 10

Outlook

Tests under real world conditions - meaning standoff distance between measurement instruments and examined substances as all as exposure to whether conditions outside of a laboratory or roofed structure - have shown that advanced preprocessing methods are necessary to enable automatic detection based on reference data.

Usage of a threshold function to estimate variable noise parameters solves problems like over- and under-performing suppression in low and high intensity noise regions respectively. However the extraction or estimation of a variable function is challenging and more prone to errors than the estimation of a single value from a set of distorted sample points. Both proposed methods limit the number of sample points available for the estimation in order to describe the variable nature of the noise modulating function. In doing so they naturally reduce their capability to correct for common noise variations and possible artifacts. The estimated threshold values thus represent less of a general noise characteristic and more the specific noise present in the examined measurement. This does not diminish the suppression characteristics, but while a global threshold value varies to a minor degree in different instances generated using identical noise parameters, a threshold function is more likely to reflect the specific characteristics of individual instances. To enhance estimation results future works could focus on typical characteristics of variable noise environments and incorporate those into the estimation process. An other approach that avoids bias introduced by modeling functions, which might be too rigid in real world scenarios, could be the incorporation of several different estimations methods in order to combine all results to a final estimation that minimizes individual errors.

By incorporating an adaptive error function and stability of estimations based on varying input parameters adaptive local regression is able to achieve good estimation results in a wide array of scenarios without relying on model functions or hard constrains. However it does not possess a global smoothness criterion and therefore is more prone to errors if signals are particularly dense. Detecting and eliminating inaccurate

estimations caused by such dense signal clusters without the need for a delicate control parameter used by other baseline estimation techniques could enhance the results in such cases. Furthermore in scenarios where runtime is extremely critical the calculation of several local regressions can become an issue. Optimizations in range and sampling distance in the bandwidth domain have already been given in chapter 7 but further studies could focus on identifying stable conditions early on and adjust calculations to be carried out only for unstable regions thus reducing the average number of operations needed.

The adaptive methods proposed here are able to enhance spectral data acquired at standoff distances significantly and reduce distortions to a degree that enables automatic identification using references recorded under laboratory conditions, which was impossible without corrective pre-processing. However reference generation is still a manual task and the selection of signal features is often critical. Highly detailed references are well suited to discern similar compounds but under real world conditions measurements often do not contain details visible in the laboratory causing details to actually make identification more difficult. Low detailed references on the other hand often lack the necessary information to identify unknown compounds fully and can only be used to categorize results into groups. These challenges are likely not solvable by post acquisition signal enhancement alone, but a possible solution could involve different detail levels of reference measurements similar to the scales used in the wavelet transform process. Furthermore new sensors and advanced focusing hardware can enhance the reliability of standoff data acquisition and reduce the limit of detection characteristics for trace detection.

List of abbreviations

a.u.	arbitrary unit
AAS	Atomic absorption spectroscopy
AIMA	Automatic iterative moving averaging
airPLS	Adaptive iteratively reweighed penalized least squares
ALR	Adaptive local regression
CCD	Charge-coupled
COV	Covariance
dll	Dynamic link llibrary
DWT	Discrete wavelet transform
EDS	Energy dispersive X-ray spectroscopy
FID	Finite inductive decay
FT	Fourier transform
GC	Gas chromatography
ICP	Inductively coupled plasma
IED	Improvised explosive device
IR	Infra red
LIBS	Laser induced breakdown spectroscopy
MAD	Median absolute deviation
MRT	Magnetic resonance spectroscopy
MSE	Mean squared error
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
STFT	Short time Fourier transform
SVD	Singular value decomposition
SWT	Shift invariant wavelet transform
VAR	Variance
WDS	Wavelength dispersive X-ray spectroscopy
XPS	X-ray photoelectron spectroscopy
XRF	X-ray Fluorescence

Bibliography

[Ack15] T.G. Ackland. On osculatory interpolation, where the given values of the function are at unequal intervals. *Journal of the Institute of Actuaries*, pages 369–375, 1915.

[Aki70] H. Akima. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM (JACM)*, 17(4):589–602, 1970.

[ANW67] J.H. Ahlberg, E.N. Nilson, and J.L. Walsh. *The theory of splines and their applications*, volume 1. Academic Press, New York, NY, USA, 1967.

[BDE+04] H.F.M. Boelens, R.J. Dijkstra, P.H.C. Eilers, F. Fitzpatrick, and J.A. Westerhuis. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *Journal of Chromatography A*, 1057(1):21 – 30, 2004.

[Bey92] G. Beylkin. On the representation of operators in bases of compactly supported wavelets. *SIAM Journal on Numerical Analysis*, 6(6):1716–1740, dec 1992.

[BGG98] C. S. Burrus, R.A. Gopinath, and H. Guo. *Introduction to wavelets and wavelet transforms: a primer*. Prentice Hall, Upper Saddle River, NJ, USA, 1998.

[Boo78] C. De Boor. *A practical guide to splines*. Springer, New York, NY, USA, 1978.

[BVMW00] C.D. Brown, L. Vega-Montoto, and P.D. Wentzell. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Applied Spectroscopy*, 54(7):1055–1068, jul 2000.

[CDW75] N.B. Colthup, L.H. Daly, and S.E. Wiberley. *Introduction to infrared and Raman spectroscopy*. Academic Press, New York, NY, USA, 1975.

[CHH+04] L. Cao, P.B. Harrington, C.S. Harden, V.M. McHugh, and M.A. Thomas. Nonlinear wavelet compression of ion mobility spectra from ion mobility spectrometers mounted in an unmanned aerial vehicle. *Analytical chemistry*, 76(4):1069–1077, 2004.

[Cle79] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[Cod13] CodePlex. R.net - documentation. https://rdotnet.codeplex.com/documentation, jan 2013.

[Dau88] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

[Dau92] I. Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1st edition, jun 1992.

[Dau93] I. Daubechies. Orthonormal bases of compactly supported wavelets ii: variations on a theme. *SIAM Journal on Mathematical Analysis*, 24:499–519, mar 1993.

[DHZ86] R.P.W. Duin, H. Haringa, and R. Zeelen. Fast percentile filtering. *Pattern Recognition Letters*, 4(4):269–272, 1986.

[DJ94] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[DJ95] D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

[DRN91] W. Dietrich, C.H. Rüdel, and M. Neumann. Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *Journal of Magnetic Resonance*, 91(1):1 – 11, 1991.

[EB05] P.H.C. Eilers and H.F.M. Boelens. Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, oct 2005.

[Efr91] B. Efron. Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1(1):93–125, 1991.

[FG96] J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, USA, 1996.

[Fri88] H. Friebolin. *Ein-und zweidimensionale NMR-Spektroskopie*, volume 71. VCH, Weinheim, Germany, 1988.

[Gab46] D. Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26):429 –441, 1946.

[GB97] H.Y. Gao and A.G. Bruce. Waveshrink with firm shrinkage. *Statistica Sinica*, 7(4):855–874, 1997.

[GEN+03] I.B. Gornushkin, P.E. Eagan, A.B. Novikov, B.W. Smith, and J.D. Winefordner. Automatic correction of continuum background in laser-induced breakdown and Raman spectrometry. *Applied Spectroscopy*, 57(2):197–207, 2003.

[Gib98] J.W. Gibbs. Fourier's series. *Nature*, 59:200, 1898.

[Gib99] J.W. Gibbs. Fourier's series. *Nature*, 59:606, 1899.

[GW92] R.C. Gonzalez and R.E. Woods. *Digital image processing*. Addison-Wesley, Boston, MA, USA, 2nd edition, 1992.

[Hil87] B.F. Hildebrand. *Introduction to numerical analysis*. Dover Publications, New York, NY, USA, 2nd edition, 1987.

[HMT83] David C Hoaglin, Frederick Mosteller, and John Wilder Tukey. *Understanding robust and exploratory data analysis*, volume 3. Wiley, New York, NY, USA, 1983.

[HP97] J.H. Hodrick and E.C. Prescott. Post-war u.s. business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, 29(1):1–16, 1997.

[IND13] INDRA. Optix project website. http://www.fp7-optix.eu, jan 2013.

[Jac91] J.E. Jackson. *A user's guide to principal components*. Wiley, New York, NY, USA, 1991.

[JPL12a] NASA Jet Propulsion Laboratory. Mars science laboratory: Instruments. http://mars.jpl.nasa.gov/msl/mission/instruments/, jul 2012.

[JPL12b] NASA Jet Propulsion Laboratory. Msl science corner: Chemistry and camera. http://msl-scicorner.jpl.nasa.gov/Instruments/ChemCam/, jul 2012.

[JSY+04] A. Jirasek, G. Schulze, M.M.L. Yu, M.W. Blades, and R.F.B. Turner. Accuracy and precision of manual baseline determination. *Applied Spectroscopy*, 58(12):1488–1499, 2004.

[Kal60] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[KKMN98] D.G. Kleinbaum, L.L. Kupper, K.E. Muller, and A. Nizam. *Applied regression analysis and multivariable methods*. Duxbury Press, Pacific Grove, CA, USA, 1998.

[LGO+95] M. Lang, H. Guo, J.E. Odegard, C.S. Burrus, and R.O. Wells. Nonlinear processing of a shift-invariant discrete wavelet transform (DWT) for noise reduction. In H.H. Szu, editor, *Proceedings of SPIE*, volume 2491, pages 640–651, 1995.

[LGO+96] M. Lang, H. Guo, J.E. Odegard, C.S. Burrus, and R.O. Wells. Noise reduction using an undecimated discrete wavelet transform. *Signal Processing Letters, IEEE*, 3(1):10 –12, jan 1996.

[LMJ03] C.A. Lieber and A. Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied Spectroscopy*, 57(11):1363–1367, 2003.

[Lon77] D.A. Long. *Raman spectroscopy*. McGraw-Hill, New York, NY, USA, 1977.

[Mad78] H.H. Madden. Comments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data. *Analytical Chemistry*, 50(9):1383–1386, 1978.

[Mal02] E.R. Malinowski. *Factor analysis in chemistry*. Wiley, New York, NY, USA, 3rd edition, 2002.

[MIT13] Massachusetts Institute of Technology MIT. MIT spectroscopy lab - history. http://web.mit.edu/spectroscopy/history/history-classical.html, jan 2013.

[MM97] D.A. McNulty and H.J.H. MacFie. The effect of different baseline estimators on the limit of quantification in chromatography. *Journal of Chemometrics*, 11(1):1–11, 1997.

[NP87] W.K. Newey and J.L. Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 55:819–847, 1987.

[OV10] A.V. Oppenheim and G.C. Verghese. 6.011 Introduction to Communication, Control, and Signal Processing. (Massachusetts Institute of Technology: MIT OpenCouseWare) http://ocw.mit.edu. License: Creative Commons BY-NC-SA, 2010.

[PH96] A.J. Phillips and P.A. Hamilton. Improved detection limits in Fourier transform spectroscopy from a maximum entropy approach to baseline estimation. *Analytical Chemistry*, 68(22):4020–4025, 1996.

[PW11] B.D. Prakash and Y.C. Wei. A fully automated iterative moving averaging (AIMA) technique for baseline correction. *Analyst*, 136:3130–3135, 2011.

[RC93] P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

[RCSW97] D.A. Rusak, B.C. Castle, B.W. Smith, and J.D. Winefordner. Fundamentals and applications of laser-induced breakdown spectroscopy. *Critical Reviews in Analytical Chemistry*, 27(4):257–290, 1997.

[RJFD01] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, and J.A. Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179 – 193, 2001.

[SB83] M.P. Seah and D. Briggs. *Practical surface analysis: By Auger and X-ray photoelectron spectroscopy*. Wiley, New York, NY, USA, 1983.

[SB13] J. Schlenke and L. Brecker. Automated signal detection as tool to evaluate magnetic field homogeneity from Fourier transformed proton NMR spectra. *Applied Magnetic Resonance*, 44(6):745–758, 2013.

[SBC97] F. Sánchez-Bajo and F.L. Cumbrera. The use of the pseudo-Voigt function in the variance method of X-ray line-broadening analysis. *Journal of applied crystallography*, 30(4):427–430, 1997.

[SG64] A. Savitzky and M.J.E Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[SH13] J. Schlenke and L. Hildebrand. Automatic variable noise suppression for laser based classification of explosive materials. In M. Fathi, editor, *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*, pages 333–339. Springer, 2013.

[She92] M.J. Shensa. The discrete wavelet transform: wedding the a trous and Mallat algorithms. *IEEE Transactions on Signal Processing*, 40(10):2464 –2482, oct 1992.

[SHML12] J. Schlenke, L. Hildebrand, J. Moros, and J.J. Laserna. Adaptive approach for variable noise suppression on laser-induced breakdown spectroscopy responses using stationary wavelet transform. *Analytica Chimica Acta*, 754:8–19, 2012.

[Sim96] J.S. Simonoff. *Smoothing methods in statistics*. Springer, New York, NY, USA, 1996.

[SS89] M.C. Seiler and F.A. Seiler. Numerical recipes in C: the art of scientific computing. *Risk Analysis*, 9(3):415–416, 1989.

[STD72] J. Steinier, Y. Termonia, and J. Deltour. Smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, 44(11):1906–1909, 1972.

[SW80] D.A. Skoog and D.M. West. *Principles of instrumental analysis*, volume 158. Saunders College, Philadelphia, PA, USA, 1980.

[Tuk77] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science. Addison-Wesley, Boston, MA, USA, 1977.

[Wal55] A. Walsh. The application of atomic absorption spectra to chemical analysis. *Spectrochimica Acta*, 7:108–117, 1955.

[Whi22] E.T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922.

[Wie64] N. Wiener. *The interpolation, extrapolation and smoothing of stationary time series*. MIT Press, London, UK, 1964.

[XR08] Y. Xi and D.M. Rocke. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, 9(1):324, 2008.

[XSH11] Z. Xu, X. Sun, and P.B. Harrington. Baseline correction method using an orthogonal basis for gas chromatography/mass spectrometry data. *Analytical Chemistry*, 83(19):7464–7471, 2011.

[ZCL10] Z.M. Zhang, S. Chen, and Y.Z. Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.