Feature Extraction in NMR Data Analysis

Dissertation

zur Erlangung des Grades eines

D o k t o r s d e r N a t u r w i s s e n s c h a f t e n

der Technischen Universität Dortmund
am Fachbereich Informatik

von
Hyung-Won Koh

Dortmund 2010

Tag der mündlichen Prüfung:
02.03.2010

Dekan:
Prof. Dr. Peter Buchholz

Gutachter:
Prof. Dr. Sven Rahmann
Prof. Dr. Eyke Hüllermeier

Dortmund, den 21.Mai 2010

For Yun-Kyong

Acknowledgments

I would like to thank my supervisor Dr. Lars Hildebrand for giving me the opportunity to work as a research assistant and PhD student at the Computer Science Department, Chair 1 at TU Dortmund University of Technology. Nothing less than his constant support, his never-ending patience, and especially the exceptionally high degree of freedom he offered me during my time as a PhD-student were in my opinion excellent conditions for unbiased exploration and research. I also thank him for all the discussions and proof-readings of our papers.

I also thank Dr. Roland Hergenröder at the Leibniz-Institute for Analytical Sciences - ISAS for all of the interesting and sometimes eye-opening discussions we had in the past years. I really appreciated the highly motivating atmosphere and already begin to miss the working environment at the institute. Furthermore, I would also like to thank Dr. Jörg Lambert for many hours of very informative and interesting discussions, and of course for patiently answering all my questions regarding the principles and characteristics of Nuclear Magnetic Resonance Spectroscopy. In addition, I also thank him together with Dr. Brendan Holland and Rafael Slodzinski for english-language proof-reading of my thesis.

Special thanks goes to Dr. Frank-Michael Schleif for the great support and the encouraging discussions about spectral data analysis and more in Jyväskylä, Finland, and also in Leipzig, Germany. I hope one day I will get the opportunity to return the favor.

I am surely also thankful for all the interesting on- and off-topic discussions and moments I shared with my colleagues and friends: Chan & family (Danke für alles!), Sloty & Doro (Danke für alles!!), Parinas, Klaus, Achim, Steffen, Gatti (die Matrix in JAVA programmieren, ne?), Wolfgang (die gute Seele des LS1, schön Dich zu kennen!), Volker W. (1979-2009, r.i.p.), Markus B. & Karsten (in Gedenken an all die unterhaltsamen 5-Minuten-Gespräche vor der OH16), Volker M., Jan (ja, bis auf die Parameter ist es dieselbe Funktion...), Seppel & Uta (keine Sorge, Eurem Teller gehts gut!), Nico (Du weißt...), Tomek, Katkat, Sassi, Andrei, Martin, Albert & Laura, and everyone else I accidentally forgot to mention. You made my time in Dortmund as much comfortable as possible, thank you all for this!

Finally, I like to thank my beloved family for supporting me during my PhD studies.

Funding

I acknowledge funding by the two projects “ICA - Integrated Cell Analysis”, co-funded by the European Union (EFRE) and supported by the Ministry of Innovation, Science, Research, and Technology of North Rhine-Westphalia, and “Second Generation Locator for Urban Search and Rescue Operations”(SGL USaR) within the Seventh Framework Programe (FP7) of the European Commission.

Contents

Abbreviations and Notation

Abbreviations in alphabetical order

CBPS	Curvature-Based Peak Selection.
CT	Computer Tomography.
DNA	deoxyribonucleic acid.
FFT	Fast Fourier Transform.
LM	Levenberg-Marquardt method.
LPR	Lorentzian Peak Reconstruction.
LSR	Lorentzian Spectrum Reconstruction.
MASE	Mean Area Symmetry Error.
MPE-Area	Mean Percentage Error of the Area Parameters.
MPE-HWHH	Mean Percentage Error of the HWHH Parameters.
MPE-Pos	Mean Percentage Error of the Position Parameters.
MPSE	Mean Position Symmetry Error.
MRI	Magnetic Resonance Imaging.
MSE-PH	Mean Squared Error at the Peak Hills.
MSE-PM	Mean Squared Error at the Peak Maxima.
MSEPP	Mean Squared Error at the Peak Points.
MSSE	Mean Shape Symmetry Error.
NMR	Nuclear Magnetic Resonance.
PA	Proportional Approximation.
PET	Positron Emission Tomography.
SDR	Signal-to-Distortion Ratio.
WPSE	Weighted Pair Symmetry Error.

Notation by topic

Chapter 2

A_j	scale parameter of a spectral component j.
B_0	static magnetic field for alignment of nuclei in NMR spectroscopy, measured in tesla.
B_1	high intensity pulses of radio frequency energy used to excite nuclei.
B_{1i}	intensity of the magnetic field B_1.
I	spin quantum number of a nucleus.
N_α, N_β	number of nuclei aligning either parallel or antiparallel to an external magnetic field.
P	angular momentum of a nucleus.
P_z	angular momentum of a nucleus in direction of an external magnetic field.
$S(\omega)$	analytical description of the NMR signal in the frequency-domain.
T_1, T_2	relaxation times, the time needed for realignment of the nuclei in accordance with the static magnetic field B_0.
Θ	pulse angle, the angle between M_0 and the direction of the static magnetic field B_0 after applying B_1.
γ	gyromagnetic ratio describing the ratio of the magnetic moment to the angular momentum of a nucleus, measured in units of $\frac{\text{Hz}}{\text{tesla}}$ with $1\ \text{Hz} = 1\ \text{s}^{-1}$.
λ_j	half width at half height (HWHH) of a spectral component j.
μ	magnetic moment of a nucleus, measured in units of $\frac{\text{J}}{\text{tesla}}$.
μs	microseconds, $1 \cdot 10^{-3}$ s (seconds).
μ_z	magnetic moment of a nucleus in direction of an external magnetic field.
ω	variable denoting the frequency in a continuous manner.

ω_1	radiation frequency, applied to excite nuclei in NMR experiments.
ω_j	variable denoting the response frequency of a spectral component j.
τ_P	pulse length of the magnetic field B_1.
φ	phase of a spectral component j.
h	PLANCK's constant, $h = 6.6256 \cdot 10^{-34}$ Js (joule second).
k_B	BOLTZMANN constant, $k_B = 1.3805 \cdot 10^{-23} \frac{\mathrm{J}}{\mathrm{K}}$.
x', y', z	rotating coordinate system around the z-axis, the direction of the static magnetic field B_0.
e	EULER's number, $e = 2.718281828459\ldots$.
J	unit of energy, named for JAMES PRESCOTT JOULE. $1J = 1\frac{\mathrm{kg\,m^2}}{\mathrm{s^2}}$, with kg = kilogram, m = meter and s = second.
K	Kelvin, measure of the absolute temperature.
MHz	megaherz, 10^6 Hz = 10^6 s^{-1}.
s	second, a measure of time.
s(t)	time-domain signal of an NMR experiment, also known as FID.

Chapter 3

Q	number of local maxima in a spectrum.
$d(Y_i, Y_j)$	distance between the maximum positions ω_i, ω_j of two Lorentz-functions Y_i, Y_j.
l	index of a local maximum.
n	number of spectral datapoints.
$\mathbf{w}$, w_i	discrete frequency vector of a spectrum with w_i denoting the frequency with index i.

Chapter 4

K_1	number of outer loops in algorithm 2.
K_2	number of inner loops in algorithm 2.
Sim_A, Sim_B, Sim_C	simulated spectra for the evaluation of algorithm 2.

Sim_{real}	real-world spectrum for the evaluation of algorithm 2.
ρ	Signal-to-Distortion Ratio (SDR).
$\widehat{Y}^{[i]}$	model of the spectrum at iteration step i.
$\widehat{Y}_j^{[i]}$	model of Lorentz-function Y_j at iteration step i.
r	height threshold in algorithm 2.
u_ω, u_λ, u_A	uniformly distributed random numbers used for the Lorentzian parameters in the simulated spectrum series.
$v, v_{\max}$	random variable representing simulated noise, uniformly distributed in the range $[0, v_{\max}]$.
$w_{j,left}, w_{j,max}, w_{j,right}$	positions of the point triplet with index j.
$w_{j,x}$	arbitrary position of the point triplet with index j.

Chapter 5

R	user-specified signal free region of a spectrum.
δ	threshold parameter of algorithm 4.
$\widehat{\omega}_j^{[i]}$, $\widehat{\lambda}_j^{[i]}$, $\widehat{A}_j^{[i]}$	model parameters of Lorentz function Y_j at iteration step i.
$\widehat{\omega}_j$, $\widehat{\lambda}_j$ and $\widehat{A}_j$	model parameters of Lorentz function Y_j.
w_l, w_m, w_r	peak triplet positions.

Chapter 6

M	peak palindrome as a set of Lorentz functions.
Y_j^*	best matching peak.
$\alpha_\omega, \alpha_\lambda, \alpha_A$	symmetry weights.
$\delta_\omega, \delta_\lambda, \delta_A$	symmetry thresholds.
δ_ω^*	spectral width threshold of a palindrome.
$d_A(Y_i, Y_j)$	area symmetry distance of a pair of Lorentz functions Y_i, Y_j.
$d_\lambda(Y_i, Y_j)$	shape symmetry distance of a pair of Lorentz functions Y_i, Y_j.

$d_\omega(Y_i, c, Y_j)$	position symmetry distance of a pair of Lorentz functions Y_i, Y_j given position c.

Summary

This thesis proposes an automated feature extraction methodology for data analysis in one-dimensional Nuclear Magnetic Resonance (NMR) spectroscopy. The aim is to reduce the amount of data while simultaneously preserving the information.

Based on the theory of NMR, the signal of an NMR experiment can ideally be described as a superposition of Lorentz functions with unknown parameters. The task of feature extraction is accomplished by decomposing a given spectrum into a set of distinct Lorentz functions in the main part of the thesis. The two major problems arising in this context are *peak selection* and *parameter approximation*. The former addresses the problem of identifying the set of Lorentz functions contained in a spectrum, and the latter stands for finding the corresponding set of parameters that best fit the data.

The main contributions of this thesis have been published as follows:

1. H.W. Koh, S. Maddula, J. Lambert, R. Hergenröder and L. Hildebrand, "Feature Selection by Lorentzian Peak Reconstruction for ^{1}NMR Post-Processing", Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS'08), Yväeskylä, Finland, pp. 608-613 (Koh *et al.*, 2008),

2. H.W. Koh, S. Maddula, J. Lambert, R. Hergenröder and L. Hildebrand, "An Approach to Automated Frequency-Domain Feature Extraction in Nuclear Magnetic Resonance Spectroscopy", Journal of Magnetic Resonance 201(2), pp. 146-156 (Koh *et al.*, 2009),

3. H.W. Koh and L. Hildebrand: "A Heuristic Approach for the Identification of Palindromic Peak Sets", Proceedings of the 2009 International Conference on Bioinformatics and Computational Biology (BIOCOMP 2009), Las Vegas NV, USA, pp. 632-638 (Koh & Hildebrand, 2009).

In the context of *peak selection*, the trivial approach of focusing only on the occurrence of local maxima has the drawback of inherently omitting "shoulders", hidden Lorentz functions which are overlapped by their neighbours such that their corresponding local maximum is not present in the spectrum. With providing a theoretical basement, this thesis proposes two alternative methods to solve this problem, one by repeated subtraction (Koh *et al.*, 2008), and the other based on the curvature information of a spectrum, i.e. by identifying each Lorentz function as a second derivative minimum (Koh *et al.*, 2009). Moreover, by exploiting the analytical solution for the parameters of a single Lorentz function, an alternative *parameter approximation* scheme is proposed (Koh *et al.*, 2009). It is empirically shown that the results highly outperform the Levenberg-Marquardt algorithm, a commonly used least-squares method in the context of NMR data analysis.

Finally, aiming at encapsulating the signal of the same origin throughout a series of NMR experiments, an approach for the identification of palindromic peak sets is proposed (Koh & Hildebrand, 2009).

All empirical studies have been carried out on a dual core 1.66 GHz notebook with 1 GByte RAM and OS Windows XP.

The thesis is structured as follows: After the introduction in Chapter 1 and the basic concepts of NMR spectroscopy in Chapter 2, general aspects regarding the task of feature extraction in NMR data processing are given in Chapter 3, including an overview of related literature. An initial, local maximum-based feature extraction approach is proposed in Chapter 4, and Chapter 5 deals with a corresponding approach incorporating the information provided by the second derivative. Subsequently, an algorithm for the identification of palindromic peak sets is proposed in Chapter 6, and finally the conclusions and comments on future work are given in Chapter 7.

Chapter 1

Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy allows one to investigate the electronic environment of atoms and the interaction between neighbouring nuclei, and yields information about the number and type of molecular substructures contained in a sample. In general, the application of NMR spectroscopy ranges from physics to various branches of chemistry, biology and medicine (Ernst *et al.*, 1987). In particular, NMR spectroscopy plays a major role in molecular structure determination and the analysis of heterogeneous mixtures of molecules. Figure 1.1 shows an example of a modern NMR device.

1.1 Motivation

Understanding the mechanisms and principles of life has always been of central interest with the purpose of not only fighting diseases and healing injuries, but generally finding answers to the fundamental question of how life actually establishes in arbitrary living organisms like plants, bacteria, insects and vertebrates including the human species. The early work on inheritance between 1856 and 1863 by Gregor Johann Mendel, leading to the Mendel's Laws of Inheritance, the discovery of the double helix structure of the deoxyribonucleic acid (DNA) molecule in 1953 by James D. Watson and Francis Crick, and the completion of the Human Genome Project in 2003 are some of the most well known milestones.

Next to a scientist's brilliant mind, diligence, excitement and sustained endurance, progress in biological research also essentially depends on the technologi-

Figure 1.1: Worlds first 1 GHz Spectrometer. The picture is taken from http://en.wikipedia.org/wiki/File:Bruker_Avance1000.jpg

cal development (we cannot prove what we cannot measure). For instance, the work of James D. Watson and Francis Crick was based on observations made by x-ray imaging, while Mendel carried out his studies only equipped with a pencil, a paper and a magnifying glass, while sequencing the whole human genome, containing approximately 3 billion base pairs, would have been unimaginable without the use of automated robotic devices and computer systems.

In the context of medical diagnosis, we nowadays benefit from the advent of non-invasive devices, e.g. Computer Tomography (CT), Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI), all of them providing insights into an organisms inner life without any physical interference. In opposition to the former two technologies, whose scans are inherently accompanied with a notable amount of radiation exposure (a single scan exposes several times more radiation than the average natural background radiation in one year), MRI technology is free of radiation and instead based on magnetic fields and radio frequency energy. Thus, MRI constitutes an important monitoring technology and is applied in many hospitals and other medical institutions all over the world. As a drawback however,

due to the magnetic field, MRI cannot be applied if magnetic materials like iron, cobalt or nickel are present in the body under investigation, e.g. contained in heart pacemakers, implants, prostheses, etc.

MRI is based on the phenomenon of NMR, and allows one to visualize the structure and function of tissue and organs. As far as one is interested in measurements at the molecular level, NMR spectroscopy as the originating technology allows to quantitatively monitor classes of molecules non-invasively and non-destructively for the same reasons as mentioned for MRI. Thus, NMR spectroscopy has much potential in analyzing molecular systems of living cells and organisms.

Regarding the complex nature of living systems in general, high-throughput monitoring of biological entities in a reliable, quantitative and global manner seems to be a necessity in order to fully understand the underlying mechanisms and principles of life. Thereby, non-invasive experiments on living cells or living organisms obviously have more potential to reveal insights of interest rather than on lifeless material. NMR spectroscopy is based on magnetic fields and radio frequency energy, its experiments are non-invasive, and the signal represents quantitative information. Therefore, it is considered to have much potential in the context of *Systems Biology*, the inter-disciplinary field of research which studies interactions in biological systems on the cellular and molecular level.

The analysis of metabolites, those small bio-molecules which occur as intermediates or products of chemical reactions within living organisms, is often referred to as *Metabonomics* or *Metabolomics*. The latter is more interested in monitoring the metabolism of an organism in whole, hence also facing the identification of not yet assigned peak patterns of unknown compounds (Fiehn, 2002), while the former is more focused on understanding the responses of known metabolites to certain external stimuli (Nicholson *et al.*, 1999).

For the measurement of metabolites, NMR Spectroscopy has become an important technology due to its non-invasive and quantitative nature (Lindon & Nicholson, 2008). However, NMR experiments yield a mixture of partly overlapping signals, altogether representing the molecular substructures which are contained in a sample at once. Thus, for samples containing hundreds of different metabolites and more, which is commonly the case for the measurement of living cells, the extraction of quantitative and even only qualitative metabolite information out of an NMR experiment often requires extensive care and manual processing due to

the complex nature of the data. Thus, analyzing NMR data of biological samples is commonly a tedious and time consuming task.

1.2 Aims of the thesis

Up to now, no golden standard for automated analysis and interpretation of NMR data has been made. Establishing automated approaches allowing to reliably extract the quantitative information provided by each experiment is a challenging task, especially with respect to the essential need of high-throughput experiments resulting in data series containing hundreds of spectra and more in order to comprehensively understand the molecular system of living cells and organisms. This thesis aims at developing an automated methodology for preserving and extracting valuable information out of an NMR spectrum in order to allow reliable, automated and quantitative analysis of NMR spectroscopic datasets, and to facilitate research in the area of *Systems Biology* in general.

Chapter 2

Basics in NMR Spectroscopy

The discovery of NMR spectroscopy goes back to the work of Rabi *et al.* (1938), Purcell *et al.* (1946) and Bloch (1946). In the first three decades, all experimental setups were one-dimensional (1D NMR), where signal intensities are displayed along a single frequency axis. This approach is called the classical continuous-wave approach. The development of pulse Fourier spectroscopy in the 1970s, in conjunction with the development of the Fast Fourier Transform (FFT) algorithm in the same time period, then revolutionized the classical approach, leading to essential improvements in acquisition time and sensitivity. This thesis focuses on pulse Fourier NMR experiments. In the remainder of this chapter, a rough overview of physical foundations in pulse Fourier 1D NMR spectroscopy is given based on Friebolin (1999) as long as not indicated otherwise.

2.1 The NMR spectrometer

In this section, a basic overview of an NMR device is given. Figure 2.1 shows an illustration of a standard NMR device. In general, a spectrometer consists of a magnet with corresponding magnetic coils, chambers of liquid helium and liquid nitrogen, vacuum chambers, the sample tube and a shimming unit.

- Magnet:
  The magnet is the essential component in a spectrometer, since the quality of the experiments do strongly depend on it. Where the NMR experiments in the 1960s based on permanent magnets or electromagnets with a magnetic

flux density of up to 1.41 tesla, todays NMR experiments are based on superconducting magnets with a magnetic flux density of up to 23.5 tesla. It will become more clear in the remainder of this chapter why a high magnetic flux density is beneficial in the context of NMR experiments, but as a rule of thumb one can say "the higher the better".

- Sample tube:
  The core of an NMR decive is given by the sample tube. It comprises the sample, the transmitter and reciever coils, and the shimming unit, with the latter being used to establish homogeneity of the magnetic field.

- Transmitter:
  The transmitter unit is basically given as a radio frequency generator and a frequency synthesizer, and produces the pulses of radio frequency energy needed for the experiments.

- Receiver:
  As the name implies, the receiver coil is used to detect the NMR signal. Commonly, the detection process is accompanied by signal amplification units similar to radio technology.

- Computer:
  The whole experimental setup of an NMR experiment is controlled by the computer. As well as supplying certain parameters for the shimming and the transmitter units in view of automation, the analogue NMR signal is finally recorded in a digital way for further processing, i.e. analysis, simulation, prediction and interpretation of an experiment. The methods proposed within this thesis belong to this category.

2.2 The NMR Experiment

A classical NMR experiment is illustrated in figure 2.2 and can be described in three steps:

1. A strong magnetic field B_0 induces macroscopic magnetization, having the effect that nuclear spins align with the field.

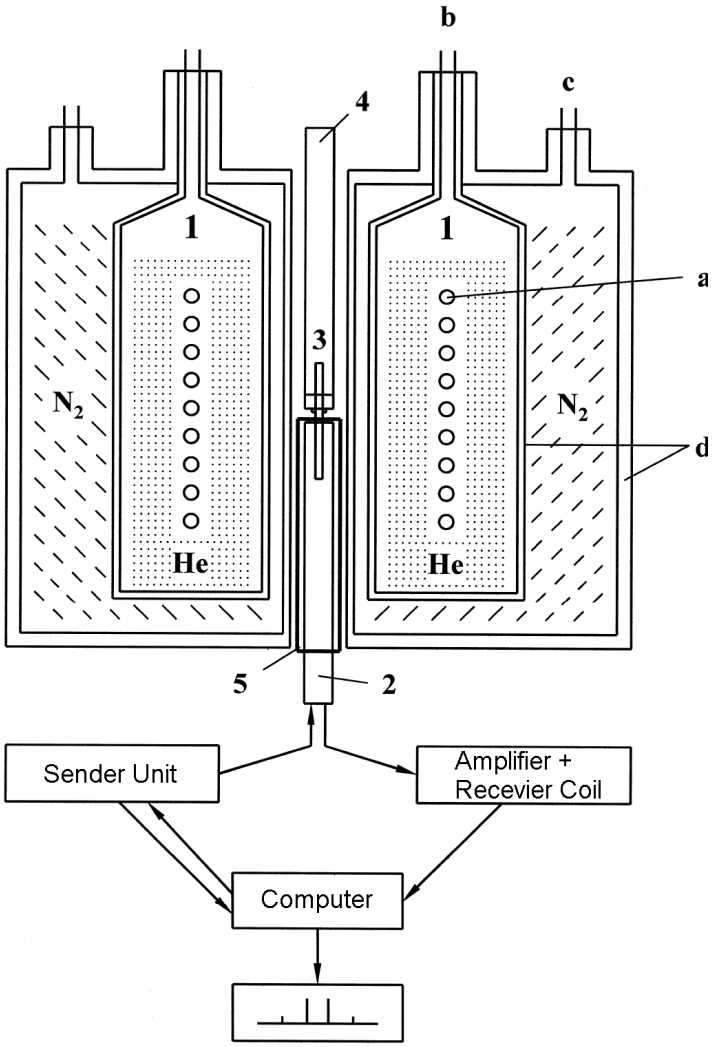


Figure 2.1: Schematic illustration of a standard NMR device (Friebolin, 1999). 1: Magnet, a) Magnetic coil, b,c) Filler for liquid helium and liquid nitrogen, respectively, d) Inner and outer vacuum chambers, 2: Sample tube, 3: Sample probe, 4: Probe changer, 5: Shimming unit

2. High intensity pulses B_1 are used to excite a particular type of the sample's nuclei.

3. After the pulse has been applied, the nuclei induce a current in the receiver coil due to the precession of the macroscopic magnetization, which leads to the signal in the time-domain called *free induction decay* (FID).

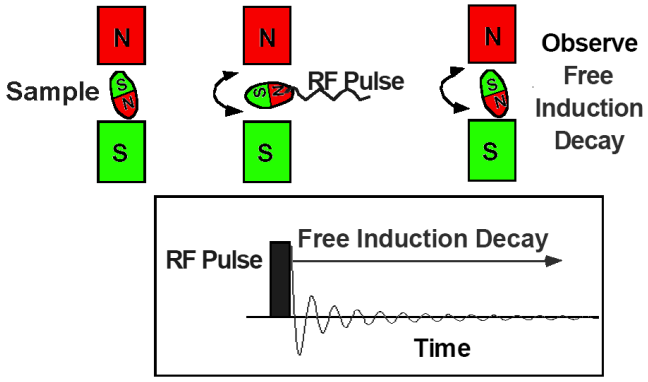


Figure 2.2: Illustration of a classical NMR experiment based on Campbell *et al.*. See the text for more details.

A more formal description of key aspects in NMR spectroscopy is given in the following subsections.

2.3 Magnetic Properties of a Nucleus

The NMR technology is based on radio waves and magnetic fields, and measures the resonance of nuclei based on their magnetic properties, with Proton (H) and Carbon-13 (^{13}C) being most commonly studied. The number "13" signalizes the isotope of Carbon. Other examples of investigated nuclei are Nitrogen-15 (^{15}N), Fluorine-19 (^{19}F) and Phosphorus-31 (^{31}P).

In a static magnetic field, the angular momentum P of a nucleus is described as

$$P = \frac{h}{2\pi}\sqrt{I(I+1)} \tag{2.1}$$

where h is PLANCK's constant h= $6.6256 \cdot 10^{-34}$ Js, with J (joule) being the unit of energy, and where I denotes the *spin quantum number*, also simply called *spin*. P is given in units of Js (joule seconds), and describes the rotational state of a nucleus. I arises in $\frac{1}{2}$ steps from the difference in the number of protons and neutrons of the nucleus, and can take values from $I \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \ldots, 6\}$. For example, Proton, Carbon-13 and Fluorine-19 have a spin of $I = \frac{1}{2}$.

Nucleus	$\gamma/2\pi$ **(MHz/tesla)**
^{1}H	42.546
^{13}C	10.705
^{15}N	-4.3156
^{19}F	40.0541
^{31}P	17.235

Table 2.1: Gyromagnetic ratios for some example nuclei (following Friebolin (1999)).

The angular momentum P is quantized in the context of nuclear spins, associated with a so called *magnetic quantum number* $m = \{-I, -I+1, \ldots, I\}$, describing in total $2I+1$ *angular momentum states* of a given nucleus. It is worth noting, that neither P nor I can be theoretically predicted.

The magnetic moment μ of a nucleus is a measure of the strength and the direction of its magnetization, in units of J/tesla, where tesla is the unit of the magnetic flux density of a magnetic field. μ is associated with the angular momentum P by the nucleus-specific gyromagnetic ratio γ as

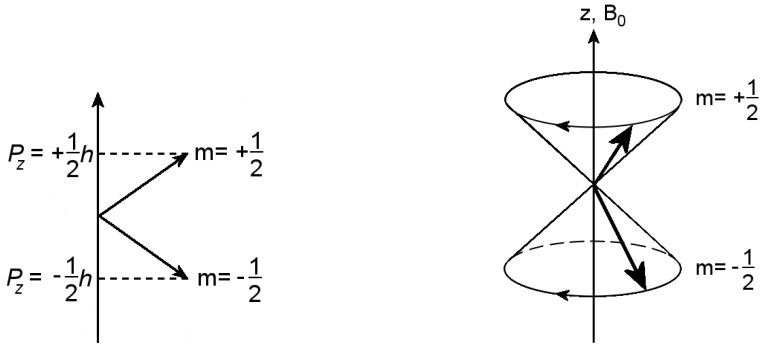
$$\mu = \gamma\, P. \tag{2.2}$$

Nuclei with an equal number of protons and neutrons have a *spin quantum number* $I = 0$, and thus zero angular momentum P and zero magnetic momentum μ. For example, the regular carbon-atom ^{12}C has zero angular momentum and thus is not detected in NMR spectroscopic experiments. All nuclei with non-zero angular and magnetic momentum maintain a characteristic gyromagnetic ratio γ, with high values for γ indicating better detection in NMR experiments. Some approximate gyromagnetic ratios are shown in table 2.1.

2.4 Alignment in a Magnetic field

For a nucleus with angular momentum P and magnetic moment μ placed in a static magnetic field B_0 the z-component P_z of P, namely the angular momentum in direction of B_0, is given as

$$P_z = m\frac{h}{2\pi}$$



(a) Angular momentum states of spins with $I = \frac{1}{2}$.

(b) Double precession cone for spins $I = \frac{1}{2}$.

Figure 2.3: Angular momentum states and the precession in a static magnetic field B_0.

where m again denotes the *magnetic quantum number* with $m = \{-I, -I + 1, \ldots, I\}$, and h stands for PLANCK's constant. Figure 2.3(a) illustrates the corresponding angular momentum states for spins $I = \frac{1}{2}$. Accordingly, the magnetic moment μ_z in direction of the magnetic field B_0 is then given as

$$\mu_z = m\gamma \frac{h}{2\pi}. \tag{2.3}$$

The *larmor precession rate* ω_L, describing the rotation frequency of a nucleus, is thereby proportional to the magnetic field strength B_0, given as

$$\omega_L = \frac{|\gamma|}{2\pi} B_0. \tag{2.4}$$

Note that γ is specified by the type of nucleus, and therefore the *larmor precession rate* ω_L is nucleus-specific as well.

In the classical view, the nuclei rotate around the z-axis, the direction of the magnetic field B_0 (see figure 2.3(b)). They behave like little gyroscopes, with the exception that only certain angles are allowed due to their quantized angular momentum P. For example, each proton (H) with $I = \frac{1}{2}$ rotates at the same angle $54°44'$ around the z-axis, the direction of the magnetic field B_0.

In a magnetic field, the energy E of a magnetic dipole is proportionally related to its magnetic moment μ_z and the magnetic flux density B_0 by

$$E = -\mu_z \, B_0,$$

and combining this with formula (2.3) gives

$$E = -\frac{m\gamma h B_0}{2\pi}.$$

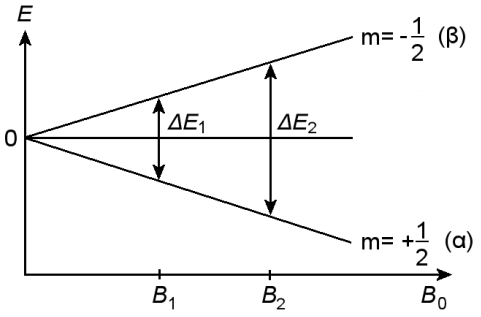
In thermal equilibrium, atoms of the same nucleus type are approximately equally distributed in the different angular momentum states m (see previous section). In a static, non-zero magnetic field of strength B_0 however, due to interactions between the field and the nuclear magnetic moments, the states m differ in their energy level. In the case of nuclei with spins $I = \frac{1}{2}$ (e.g. ${}^1H, {}^{13}C$), the natural distribution in the number of atoms N_α and N_β of the respective two states $m = -\frac{1}{2}$ and $m = \frac{1}{2}$ in equilibrium is described by the BOLTZMANN distribution as

$$\frac{N_\alpha}{N_\beta} = e^{-\frac{\Delta E}{k_B \mathrm{K}}} \approx 1 - \frac{\Delta E}{k_B T} \tag{2.5}$$

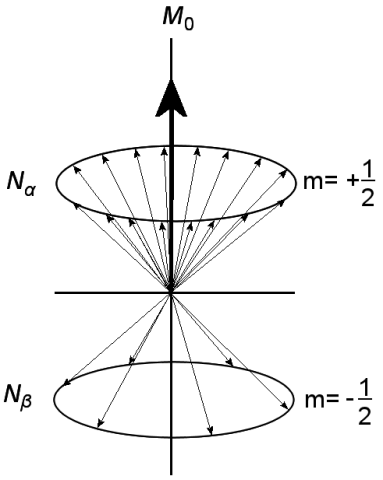
where K denotes the absolute temperature in Kelvin, k_B is the BOLTZMANN constant $k_B = 1.3805 \cdot 10^{-23} \frac{\mathrm{J}}{\mathrm{K}}$, and $\Delta E = \gamma \frac{h}{2\pi} B_0$ stands for the energy difference between the two states $m = -\frac{1}{2}$ and $m = \frac{1}{2}$ (see figure 2.4(a)). In other words, the two spin states of a spin $I = \frac{1}{2}$ align either *parallel* or *antiparallel* to the magnetic field, according to (2.5). In total, the lower energy state occurs more often than the higher energy state. Summation over all magnetic moments μ_z in direction of the magnetic field then leads to a tiny but measurable macroscopic magnetization M_0 of the sample (see figure 2.4(b)). Thus, M_0 gives quantitative information concerning the number of nuclei and in general plays a key role in describing pulse NMR experiments, as we will see in the following sections.

2.5 Magnetic Resonance of the Nuclei

In NMR experiments, electromagnetic radiation of radio frequency is applied in order to induce transitions between the angular momentum states of a particular type of nucleus. Thereby, transitions from states with lesser to those with higher energy are called energy absorption, and vice versa energy emission. Since more spins initially align in the angular momentum state of lower energy ($N_\alpha > N_\beta$), energy absorption is observed in total, whereby the intensity of the absorption is proportional to $N_\alpha - N_\beta$ and is therefore also proportional to the total number of



(a) Energy difference between two states in dependence to the magnetic field strength B_0.



(b) State distribution on the double precession cone. $N_\alpha > N_\beta$ results in a macroscopic magnetization.

Figure 2.4: Energy difference and distribution of the angular momentum states for spins with $I = \frac{1}{2}$.

respective nuclei contained in a given sample. This magnetic resonance absorption is detected as the quantitative signal in NMR experiments.

A transition only occurs, if the corresponding radiation frequency ω_1 equals the *larmor precession rate* ω_L (2.4) of the respective nucleus under investigation, formally described as the *resonance condition*:

$$\omega_1 = \omega_L = \frac{|\gamma|}{2\pi} B_0. \tag{2.6}$$

ω_L is also called the *larmor-frequency.*

As a result, the investigation can be focused on specific nuclei depending on which frequency ω_1 is applied in an experiment. For Protons, the gyromagnetic ratio γ is known to be 42,546 MHz/tesla, and NMR experiments with $\omega_1 = |\frac{42{,}546}{2\pi}|B_0$ are denoted as ^{1}H NMR experiments. Analogously, ^{13}C NMR experiments operate at the frequency 10.705 MHz/tesla for ^{13}C (see table 2.1).

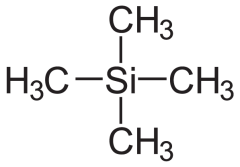


Figure 2.5: Chemical structure of *tetramethylsilane* (TMS).

2.6 Chemical Shift

The local magnetic field experienced by nuclei differs from the total magnetic field B_0. The nuclei are said to be *shielded*. A measure for this effect is given by the *shielding constant* σ, leading to an expansion of the *resonance condition* (2.6) as

$$\omega_1 = \frac{|\gamma|}{2\pi}(1-\sigma)B_0. \tag{2.7}$$

Theoretical approaches are still unable to precisely calculate σ with respect to a given sample. As a result, the effective local magnetic field and thus the effective resonance frequency of a particular substrate cannot be specified prior to an experiment. However, theory and experiments indicate that the differences between the local and the total magnetic fields, and thus changes in the resonance frequencies, mainly depend on the electron density distribution of the molecules, but are independent to the total magnetic field B_0. Next to the structure and the type of a molecule, molecular interactions due to concentration and pH-value[1] are the main contributors to the *shielding* effect.

NMR spectroscopic datasets are commonly free of absolute values, since the magnetic field B_0 and the resonance frequencies ω_i are proportionally related to each other (2.6, 2.7). Instead, resonance frequencies are measured relative to a reference. For this purpose, *tetramethylsilane* (TMS) ($Si(CH_3)_4$) is commonly used for its beneficial properties of being chemically inert[2] and being easily removed from a solution due to its low boiling point of $26,5°$ C under normal conditions[3]. Figure 2.5 shows the chemical structure of TMS.

[1]The pH-value is a measure of the acidity or basicity of a solution, and expresses the activity of dissolved hydrogen ions H^+.

[2]A chemically inert substrate is chemically not active.

[3]Normal conditions in terms of approx. 1013 hPa = 1.013 bar of air pressure

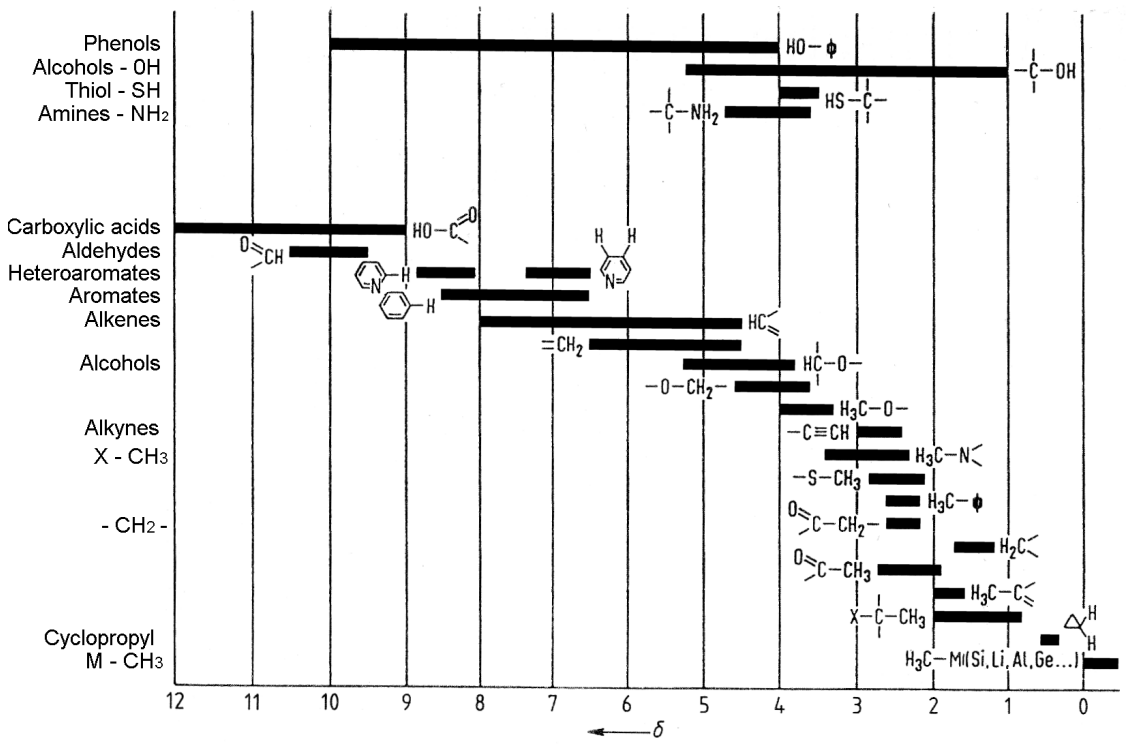


Figure 2.6: Chemical shifts for protons of organic compounds, based on Friebolin (1999).

In order to compare experimental results obtained with different external magnetic fields B_0, the resonance signals are commonly given by the dimensionless value δ as

$$\delta = \frac{\omega_{substrate} - \omega_{reference}}{\omega_{reference}} 10^6.$$

The factor 10^6 has the purpose of simplifying the resulting numbers. Thus, δ is given in "parts per million" (ppm), and called the *chemical shift* of a substrate. For ^{1}H NMR experiments, δ commonly ranges from 14 ppm to -3 ppm, and is given in descending order to display the NMR signal in terms of ascending shielding constants σ. Figure 2.6 provides an overview to frequency ranges of several classes of organic substrates.

2.7 Relaxation and the Time-Domain Signal

Applying B_1 at frequency ω_1 on a given sample has the effect that the initial macroscopic magnetization vector M_0 changes its direction by pulse angle Θ as

$$\Theta = \gamma B_{1i} \tau_P \tag{2.8}$$

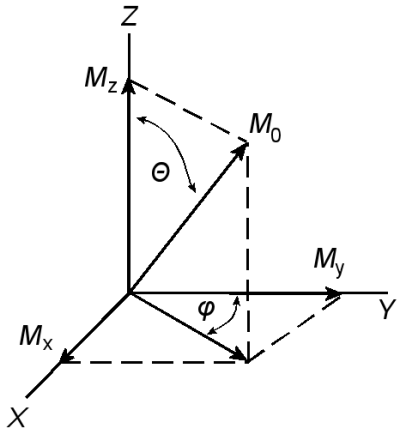
with B_{1i} as the intensity and τ_P as the pulse length of the high-frequency pulse B_1, commonly in the range of several μs. The change of M_0 is illustrated in figure 2.7(a). However, the movement of M_0 is hard to comprehend with regard to a static coordinate system x, y, z. The whole process becomes more understandable by changing the view to a rotating coordinate system x', y', z around the z-axis at the same frequency as ω_1. In this way the precession around the z-axis can be neglected. Remember, z stands for the direction of the static magnetic field B_0, the axis around which the precession takes place.

Common pulse angles are $90^\circ_{x'}$ and $180^\circ_{x'}$, with x' denoting the direction of the pulse in the rotating frame. Figure 2.7(b) shows the effect of applying a $90^\circ_{x'}$-pulse. Due to the impact of B_1, the precession of the nuclei around the z-axis is not equally distributed on the surface of the double precession cone anymore, but slightly biased toward the $M_{y'}$ vector, resulting in a transverse magnetization in direction of the y'-axis. In total, the macroscopic magnetization vector M_0 changes its direction in the y', z-plane around the x'-axis by pulse angle Θ, and precesses around the z-axis at the *larmor-frequency* ω_L (2.4, 2.7). As soon as B_1 is switched off, the macroscopic magnetization vector M_0 returns back to its initial alignment in the static magnetic field B_0, a process known as *relaxation*.

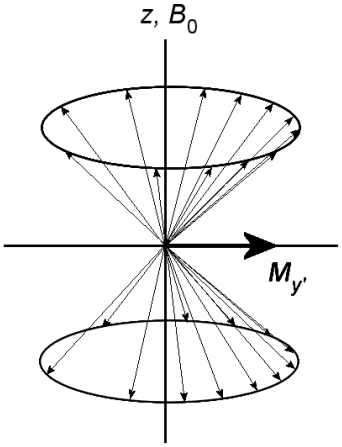
Back in the equilibrium state after relaxation, it holds $M_0 = M_z$ and $M_x, M_y = 0$. The behavior of an isolated magnetization vector M_0 during the relaxation process has been mathematically analyzed by Bloch (1946) on the basis of two simplifications:

1. "That the changes of orientation of each nucleus are solely due to the presence of the external magnetic fields", and
2. "That the external fields are uniform throughout the sample,"

which amongst other things means that interactions between nuclei are omitted. Nevertheless, by changing the spatially stationary coordinate system as shown in



(a) Applying pulse B_1 changes the orientation of the macroscopic magnetization vector M_0 by pulse angle Θ and leads to a precession of M_0 around the $z-$axis at *larmor-frequency* v_L.



(b) Magnetization bias toward the $M_{y'}$ vector as an effect of applying a $90^\circ_{x'}$-pulse.

Figure 2.7: Change of the magnetization vector M_0 after applying a high-frequency pulse.

figure 2.7(a) into a rotating coordinate system x', y', z as mentioned before, the precession around the z-axis can be neglected, and the relaxation process for an isolated spin system is described by the following *Bloch equations*:

$$\frac{\mathrm{d}M_z(t)}{\mathrm{d}t} = -\frac{M_z(t) - M_0}{T_1}$$
$$\frac{\mathrm{d}M_{x'}(t)}{\mathrm{d}t} = -\frac{M_{x'}(t)}{T_2}$$
$$\text{and} \quad \frac{\mathrm{d}M_{y'}(t)}{\mathrm{d}t} = -\frac{M_{y'}(t)}{T_2}$$

with T_1 and T_2 denoting the time constants for the relaxation processes along the z-axis and the $x-$ and y-axes, respectively. Following Schorn (2001), the solutions for $M_{x'}(t)$ and $M_{y'}(t)$ are given as

$$M_{x'}(t) = M_0 \sin\Theta \cos(\omega_1 t)\, e^{\frac{t}{T_2}}$$
$$\text{and} \quad M_{y'}(t) = M_0 \sin\Theta \sin(\omega_1 t)\, e^{\frac{t}{T_2}}.$$

In terms of practical experiments however, inhomogeneities of the magnetic field, interactions with neighbouring nuclei, and even instrumental contributions result in varying precession rates with varying phases and varying relaxation times even for chemically equivalent nuclei. A thorough mathematical description of the underlying mechanisms is provided by the *quantum-mechanical relaxation theory* (Ernst *et al.*, 1987), but is beyond the scope of the thesis.

In the classical description, the resulting signal is given by $|J|$ distinct groups of nuclei, maintaining their own precession rate ω_j, their own phase φ_j and their own relaxation time $T_{2_j}^*$. Following Ernst *et al.* (1987) and Jarvi *et al.* (1997), the subsequent signal along the axes x' and y' can be described as

$$\begin{aligned} M_{x'}^j(t) &= M_j \sin\Theta \, \cos(\omega_j \, t + \varphi_j) e^{-\frac{t}{T_{2_j}^*}} \\ \text{and} \quad M_{y'}^j(t) &= M_j \sin\Theta \, \sin(\omega_j \, t + \varphi_j) e^{-\frac{t}{T_{2_j}^*}}, \end{aligned}$$

with Θ denoting the pulse angle (2.8). In complex notation, $M_{x'}^j(t)$ and $M_{y'}^j(t)$ are summarized as

$$M^j(t) = M_{x'}^j(t) + \mathit{i} M_{y'}^j(t) = M_j \sin\Theta \, e^{\mathit{i}(\omega_j \, t + \varphi_j) - \frac{t}{T_{2_j}^*}}$$

with i denoting the imaginary number with $\mathit{i}^2 = -1$. The measured time signal $s(t)$, obtained by simultaneous observation of both x- and y-axes, is directly proportional to the complex magnetization $M^j(t)$. It is for each group of nuclei j given as

$$s(t) = \sum_j^{|J|} A_j \, \sin\Theta \, e^{\mathit{i}(\omega_j \, t + \varphi_j) - \frac{t}{T_{2_j}^*}}, \tag{2.9}$$

and thus reflects the signal proportional to the number of responding nuclei of group j. The signal $s(t)$ is also known as the *free induction decay* (FID).

With applying a $90^\circ_{x'}$-pulse ($\Theta = 90^\circ$), the idealized[4] time domain signal $s(t)$ as a sum of exponential decays is given as

$$s(t) = \sum_j^J A_j e^{\mathit{i}(\omega_j \, t + \varphi_j) - t/T_{2_j}^*}.$$

[4]The model analytically describes the pure NMR signal without considering effects of noise and other distortions.

In summary, the recorded time-domain signal quantitatively reflects the differences in the relaxation process of all responding nuclei, thus allowing different molecular substructures to be distinguished from one other.

2.8 The Signal in the Frequency Domain

As shown in the previous section, the signal of an NMR measurement is given as a weighted sum of oscillating functions whose envelopes are exponentially decaying in time (2.9). For reasons of better visual analysis and interpretation, the Fourier Transform (FT) is applied on the signal, resulting in a spectrum in the frequency domain, which basically means a separation of the resonance frequencies along the frequency axis.

The Fourier Transform goes back to the work of the french mathematician Jean Baptiste Joseph Forier (1768-1830), who amongst other things introduced the *Fourier Series*. In basic summary, they represent the decomposition of any periodic function $f(t)$ with basic frequency $v = \frac{1}{T}$ as a superposition of *sine* and *cosine* functions (Popov, 1990):

$$f(t) = f(t - mT) = \sum_{j=1}^{\infty} \big(b_j \sin(2\pi jvt) + c_j \cos(2\pi jvt)\big), \tag{2.10}$$

where m, j are integers. Equation (2.10) is also known as the *fourier series expansion*. In case of a non-periodic function, an analogous relationship exists:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\big(\cos(\omega t) - i\sin(\omega t)\big)dt = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, \tag{2.11}$$

with $\omega = 2\pi v$. Equation (2.11) represents the *Fourier Transform*, and the *inverse* FT is defined as

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)\, e^{i\omega t}d\omega.$$

In the context of NMR, fourier transforming the FID results in the signal function $S(\omega)$ in the frequency domain, with the real part $a_j(\omega)$ and imaginary part $d_j(\omega)$ of each component j (Ernst *et al.*, 1987; Popov, 1990):

$$S(\omega) = \int_0^{+\infty} s(t)\, e^{-i\omega t} dt = \sum_j^{|J|} e^{i\varphi_j} \left(a_j(\omega) + i d_j(\omega)\right),$$

$$\text{with } a_j(\omega) = A_j \frac{T_{2_j}^*}{1 + (\omega - \omega_j)^2\, T_{2_j}^{*2}}$$

$$\text{and } d_j(\omega) = -A_j \frac{T_{2_j}^{*2}(\omega - \omega_j)}{1 + (\omega - \omega_j)^2\, T_{2_j}^{*2}}$$

$a_j(\omega)$ is called *absorption signal*, $d_j(\omega)$ is known as the *dispersive signal*, and both differ from each other by 90° in their phase. Figure 2.8 provides a schematic illustration of example FIDs and the result after FT.

With the assumption, that a given spectrum can easily be phase corrected ($\varphi_j = 0 \quad \forall j \in \{1, \ldots, |J|\}$), (2.12) can be rewritten as

$$S(\omega) = \sum_j^{|J|} A_j \frac{T_{2_j}^*}{1 + (\omega - \omega_j)^2 T_{2_j}^{*2}},$$

and by replacing $T_{2_j}^* = \frac{1}{\lambda_j}$, it follows

$$S(\omega) = \sum_j^{|J|} A_j \frac{\lambda_j}{\lambda_j^2 + (\omega - \omega_j)^2} = \sum_j^{|J|} Y_j(\omega). \tag{2.12}$$

$Y_j(\omega)$ is also known as a Lorentz function, and denotes the contribution of a single spectral component, namely a single peak of the spectrum. Figure 2.9 shows an example metabolite NMR spectrum containing 32768 datapoints. The center and bottom graphs show horizontally zoomed regions from the respective spectrum above.

2.9 Multiplet patterns

As described in the previous section, the signal of an NMR spectrum is given as a superposition of Lorentz functions. However, each Lorentz function alone does not necessarily represent a single molecular substrate, as shown in the following. Figure

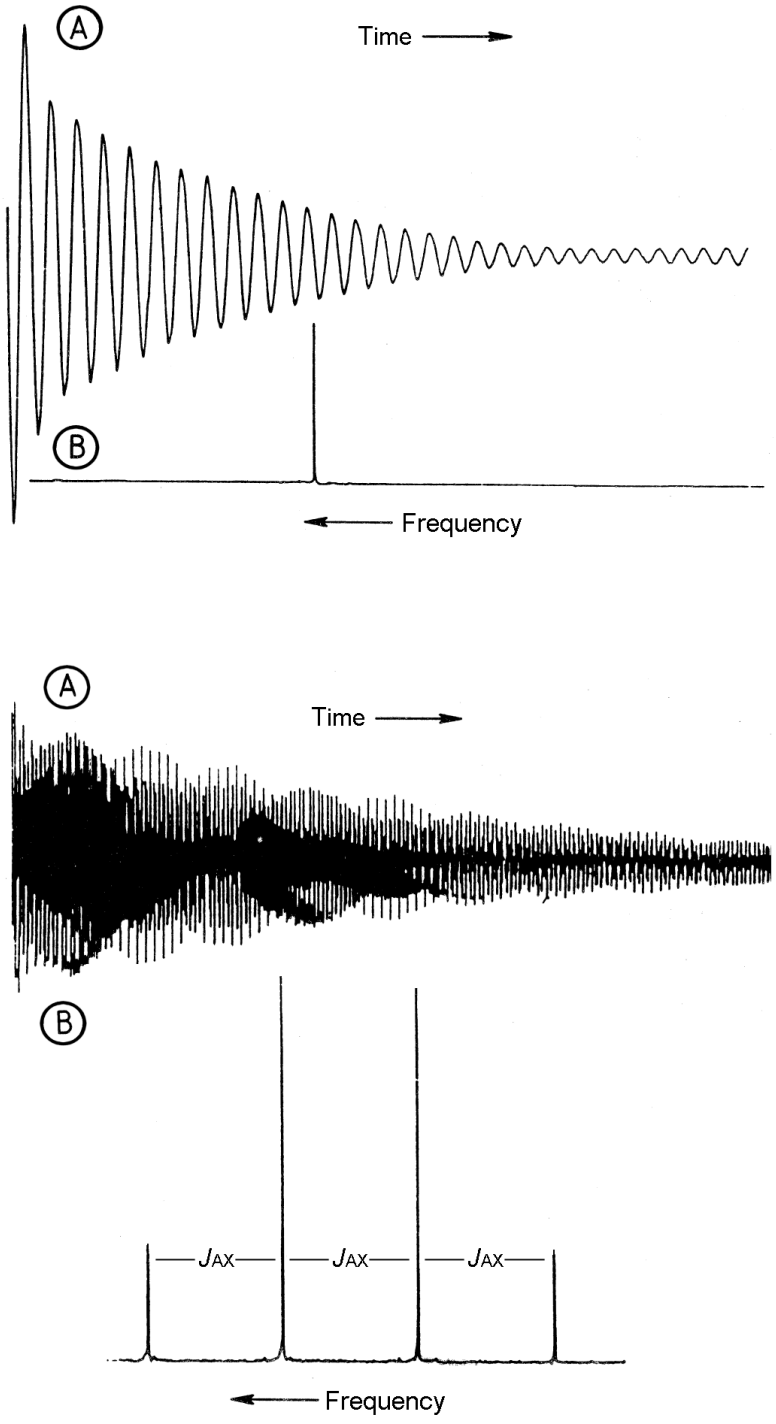


Figure 2.8: Schematic illustration of the Fourier Transformation, based on Friebolin (1999). Top: A single frequency in the time domain results as a single peak in the frequency domain. Bottom: Signals at four different frequencies and their corresponding spectrum. The frequency difference is denoted by J_{AX}.

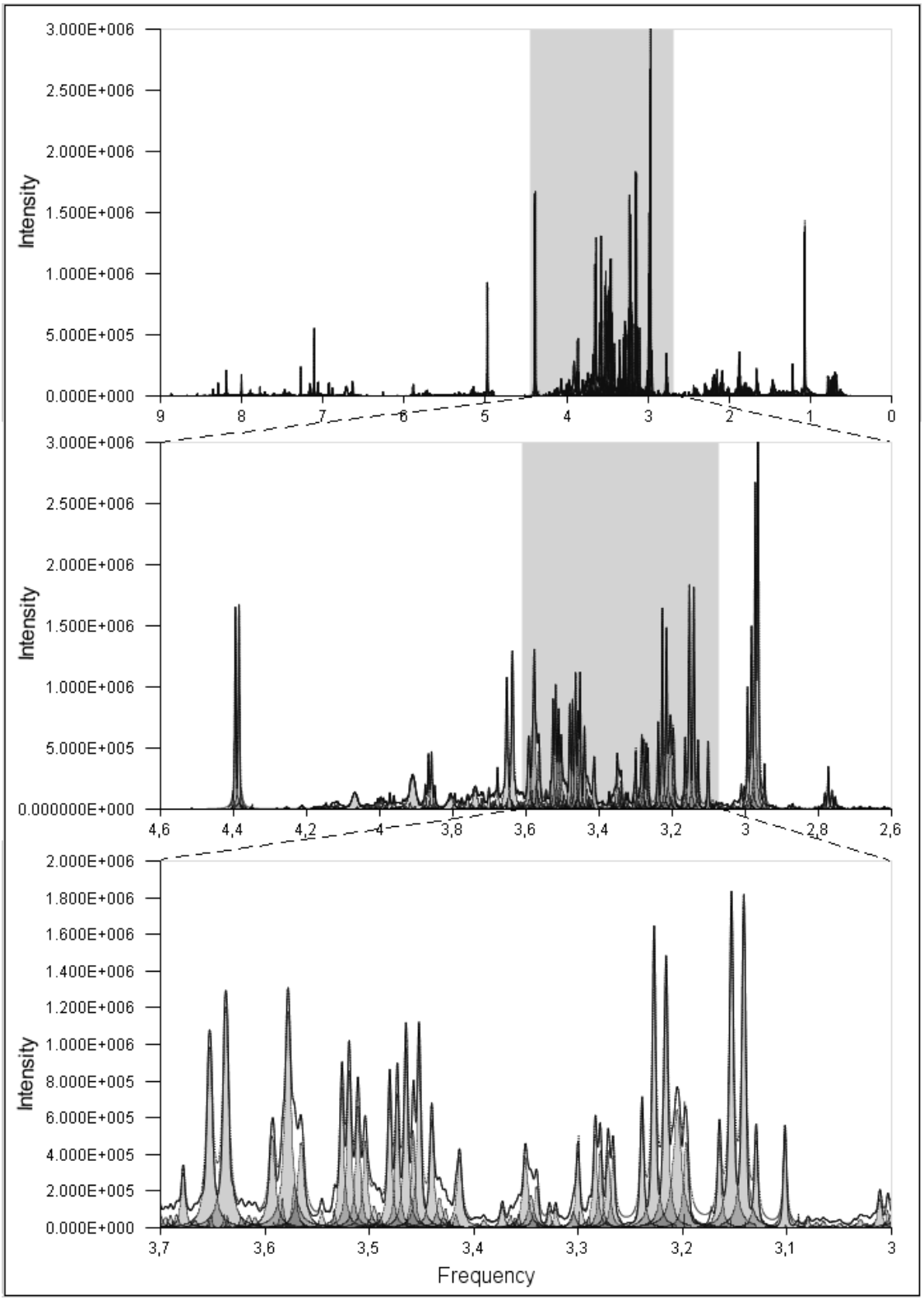


Figure 2.9: Example regions of a real-world NMR spectrum (solid line). A suggestion for the underlying Lorentz functions is given in grey colour.

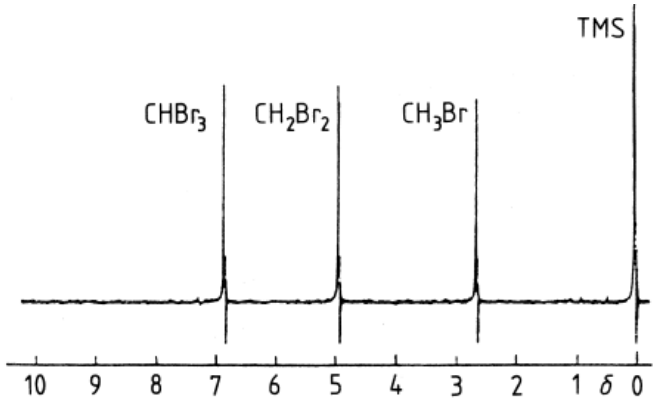


Figure 2.10: 90 MHz ^{1}H NMR spectrum of a mixture of $CHBr_3$, CH_2Br_2, CH_3Br and TMS (Friebolin, 1999).

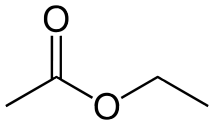
2.10 for example shows an ^{1}H NMR spectrum of a mixture of *bromoform* ($CHBr_3$), *methylene bromide* (CH_2Br_2) and *methyl bromide* (CH_3Br). Each component of the mixture results in a single peak called *singlet*, which can easily be explained by chemical equivalence of the respective nuclei in each molecule (they are all protons). However, NMR signals of more complex molecules commonly result in a specific pattern of peaks, as e.g. shown in figure 2.11 for the ^{1}H NMR signal of *ethyl acetate* ($CH_3COOCH_2CH_3$). From left to right, a *quartet*, a *singlet* and a *triplet* can be observed, although the protons within each of the denoted groups are chemically equivalent as well. The reason lies in the phenomenon called *spin-spin coupling*, basically referring to the fact that neighbouring nuclei affect the local magnetic field through the chemical bonds of a molecule, leading to increased or decreased field strengths, and therefore to changes in the observed resonance frequency (2.6). This interaction mainly evolves within 1-3 chemical bonds, and further distances have almost no impact on the observed signal.

In general, the number of signals in a multiplet, further denoted as the *multiplicity* M, mainly depends on the number of equivalent neighbouring nuclei and is given as

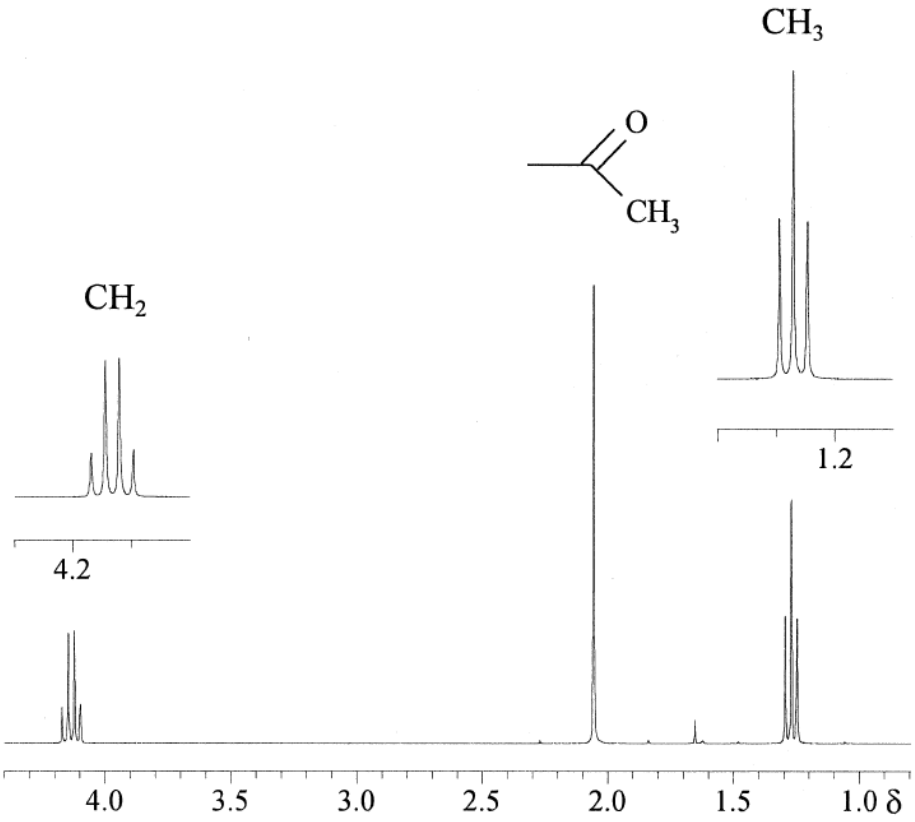
$$M = 2nI + 1,$$

with n denoting the number of neighbouring nuclei, and I denoting the *spin quantum number* (2.1). For protons ($I = \frac{1}{2}$), the above formula reduces to

$$M = n + 1,$$



(a) Chemical structure of ethyl acetate ($CH_3COOCH_2CH_3$).



(b) 300 MHz ^{1}H NMR spectrum of ethyl acetate

Figure 2.11: Chemical structure of ethyl acetate (top) and its corresponding ^{1}H NMR spectrum (bottom) as an example of spin-spin coupling effects between *equivalent* neighbouring nuclei, based on Friebolin (1999).

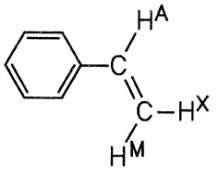
leading to intensity ratios within multiplets following the coefficients of the binomial series, which are provided by the Pascal's triangle. Table 2.2 shows the ratios for the first six multiplets.

The effect of *spin-spin coupling* between non-equivalent neighbouring nuclei commonly occurs differently, and Pascal's triangle cannot be applied anymore. Figure 2.12 shows the chemical structure and the ^{1}H NMR spectrum for three

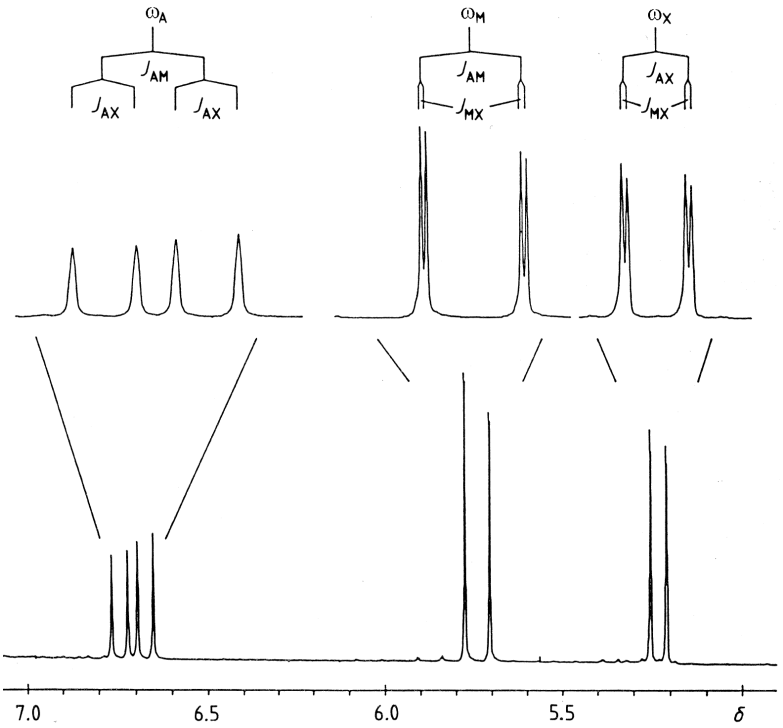
Multiplicity	**Intensity ratio**
Singlet	1
Doublet	1 : 1
Triplet	1 : 2 : 1
Quartet	1 : 3 : 3 : 1
Quintet	1 : 4 : 6 : 4 : 1
Sextet	1 : 5 : 10 : 10 : 5 : 1

Table 2.2: Pascal's triangle providing the coefficients of the binomial series. They are equal to the intensity ratios within multiplets of equivalent nuclei with spin $I = \frac{1}{2}$.

protons H_A, H_M and H_X of *styrene*. The initial resonance frequencies ω_A, ω_M and ω_X without coupling effects split two times for each neighbouring proton but with different impact, indicated by J_{AM}, J_{AX} and J_{MX}, the so called *coupling constants*. Interestingly, the value of the *coupling constants* only depends on the *angular moments* P of the respective nuclei, and is in opposition to the *chemical shift* therefore independent of the magnetic field B_0 (2.1). This also holds for *spin-spin coupling* between equivalent nuclei, and basically means that the splitting into multiplet patterns can in general be considered to be robust against variations in experimental conditions, e.g. concentration or pH-value. This characteristic property will become important in Chapter 6.



(a) Chemical structure of styrene



(b) Spin-spin coupling resulting in a splitting of signals

Figure 2.12: Chemical structure of styrene (top) and its corresponding ^{1}H NMR spectrum (bottom) as an example of spin-spin coupling effects between *non-equivalent* neighbouring nuclei, based on Friebolin (1999).

Chapter 3

NMR Feature Extraction

The term *Feature Extraction* basically means the process of constructing and selecting a set of relevant *features* in a given dataset, commonly aiming at reducing the amount of data under investigation while preserving the information content as well as possible. Beginning with a rough overview of related literature, this chapter presents some basic properties of a Lorentz function, and discusses some key aspects regarding the task of feature extraction in NMR data analysis.

With regard to the fact that each measured spectrum results in a finite list of n discrete datapoints, $\mathbf{w} = \{w_1, \ldots, w_n\}$ denotes their positions in descending order $w_1 > \cdots > w_n$ according to the convention in NMR spectroscopy (Friebolin, 1999). The corresponding intensity values are in the remainder of this work denoted as $S(w_1), \ldots, S(w_n)$, and the spectrum is given as

$$S = \big\{\big(w_1, S(w_1)\big), \ldots, \big(w_n, S(w_n)\big)\big\}.$$

Further, the first discrete derivative $S^{'}$ of S is given as

$$\begin{aligned} S^{'} &= \big\{\big(w_1^{'}, S^{'}(w_1)\big), \ldots, \big(w_{n-1}^{'}, S^{'}(w_{n-1})\big)\big\} \\ &= \left\{\left(\frac{w_1 + w_2}{2}, S(w_2) - S(w_1)\right), \ldots, \left(\frac{w_{n-1} + w_n}{2}, S(w_n) - S(w_{n-1})\right)\right\}, \end{aligned}$$

and under the assumption of equal distance between any consecutive pair of frequencies $w_i, w_{i+1}, i \in \{1, \ldots, n\}$, the second discrete derivative S'' is given as

$$\begin{aligned} & S'' = \left\{ \left(w_2'', S''(w_2)\right), \ldots, \left(w_{n-1}'', S''(w_{n-1})\right) \right\} \\ \text{with} \quad & w_i'' = w_i \\ \text{and} \quad & S''(w_i) = S(w_{i-1}) + S(w_{i+1}) - 2S(w_i). \end{aligned}$$

Note, that the frequency information is here for datapoints denoted as w_i instead of ω, whereas the position parameters of a spectrum's Lorentz functions are denoted in a continuous manner by ω_j (2.12), accounting for the effects of discretization. Also note, that it is rather unlikely that the resonance frequency of a spectral component coincides with a discretely given frequency value of the resulting digitized spectrum.

3.1 Related Work

In practice, an NMR signal is distorted in many ways, ranging from inhomogeneous magnetization (Spielman *et al.*, 1988) and noise arising from different sources during the measurement (Hoult & Lauterbur, 1979) through artifacts of the Fourier Transform (Giancaspro & Comisarow, 1983; Verdun *et al.*, 1988) to frequency shifts mainly induced by molecular interactions within the sample itself. Many approaches to enhance the quality of the signal have been proposed, e.g. zero-filling and apodization (Verdun *et al.*, 1988; Bartholdi & Ernst, 1973; Ebel *et al.*, 2006), spectral noise filtering (Massaro *et al.*, 1989; Asfour *et al.*, 2000), phase correction and exact interpolation (Giancaspro & Comisarow, 1983; van Vaals & van Gerwen, 1990; Goto, 1998) and reference deconvolution (Morris *et al.*, 1997; Metz *et al.*, 2000; Li *et al.*, 2003).

Concerning the extraction of relevant features with regard to multivariate analysis and classification, various methods have been proposed in the literature and are essentially grouped into three classes:

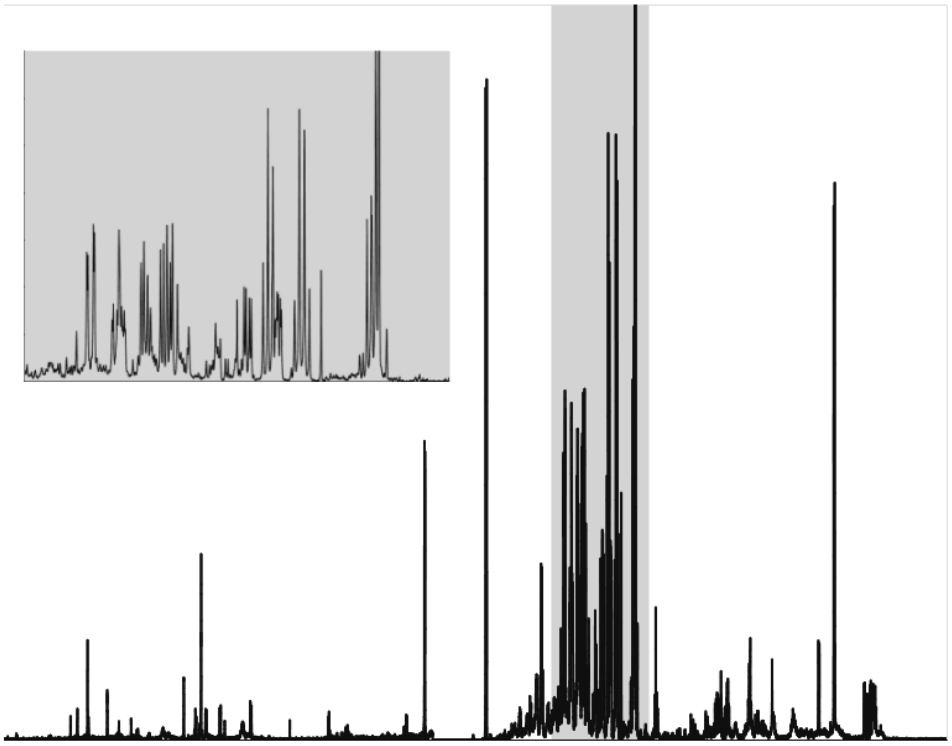
1. Spectral Binning
2. Targeted Profiling
3. Peak Selection and Parameter Approximation

3.1.1 Spectral Binning

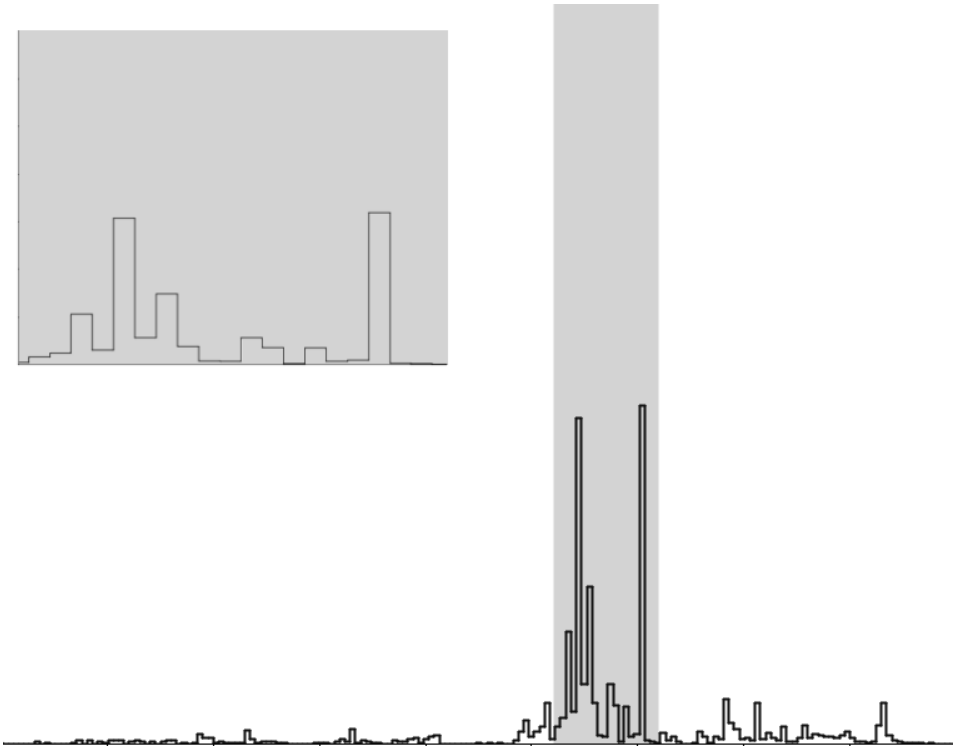
Spectral binning is often applied after Fourier Transform prior to *principal component analysis* (PCA) or *partial least squares - discriminant analysis* (PLS-DA) (Potts *et al.*, 2001; Wang *et al.*, 2003; Viant, 2003; Stoyanova *et al.*, 2004; Jansen *et al.*, 2005; Dieterle *et al.*, 2006). It is considered to potentially mitigate effects of peak shifts and other variations by averaging over a certain number of datapoints (Chang *et al.*, 2007). However, since these shifts in the frequency domain may commonly occur for each peak or peak pattern in each spectrum differently, single bins at same positions of different spectra may not contain signal from the same source of origin at all. Potential results are dramatic loss of spectral resolution, obscured feature vectors, and hence misinterpretation of the data, as also reported by Viant (2003) and Dijkstra *et al.* (2006). Figure 3.1 shows an example spectrum before and after binning to the bin size of 100 datapoints.

3.1.2 Targeted Profiling

As the NMR chemical shifts are sensitive to concentration, temperature and the pH-value of the metabolite solutions under investigation, the spectral response for a given metabolite differs from spectrum to spectrum. To circumvent this problem, compound-specific peak patterns are manually assigned in a process known as *Targeted Profiling* (Weljie *et al.*, 2006). As a result, an NMR spectrum series M containing m spectra is turned into a set of compounds $\{c_1, \ldots, c_k\}$, with the feature vector $\{v_{i,1}, \ldots, v_{i,m}\}$ for each compound i denoting its relative concentration in each of the m experiments. Subsequently, multivariate methods can be applied based on the achieved metabolite concentrations. The crucial part of this approach is the pattern assignment itself, which has been performed manually e.g. in Weljie *et al.* (2006). Although this approach seems to be very promising, the limitations are clear: manual assignment demands expert knowledge, is time consuming, and the outcome is restricted to the reference compounds database ab initio. Figure 2.6 of the previous chapter shows potential ranges of proton signals of typical molecular substructures. The huge degrees of overlap indicate, that the assignment is not a trivial problem, and generally considered time-consuming and error-prone.



(a) Raw spectrum containing 32768 dps.



(b) After binning to the bin size of 100 dps.

Figure 3.1: Example on effects of spectral binning, here with a bin size of 100 datapoints.

3.1.3 Peak Selection and Parameter Approximation

The last class of methods for extracting features out of a given NMR spectrum is also known as *Quantification*, comprising the two tasks of *peak selection* and *parameter approximation*.

3.1.3.1 Peak Selection

Peak selection, also known as *model selection*, stands for the identification of spectral components in a measured sample, i.e. the number of addends in (2.12). As far as frequency-domain methods are concerned, they are commonly found based on the occurrence of local maxima and applying a height threshold afterwards, a strategy which is often proposed in the context of spectroscopic data analysis (Yasui *et al.*, 2003; Jarvi *et al.*, 1997; Koradi *et al.*, 1998; Moseley *et al.*, 2004; Nguyen *et al.*, 2009; Dijkstra *et al.*, 2006). By this, the identification of overlapped Lorentz functions called "shoulders" is not well supported, which will be discussed in more detail in Section 3.3.2. As shown by Dijkstra *et al.* (2006), neglecting peak overlap may not only yield an incomplete model but also substantially falsify the height and area information of the detected peaks, and thus constitutes a major source of error.

3.1.3.2 Parameter Approximation

Motivated by the fact that ideally each resonance frequency of the measured time signal corresponds to a known analytical expression, the aim of these approaches is to approximate the corresponding parameters to accurately model the signal. Methods available in the literature are either directly operating on the measured FID in the time domain or after Fourier Transform. Approaches for the former class of methods are for example given by (Spielman *et al.*, 1988; Neil & Bretthorst, 1993; Miller & Greene, 1989; Vanhamme *et al.*, 2000b; Bretthorst *et al.*, 2005), a review is provided by Vanhamme *et al.* (2000a). Approaches of the latter class of methods are e.g. based on exact interpolation (Giancaspro & Comisarow, 1983; Goto, 1998), on the levenberg-marquardt algorithm (Marquardt, 1963; Jarvi *et al.*, 1997; Pons *et al.*, 1996), or on genetic algorithms (Metzger *et al.*, 1996; Karakaplan, 2007). A review can be found in Mierisov & Ala-Korpela (2001).

3.1.3.3 Levenberg-Marquardt Algorithm

A prominent method for the task of parameter approximation is given by the Levenberg-Marquardt method (LM). In the following, an outline of the method is given. The more interested reader is referred to Press *et al.* (2007).

Definition 3.1 (Least-Squares Problem)
Given a vector of real-valued datapoints $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ *with* $x_i, y_i \in \mathbb{R}$, $i \in \{1, \ldots, n\}$, *given the parameterized model function*[1] $f : \mathbb{R} \to \mathbb{R}$ *with* m *parameters* $\mathbf{a} = \{a_1, \ldots, a_m\}$, *find the parameter set* $\boldsymbol{a}^* = \{a_1^*, \ldots, a_m^*\}$ *which (locally) minimizes the following merit function:*

$$\chi(\mathbf{a}) = \sum_{i=1}^{n} \big(y_i - f(x_i, \mathbf{a})\big)^2. \tag{3.1}$$

Several methods exist to solve the aforementioned problem, e.g. the simplex method, singular value decomposition, simulated annealing, and also the Levenberg-Marquardt method.

In the context of spectral NMR data, the model function can be described based on (2.12):

$$f(\omega, \mathbf{a}) = \sum_{j}^{|J|} A_j \frac{\lambda_j}{\lambda_j^2 + (\omega - \omega_j)^2} = \sum_{j}^{|J|} Y_j(\omega, \omega_j, \lambda_j, A_j),$$
$$\text{with parameters } \mathbf{a} = \{\omega_1, \ldots, \omega_{|J|}, \lambda_1, \ldots, \lambda_{|J|}, A_1, \ldots, A_{|J|}\}.$$

The functions $Y_j(\omega, \omega_j, \lambda_j, A_j)$ are called the *basis* functions of f. For reasons of simplicity, they are in the remainder of the thesis denoted as $Y_j(\omega)$.

As can be observed, the model f depends *nonlinearly* on the parameter set $\mathbf{a}$. Finding the parameter set which minimizes the *merit* function χ is then called a *nonlinear* least-squares problem. LM is considered as the standard method for this type of problems, and has commonly been used for parameter approximation of spectral NMR datasets in the literature.

[1] With regard to equation (2.12), only the case of one independent variable is considered here. The *least-squares* problem can however easily be generalized to l-dimensional model functions $f : \mathbb{R}^l \to \mathbb{R}$ as well.

Basically, LM is a combination of two methods: The *Gauss-Newton* method and the *steepest descent* method. The former assumes, that the χ function can at any arbitrary parameter vector $\mathbf{P}$ be well approximated by a quadratic form:

$$\chi(\mathbf{a}) = \chi(\mathbf{P}) + \sum_i \frac{\partial \chi}{\partial a_i} a_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \chi}{\partial a_i \partial a_j} a_i a_j + \dots \tag{3.2}$$

$$\approx \chi(\mathbf{P}) - \mathbf{b} \cdot \mathbf{a} + \frac{1}{2} \mathbf{a} \cdot \mathbf{A} \cdot \mathbf{a}, \tag{3.3}$$

$$\text{where } \mathbf{b} := -\nabla \chi = -\left(\frac{\partial \chi}{\partial a_1}, \dots, \frac{\partial \chi}{\partial a_m} \right),$$

$$\text{and where } [\mathbf{A}]_{i,j} := \frac{\partial^2 \chi}{\partial a_i \partial a_j}.$$

(3.2) (first line) is the *taylor series* of χ, which is approximated by (3.3). The vector $\mathbf{b}$ denotes the *gradient* of χ at $\mathbf{P}$, namely the vector of first partial derivatives, and the matrix $\mathbf{A}$ is the second partial derivative matrix of χ, also known as the *Hessian matrix* of χ at $\mathbf{P}$.

Given a parameter vector $\mathbf{a}_{cur}$, the parameter vector $\mathbf{a}_{\min}$ which minimizes (3.3) can then directly be obtained by the following equation (*Gauss-Newton*):

$$\mathbf{a}_{\min} = \mathbf{a}_{cur} + \mathbf{A}^{-1} \cdot [-\nabla \chi(\mathbf{a}_{cur})]. \tag{3.4}$$

In case where (3.3) is a poor approximation of the χ function, the next set of parameters $\mathbf{a}_{next}$ can also be found by taking a step down the gradient (*steepest descent*), e.g. for parameter a_l given as:

$$a_{l,next} = a_{l,cur} - c \cdot \frac{\partial \chi}{\partial a_l}(a_{l,cur}). \tag{3.5}$$

c is a constant describing the size of the step.

(3.4) can be rewritten as a set of linear equations, e.g.

$$\sum_l^m \alpha_{k,l} \, \delta a_l = \beta_k, \tag{3.6}$$

$$\text{with} \qquad \delta a_l = a_{l,min} - a_{l,cur},$$

$$\beta_k = -\frac{1}{2}\frac{\partial \chi}{\partial a_k} = -2\sum_{i=1}^{n}\big(y_i - f(x_i,\mathbf{a})\big)\frac{\partial f(x_i,\mathbf{a})}{\partial a_k}, \tag{3.7}$$

$$\text{and} \qquad \alpha_{k,l} = \frac{1}{2}\frac{\partial^2 \chi}{\partial a_l\,\partial a_k} = 2\sum_{i=1}^{n}\left(\frac{\partial f(x_i,\mathbf{a})}{\partial a_k}\frac{\partial f(x_i,\mathbf{a})}{\partial a_l} - \big(y_i - f(x_i,\mathbf{a})\big)\frac{\partial^2 f(x_i,\mathbf{a})}{\partial a_l\,\partial a_k}\right). \tag{3.8}$$

Accordingly, (3.5) can be rewritten to give

$$\delta a_l = c \cdot \beta_l \tag{3.9}$$

Note that the term δa_l describes the change of parameter $a_{l,cur}$ in order to receive the next parameter $a_{l,next}$, either by (3.6) or by (3.9). In the Levenberg-Marquardt method, both equations are combined as described in the following:

With noting that the quantity χ is dimensionless, with further noting that in case of parameter a_l, δa_l has the dimensions of a_l, and with also noting that β_l has the dimensions of $\frac{1}{a_l}$, the constant c in equation (3.9) must have the dimensions of a_l^2. The only obvious quantity with these dimensions are given by $\alpha_{l,l}$, the diagonal elements of matrix $[\alpha]$. As a result, the step size c in (3.9) can be replaced by the term $\frac{1}{\mu\alpha_{l,l}}$, resulting in

$$\delta a_l = \frac{1}{\mu\,\alpha_{l,l}}\beta_l \quad \Leftrightarrow \quad \mu\,\alpha_{l,l}\,\delta a_l = \beta_l. \tag{3.10}$$

μ is used as a fudge factor in order to adjust the scale. With rewriting (3.6) as

$$\sum_{l}^{m}\alpha'_{k,l}\,\delta a_l = \beta_k, \tag{3.11}$$

$$\text{with } \alpha'_{i,j} = \begin{cases} \alpha_{i,i}(1+\mu), & \text{for } i = j, \\ \alpha_{i,j}, & \text{else,} \end{cases}$$

the variation term δa_l results as a smooth combination of both equations (3.6) and (3.9). For large μ, the diagonal elements dominate, and equation (3.11) becomes

similar to (3.10). Otherwise, (3.11) goes over to (3.6). The corresponding algorithm is outlined in algorithm 1.

Algorithm 1 Levenberg-Marquardt Algorithm (LM)

Input: Initial guess for the parameters $\mathbf{a}$,
Output: parameters $\mathbf{a}^*$ minimizing χ

1: **for** $b = 1$ to K **do**
2: Compute $\chi(\mathbf{a})$
3: Choose a modest value for μ, e.g. $\mu = 0.001$
4: Calculate $\delta\mathbf{a}$ by (3.11)
5: **if** $\chi(\mathbf{a} + \delta\mathbf{a}) \geq \chi(\mathbf{a})$ **then**
6: $\mu = 10\mu$
7: **else**
8: $\mu = \frac{1}{10}\mu$
9: **end if**
10: **end for**
11: **return** $\mathbf{a}^*$

A worst-case runtime estimate of algorithm 1 is given as $O\big((n\,|J|^2 + |J|^3)\,K\big)$, since evaluating a parameter set in line 2 takes time $O(n\,|J|)$, and since the execution of line 4 takes time $O(n|J|^2 + |J|^3)$ on its own, namely by calculating the second partial derivative matrix $\mathbf{A}$ in time $O(n|J|^2)$ and calculating the inverse matrix $\mathbf{A}^{-1}$ (3.4) in time $O(|J|^3)$ (Cormen *et al.*, 2001).

In summary, LM finds the set of minimizing parameters $\mathbf{a}^*$ by continuously switching between the methods of *Gauss-Newton* and *steepest descent*, based on the squared error of all datapoints (3.1). Close to a minimum, the former is emphasized, and for those far away the latter is predominantly applied. In general, LM results in a local minimum as a result of its *hill-climbing* nature.

3.2 The Lorentz function

In Chapter 2, we have seen that the elementary NMR signal after Fourier transformation is given as a Lorentz function with particular parameters, reflecting the relaxation process of a particular group of nuclei. In the following, some basic properties of the Lorentz function are given.

A single Lorentz function $Y(\omega)$ is given as

$$Y(\omega) = A\frac{\lambda}{\lambda^2 + (\omega - \omega_0)^2}, \tag{3.12}$$

where ω_0 stands for the position of the maximum, and A is a scaling factor (see figure 3.2(a)). λ stands for the half width at half height (HWHH), as shown in the following:

Let ω_λ be the frequency, for which $Y(\omega_\lambda) = \frac{1}{2}Y(\omega_0)$ holds. It follows that

$$\begin{aligned} Y(\omega_\lambda) &= \frac{1}{2}Y(\omega_0), \\ \Leftrightarrow A\frac{\lambda}{\lambda^2 + (\omega_\lambda - \omega_0)^2} &= A\frac{1}{2\lambda} \\ \Leftrightarrow \lambda^2 &= (\omega_\lambda - \omega_0)^2 \\ \Leftrightarrow \lambda = \omega_\lambda - \omega_0 &\quad \vee \quad \lambda = \omega_0 - \omega_\lambda. \end{aligned}$$

The maximum of a Lorentz function $Y_j(\omega)$ is given by $Y_j(\omega_j) = \frac{A_j}{\lambda_j}$, and the first three derivatives of Y are given as

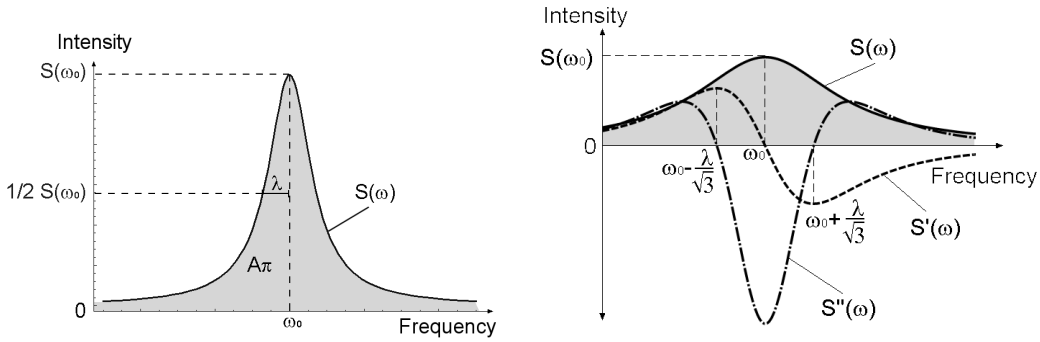
$$\begin{aligned} Y'(\omega) &= -\frac{2A\lambda(\omega - \omega_0)}{(\lambda^2 + (\omega - \omega_0)^2)^2} \\ Y''(\omega) &= \frac{8A\lambda(\omega - \omega_0)^2}{(\lambda^2 + (\omega - \omega_0)^2)^3} - \frac{2A\lambda}{(\lambda^2 + (\omega - \omega_0)^2)^2}, \\ \text{and } Y'''(\omega) &= \frac{-48\,A\,\lambda\,(\omega - \omega_0)^3}{\left(\lambda^2 + (\omega - \omega_0)^2\right)^4} + \frac{24\,A\,\lambda\,(\omega - \omega_0)}{\left(\lambda^2 + (\omega - \omega_0)^2\right)^3}, \end{aligned} \tag{3.13}$$

and the corresponding roots of the derivatives are given as

$$\begin{aligned} Y'(\omega) = 0 &\Leftrightarrow \omega = \omega_0, \\ Y''(\omega) = 0 &\Leftrightarrow \omega = \omega_0 \pm \frac{1}{\sqrt{3}}\lambda, \\ \text{and } Y'''(\omega) = 0 &\Leftrightarrow \omega = \omega_0 \vee \omega = \omega_0 \pm \lambda. \end{aligned} \tag{3.14}$$

Further, with $Y^{(-1)}$ denoting the first order integral function of Y, given as

$$Y^{(-1)}(\omega) = A \arctan\left(\frac{\omega - \omega_0}{\lambda}\right),$$



(a) Parameters of a Lorentz function

(b) A Lorentz function and its first two derivatives

Figure 3.2: The Lorentz function as the spectral component of an NMR spectrum. a) The parameters of a Lorentz function: ω_0 stands for the spectral position of the maximum, λ stands for the *half width at half height* (HWHH), and the area under the curve is given by $A\pi$. b) The first (dashed line) and second (dot-dashed line) derivatives of a Lorentz function (solid line) and the corresponding roots.

the area of Y equals $A\pi$ as shown

$$\begin{aligned}\int_{-\infty}^{\infty} Y(\omega)d\omega &= \lim_{\substack{a\to\infty\\ b\to-\infty}} \left[Y^{(-1)}(\omega)\right]_b^a = \lim_{a\to\infty} Y^{(-1)}(a) - \lim_{b\to-\infty} Y^{(-1)}(b)\\ &= A \lim_{a\to\infty} \arctan\left(\frac{a-\omega_0}{\lambda}\right) - A \lim_{b\to-\infty} \arctan\left(\frac{b-\omega_0}{\lambda}\right)\\ &= A \lim_{a\to\infty} \arctan(a) - A \lim_{b\to-\infty} \arctan(b)\\ &= A\frac{\pi}{2} + A\frac{\pi}{2} = A\pi.\end{aligned}$$

Figure 3.2(a) illustrates a Lorentz functions and its three parameters.

3.3 Key considerations

An NMR spectrum can be described as a superposition of Lorentz functions with unknown parameters. Regarding an automated approach for feature extraction of NMR data, several problems arise making it difficult to accurately extract the information and therefore gain valuable knowledge out of a series of NMR experiments.

A correct identification and separation of signal peaks from noise and artifacts plays a key-role to successful approximation. Following Bretthorst *et al.* (2005,

p.67), model selection is the crucial part of any approach operating in the time domain, since "Proposals for the model indicator are more difficult because when the model indicator changes, all of the parameters change". Furthermore, it is stated that "a change (is proposed) in the model indicator by increasing or decreasing the model indicator randomly using a Gaussian random number generator". Solving the task of model selection in the frequency domain probably has more potential, since signals occurring at different response frequencies become intuitively distinguishable from one another. Thus, this thesis focuses on methods operating in the frequency domain.

The following subsections discuss general problems in the context of frequency domain feature extraction. They are:

1. Data Complexity
2. Peak Overlap
3. Sample dependence of chemical shifts

3.3.1 Data Complexity

A sufficiently high spectral resolution is needed to minimize the effects of digitization. For example, figure 3.3 illustrates the drawbacks of poor resolution. The trivial approach of graphically determining the parameters of a Lorentz function clearly leads to highly falsified height, width, area and position information for the poorly resolved spectrum (solid line). With increasing the resolution (dashed line), the actual line shape is represented more accurately, and the parameters can be found more accurately as well. As a result, the number of datapoints in a single NMR spectrum is typically in the range of $10^4 - 10^6$, whereas in opposition the number of Lorentz functions in an NMR spectrum of heterogeneous biological samples is typically less than 10^3. Thus, data reduction has the potential to reduce computational efforts in further steps of the analysis, and hence plays an important role in NMR data analysis.

3.3.2 Distortions and Peak Overlap

As already mentioned in Section 2.2, the spectrum is commonly distorted in many ways, but even in the ideal case of (2.12) the number of local maxima does not

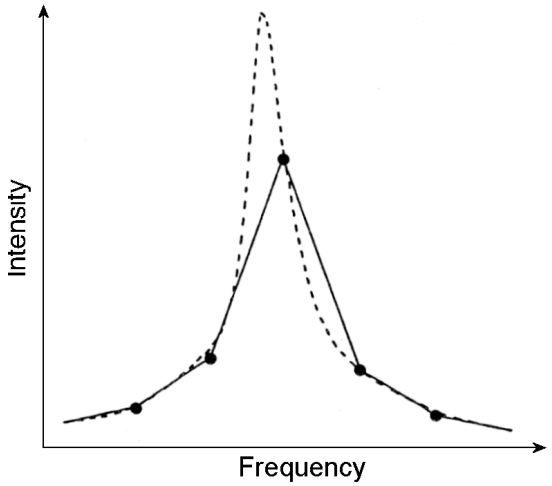


Figure 3.3: ^{1}H NMR signal, acquired with 32 K (dashed line) and with 2 K (solid line) datapoints (Friebolin, 1999).

necessarily equal to the number of single Lorentz functions. The reason for this is given by the occurrence of peak overlap, which will be discussed further in the remainder of this subsection.

Definition 3.2 (Peak Distance)
The peak distance $d(Y_i, Y_j)$ between two Lorentz functions Y_i and Y_j is defined by the absolute distance in their respective position parameters ω_i and ω_j as

$$d(Y_i, Y_j) = |\omega_j - \omega_i|$$

In conjunction, the distance between a Lorentz function Y_i and an arbitrary position $v \in \mathbb{R}$ is defined as

$$d(Y_i, v) = |v - \omega_i|$$

The distance between two arbitrary positions v_1, v_2 is given as

$$d(v_1, v_2) = |v_2 - v_1|$$

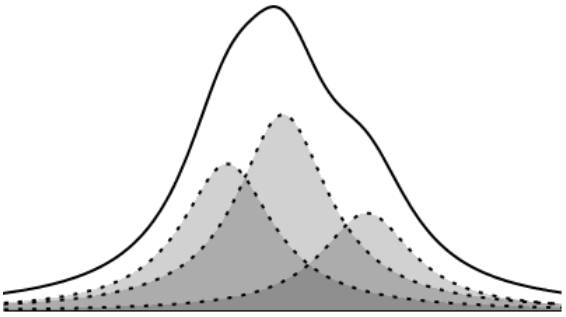


Figure 3.4: An example overlapping of three Lorentzians (dashed lines), resulting in a single maximum and two "shoulders" of the spectrum (solid line).

Definition 3.3 (Nearest Maximum)
Let $\{m_1, \ldots, m_Q\} \subset \mathbf{w}$ *be the position vector of local maxima in a discrete spectrum* S*. The nearest maximum* $m(Y_j)$ *to a given Lorentz function* Y_j *is defined as*

$$m(Y_j) = m(\omega_j) = \min_{l \in Q}\big(d(Y_j, m_l)\big) = \min_{l \in Q}\big(|m_l - \omega_j|\big).$$

with $l \in Q$ *denoting the index of the local maximum.*

Definition 3.4 (Hidden Peak)
Given a sum of Lorentz functions $Y = Y_1 + \cdots + Y_{|J|}$ *(2.12), and given an addend* Y_j *of* Y *at position* ω_j*, and given* $m(\omega_j)$ *denoting the position of the nearest maximum in* Y *to* Y_j*, a hidden peak is defined as*

$$\begin{aligned} &Y_j \text{ is hidden in } Y \Leftrightarrow \\ &\quad \exists i \in \{1, \ldots, |J|\} \Rightarrow m(Y_i) = m(Y_j) \wedge d(\omega_i, m(\omega_i)) \leq d(\omega_j, m(\omega_j)) \end{aligned}$$

Y_j *is then called a shoulder in* Y*, or simply a shoulder.*

In other words, a Lorentz function is *hidden*, if there exists another Lorentz function with the same nearest local maximum (Definition 3.3) in the spectrum but with lesser distance. In the case of $d(Y_i, m_i) = d(Y_j, m_i)$, both Y_i and Y_j might be intuitively considered as *hidden* and *unhidden* at the same time, but this is for real-world experiments rarely the case due to discretization and is therefore neglected in further considerations. In the following, a peak which is not *hidden* in Y is called a *maximum* peak. Figure 3.4 provides an example overlap situation for two *hidden* peaks and one *maximum* peak.

Proposition 3.1

Given two equally scaled and shaped Lorentz functions $Y_1(x)$ *and* $Y_2(x)$ *as*

$$Y_1(x) = A\frac{\lambda}{\lambda^2 + (x-\omega_1)^2}$$
$$Y_2(x) = A\frac{\lambda}{\lambda^2 + (x-\omega_2)^2},$$

with $x \in \mathbb{R}$*, and given their sum* $Y(x) = Y_1(x) + Y_2(x)$*, the total number of local optima in* Y *equals 1 for*

$$|\omega_1 - \omega_2| \leq \frac{2}{\sqrt{3}}\lambda$$

Proof:

The positions of the local optima in Y *equal the zero positions of the first derivative. They can be found by solving the following equation for* x*:*

$$Y'(x) = -2A\lambda\left(\frac{(x-\omega_1)}{\left(\lambda^2 + (x-\omega_1)^2\right)^2} + \frac{(x-\omega_2)}{\left(\lambda^2 + (x-\omega_2)^2\right)^2}\right) = 0,$$

resulting in three solutions x_1, x_2 *and* x_3*, given as*

$$x_1 = \frac{\omega_1 + \omega_2}{2}$$
$$\vee\, x_2 = x_1 \pm \frac{1}{2}\sqrt{-4\,\lambda^2 + (\omega_2 - \omega_1)\left(\omega_1 - \omega_2 + 2\sqrt{4\,\lambda^2 + (\omega_1 - \omega_2)^2}\right)}$$
$$\vee\, x_3 = x_1 \pm \frac{1}{2}\sqrt{-4\,\lambda^2 + (\omega_1 - \omega_2)\left(\omega_2 - \omega_1 + 2\sqrt{4\,\lambda^2 + (\omega_2 - \omega_1)^2}\right)}$$

x_1 *stands for the position of the local optimum of* Y *in the middle of* ω_1 *and* ω_2*, which either is a minimum in the separated case, or the single maximum in the overlapped case.* x_2 *and* x_3 *stand for the two maxima in the separated case for* $\omega_1 > \omega_2$ *and* $\omega_1 < \omega_2$*, respectively. Their root terms equal zero for*

$$\omega_1 = \omega_2 + \frac{2}{\sqrt{3}}\,\lambda \quad \vee \quad \omega_1 = \omega_2 - \frac{2}{\sqrt{3}}\,\lambda$$

which concludes that only one maximum occurs in Y if this particular relationship between the positions ω_1, ω_2 and the width parameter λ applies. There are two results for ω_1, distinguishing between $\omega_1 > \omega_2$ and $\omega_1 < \omega_2$ again.

□

Proposition 3.2

Given two equally scaled and shaped Lorentz functions Y_1, Y_2, and given their sum Y as in proposition 3.1, and let w.l.o.g. be $\omega_1 < \omega_2$, then it holds that

$$\omega_2 - \omega_1 < \frac{2\lambda}{\sqrt{3}} \;\Rightarrow\; Y'' \textit{has maximal two roots.}$$

Proof:
Given any Lorentz function Y_0 as

$$Y_0(\omega) = A\frac{\lambda}{\lambda^2 + (\omega - \omega_0)^2},$$

the corresponding second derivative Y_0'' of Y_0 has two roots, i.e. $\omega = \omega_0 \pm \frac{1}{\sqrt{3}}\lambda$ (3.14), and is negative only in the interval $[\omega_0 - \frac{1}{\sqrt{3}}\lambda, \omega_0 + \frac{1}{\sqrt{3}}\lambda]$ due to

$$\begin{aligned}
& & Y_0''(\omega) &< 0 \\
& & \frac{8\,A\lambda(\omega-\omega_0)^2}{(\lambda^2+(\omega-\omega_0)^2)^3} - \frac{2\,A\lambda}{(\lambda^2+(\omega-\omega_0)^2)^2} &< 0 \\
\Leftrightarrow & & \frac{2A\lambda}{(\lambda^2+(\omega-\omega_0)^2)^2}\left(\frac{4(\omega-\omega_0)^2}{\lambda^2+(\omega-\omega_0)^2} - 1\right) &< 0 \\
\Leftrightarrow & & \frac{4(\omega-\omega_0)^2}{\lambda^2+(\omega-\omega_0)^2} &< 1 \\
\Leftrightarrow & & 3(\omega-\omega_0)^2 &< \lambda^2 \\
\Leftrightarrow & & \omega > \omega_0 - \frac{1}{\sqrt{3}}\,\lambda \quad \wedge \quad \omega &< \omega_0 + \frac{1}{\sqrt{3}}\,\lambda
\end{aligned}$$

In consequence, Y_0'' is positive for

$$\omega < \omega_0 - \frac{1}{\sqrt{3}}\,\lambda \quad \wedge \quad \omega > \omega_0 + \frac{1}{\sqrt{3}}\,\lambda$$

To prove that the summation of two equally shaped and scaled Lorentz functions Y_1, Y_2 with $d(Y_1, Y_2) \leq \frac{2}{\sqrt{3}}\lambda$ result in a sum Y, of which the second derivative Y''

contains at most two roots, we focus w.l.o.g. at and around the position $\omega_0 - \frac{\lambda}{\sqrt{3}}$. *Let for this purpose* y_1 *be given as*

$$\begin{aligned} y_1 &= Y''(\omega_0 - \frac{\lambda}{\sqrt{3}} + c) \\ &= \frac{8\,A\lambda(c - \frac{\lambda}{\sqrt{3}})^2}{(\lambda^2 + (c - \frac{\lambda}{\sqrt{3}})^2)^3} - \frac{2\,A\lambda}{(\lambda^2 + (c - \frac{\lambda}{\sqrt{3}})^2)^2} \\ &= \frac{54\,A\lambda\,c\,\overbrace{(3c - 2\sqrt{3}\lambda)}^{\alpha}}{\big(4\lambda^2 + c\,\underbrace{(3c - 2\sqrt{3}\lambda)}_{\alpha}\big)^3}, \end{aligned}$$

with $c \in \mathbb{R}$, $c < 0$, *and let* y_2 *be given as*

$$\begin{aligned} y_2 &= Y''(\omega_0 - \frac{\lambda}{\sqrt{3}} + d) \\ &= \frac{8\,A\lambda(d - \frac{\lambda}{\sqrt{3}})^2}{(\lambda^2 + (d - \frac{\lambda}{\sqrt{3}})^2)^3} - \frac{2\,A\lambda}{(\lambda^2 + (d - \frac{\lambda}{\sqrt{3}})^2)^2} \\ &= \frac{54\,A\lambda\,d\,\overbrace{(3d - 2\sqrt{3}\lambda)}^{\beta}}{\big(4\lambda^2 + d\,\underbrace{(3d - 2\sqrt{3}\lambda)}_{\beta}\big)^3}, \end{aligned}$$

with $d \in \mathbb{R}$, $d > 0$. *By noting that* Y_0'' *is axis-symmetric in* ω_0, *since the third integral function* arctan *is rotation-symmetric (without proof), it is sufficient to show that the following holds:*

$$c < 0 < d \leq \frac{\lambda}{\sqrt{3}} \wedge |c| = |d| \quad \Rightarrow \qquad |y_1| < |y_2|$$

For $\lambda, A > 0$, the proof follows as

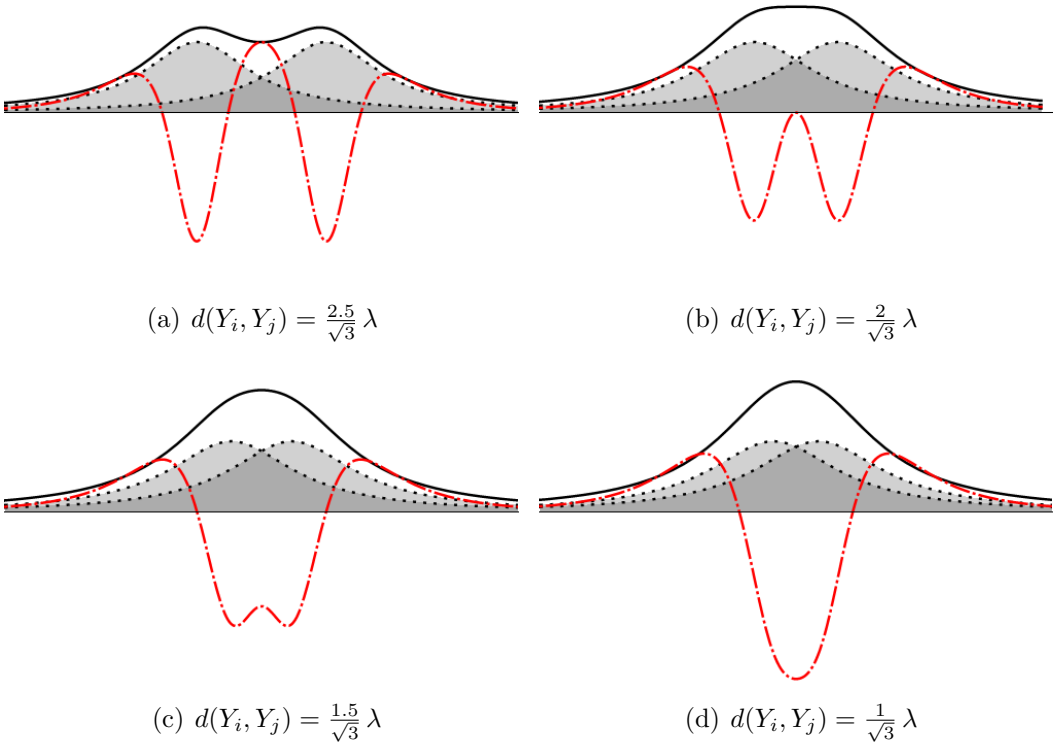
$$
\begin{aligned}
& & c < 0 < d \leq \frac{\lambda}{\sqrt{3}} & & \Rightarrow & & \alpha < 0 & \wedge & \beta < 0 \\
\wedge & \quad c < 0 < d \leq \frac{\lambda}{\sqrt{3}} & \wedge \quad |c| = |d| & & \Rightarrow & & |\alpha| > |\beta| & \wedge & |c\,\alpha| > |d\,\beta| \\
\wedge & & c < 0 & \wedge \quad \alpha < 0 & \Rightarrow & & c\alpha > 0 & & \\
\wedge & \quad 0 < d \leq \frac{\lambda}{\sqrt{3}} & \wedge \quad \beta < 0 & & \Rightarrow & & d\beta < 0 & \wedge & |d\beta| < 4\lambda^2 \\
& & & & \Rightarrow & & 4\lambda^2 + c\,\alpha & > & 4\lambda^2 + d\,\beta \\
& & & & \Rightarrow & & \frac{c\alpha}{(4\lambda^2 + c\alpha)^3} & < & \frac{d\beta}{(4\lambda^2 + d\beta)^3} \\
& & & & \Rightarrow & & |y_1| & < & |y_2|
\end{aligned}
$$

□

In summary, peak overlap between a pair of equally scaled and shaped Lorentz functions results in the loss of a local maximum in their sum for the distance $d(Y_1, Y_2) \leq \frac{2}{\sqrt{3}}\lambda$, and furthermore, the number of roots in the second derivative Y'' of the sum Y is two for $d(Y_1, Y_2) < \frac{2}{\sqrt{3}}\lambda$. Figures 3.5(a) - 3.5(d) show example overlapping scenarios of Lorentz functions, their resulting sum and the corresponding first and second derivatives. One clearly observes, that the minima of the second derivative function (red dot-dashed line) are much better preserved than the maxima of the spectrum (black solid line) (compare figures 3.5(a), 3.5(b) and 3.5(c)). Based on this observation, Chapter 5 proposes an alternative method for the identification of peak overlap.

3.3.3 Sample dependence of chemical shifts

With respect to the analysis of a whole series of spectra resulting from multiple NMR experiments, an additional drawback is given by the sample dependence of chemical shifts (compare Section 2.6). As well as to the type and the atomic neighbourhood of a nucleus, the chemical shift also depends on the temperature, the pH-value and the concentrations of the solutions. These conditions generally vary from experiment to experiment, as does the chemical shift. As a result, Lorentz functions of the same molecular source may occur at different positions along the



(a) $d(Y_i, Y_j) = \frac{2.5}{\sqrt{3}}\,\lambda$ (b) $d(Y_i, Y_j) = \frac{2}{\sqrt{3}}\,\lambda$

(c) $d(Y_i, Y_j) = \frac{1.5}{\sqrt{3}}\,\lambda$ (d) $d(Y_i, Y_j) = \frac{1}{\sqrt{3}}\,\lambda$

Figure 3.5: Two overlapping Lorentz functions Y_i, Y_j(grey areas) with equal HWHH λ and area A, the corresponding sum (solid line) and the resulting second derivative (dot-dashed red line), with distances varying as indicated.

frequency axis for different experiments, and a direct point-by-point comparison of multiple NMR spectra is thus prohibited.

For a series of NMR spectra $S_1(\omega), \ldots, S_K(\omega)$ containing an equal set of spectral components J, equation (2.12) can be extended as

$$S_k(\omega) = \sum_j^{|J|} A_j \frac{\lambda_j}{\lambda_j^2 + (\omega - \omega_{j,k})^2} = \sum_j^{|J|} Y_{j,k}(\omega),$$

$$\text{with} \quad \omega_{j,k} = \omega_j + c_{j,k}$$

denoting the particular frequency variation of component j in experiment k. This leads to the two-dimensional $|J| \times K$ matrix $\mathbf{C}$ containing the variation terms as

$$\begin{matrix} c_{1,1} & \cdots & c_{1,k} & \cdots & c_{1,K} \\ \vdots & & \vdots & & \vdots \\ c_{j,1} & \cdots & c_{j,k} & \cdots & c_{j,K} \\ \vdots & & \vdots & & \vdots \\ c_{|J|,1} & \cdots & c_{|J|,k} & \cdots & c_{|J|,K} \end{matrix}$$

As already mentioned in Section 2.6, the actual values of $\mathbf{C}$ cannot be calculated in advance. In this context, it is worth mentioning that the frequency distribution within multiplets (see Section 2.9), namely the peak-to-peak distances in the frequency domain, is given by the component-specific *coupling constants*, which themselves only depend on the external magnetic field. Thus, chemical shift variations occur for all multiplets only as a whole, more formally described as

$$c_{p,k} = c_{q,k} \quad \text{for all} \quad p, q \in M, k \in K$$

with $M \subseteq J$ denoting a multiplet, and K denoting the spectrum set. In this context, a method for the identification of multiplets is proposed in Chapter 6.

Chapter 4

Approach I: Lorentzian Peak Reconstruction

4.1 The Exact Solution

For a spectrum S containing a single peak ($|J|$ equals 1), it holds $Y(\omega) = S(\omega)$. By providing three points of the function $(\omega_1, Y(\omega_1))$, $(\omega_2, Y(\omega_2))$ and $(\omega_3, Y(\omega_3))$ (see 4.1), a quadratic equation system can be obtained as

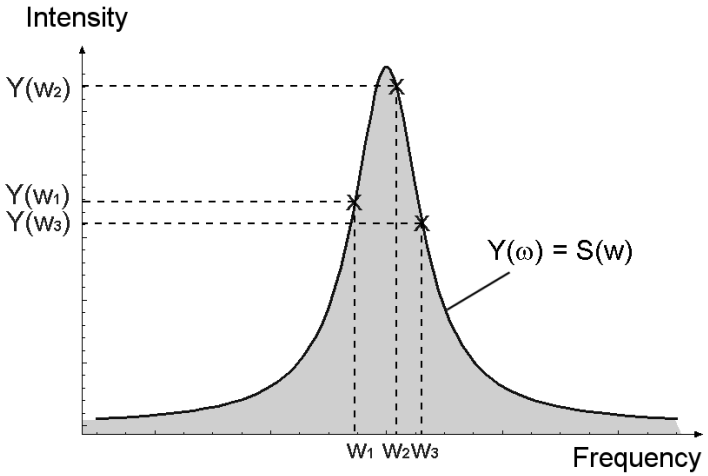


Figure 4.1: Three points are chosen to directly calculate the parameters of the Lorentz function.

$$\begin{aligned} & Y(\omega_1) = A\frac{\lambda}{\lambda^2 + (\omega_1 - \omega)^2} \\ \wedge \quad & Y(\omega_2) = A\frac{\lambda}{\lambda^2 + (\omega_2 - \omega)^2} \\ \wedge \quad & Y(\omega_3) = A\frac{\lambda}{\lambda^2 + (\omega_3 - \omega)^2}, \end{aligned} \tag{4.1}$$

With denoting $Y(\omega_1) = y_1$, $Y(\omega_2) = y_2$ and $Y(\omega_3) = y_3$, the parameter solutions for A, λ and ω are found by the software *Mathematica 5.0, Wolfram Research* as follows:

$$\begin{aligned} \omega &= \frac{{\omega_1}^2\, y_1\, y_{2,3} + {\omega_3}^2\, y_{1,2}\, y_3 + {\omega_2}^2\, y_2\ (-y_{1,3})}{2\,\omega_{1,2}\, y_1\, y_2 - 2\ (\omega_{1,3}\, y_1 + (-\omega_{2,3})\ y_2)\ y_3}, \\ \lambda &= \frac{1}{\sqrt{y_{2,3}}}\sqrt{{\omega_3}^2\, y_3 + \frac{\alpha}{4\,(\omega_1\, y_1\, y_{2,3} + \omega_3\, y_{1,2}\, y_3 + \omega_2\, y_2\ (-y_{1,3}))^2}}, \\ A &= \frac{-4\,\omega_{1,2}\omega_{1,3}\,\omega_{2,3}\, y_1\, y_2\, y_3\ (\omega_1\, y_1\, y_{2,3} + \omega_3\, y_{1,2}\, y_3 + \omega_2\, y_2\ (-y_{1,3}))\,\lambda}{\left(\omega_{1,2}^4\, {y_1}^2\, {y_2}^2 - 2\,\omega_{1,2}^2\, y_1\, y_2\ \left(\omega_{1,3}^2\, y_1 + \omega_{2,3}^2\, y_2\right)\ y_3 + \left(\omega_{1,3}^2\, y_1 - \omega_{2,3}^2\, y_2\right)^2\, {y_3}^2\right)}, \end{aligned}$$

where

$$\begin{aligned} \alpha = & -\left(\omega_{1,2}^4\, {y_1}^2\, {y_2}^3\right) + \omega_{1,2}^2\, y_1\, {y_2}^2\, \beta\, y_3 - y_2\, \gamma\, {y_3}^2 \\ & + ((\omega_1 - 3\,\omega_3)\ \omega_{1,3}\, y_1 - (\omega_2 - 3\,\omega_3)\ \omega_{2,3}\, y_2)\ \left({\omega_1}^2\, y_1 - {\omega_2}^2\, y_2 + {\omega_3}^2\ - y_{1,2}\right)\ {y_3}^3, \\ \beta = & \left(3\,{\omega_1}^2 + {\omega_2}^2 - 2\,{\omega_3}^2 - 2\,\omega_1\ (\omega_2 + 2\,\omega_3)\right)\ y_1 + 2\,\omega_{2,3}^2\, y_2, \\ \gamma = & \ \omega_{1,3}\left(3\,{\omega_1}^3 - {\omega_1}^2\ (4\,\omega_2 + 5\,\omega_3) + \omega_1\ \left(2\,{\omega_2}^2 + 4\,\omega_2\,\omega_3 - 5\,{\omega_3}^2\right)\right. \\ & \left. - \omega_3\ \left(2\,{\omega_2}^2 - 8\,\omega_2\,\omega_3 + {\omega_3}^2\right)\right)\ {y_1}^2 \\ & + 2\ (\omega_2 - \omega_3)\ \left({\omega_2}^2\ (-2\,\omega_1 + \omega_2) + 4\,\omega_{1,2}\,\omega_2\,\omega_3 + (2\,\omega_1 - 5\,\omega_2)\ {\omega_3}^2 + {\omega_3}^3\right)\ y_1\, y_2 \\ & + \omega_{2,3}^4\, {y_2}^2, \\ \omega_{1,2} = \omega_1 - \omega_2, \quad & \omega_{1,3} = \omega_1 - \omega_3, \quad \omega_{2,3} = \omega_2 - \omega_3, \\ y_{1,2} = y_1 - y_2, \quad & y_{2,3} = y_2 - y_3, \text{ and } y_{1,3} = y_1 - y_3. \end{aligned}$$

The equations are well defined for

$$\omega_1 < \omega_2 < \omega_3 \quad \wedge \quad y_1 < y_2 > y_3 \quad \wedge \quad y_1, y_2, y_3 > 0 \tag{4.2}$$

and for further conditions describing the general relationships between the given points expressible by a single Lorentz function:

$$
\begin{aligned}
&\left(y_2 < \frac{(\omega_1-\omega_3)^2 y_1}{(\omega_2-\omega_3)^2} \;\vee\; \frac{(\omega_1-\omega_3)^2 y_1}{(\omega_2-\omega_3)^2} < y_2 \leq \frac{(-2\,\omega_1+\omega_2+\omega_3)^2 y_1}{(\omega_2-\omega_3)^2}\right) \\
&\wedge\, y_3 > \frac{(\omega_1-\omega_2)^2 y_1 y_2 \left((\omega_1-\omega_3)^2 y_1 + (\omega_2-\omega_3)^2 y_2\right)}{\left((\omega_1-\omega_3)^2 y_1 - (\omega_2-\omega_3)^2 y_2\right)^2} \\
&\qquad - 2\sqrt{\frac{(\omega_1-\omega_2)^4 (\omega_1-\omega_3)^2 (\omega_2-\omega_3)^2 {y_1}^3 {y_2}^3}{\left((\omega_1-\omega_3)^2 y_1 - (\omega_2-\omega_3)^2 y_2\right)^4}} \\
\bigvee\quad & y_2 = \frac{(\omega_1-\omega_3)^2 y_1}{(\omega_2-\omega_3)^2} \;\wedge\; y_3 > \frac{(\omega_1-\omega_2)^2 y_1 y_2}{2\,(\omega_1-\omega_3)^2 y_1 + 2\,(\omega_2-\omega_3)^2 y_2} \\
\bigvee\quad & y_2 > \frac{(-2\,\omega_1+\omega_2+\omega_3)^2 y_1}{(\omega_2-\omega_3)^2} \\
&\wedge\, y_3 < 2\sqrt{\frac{(\omega_1-\omega_2)^4 (\omega_1-\omega_3)^2 (\omega_2-\omega_3)^2 {y_1}^3 {y_2}^3}{\left((\omega_1-\omega_3)^2 y_1 - (\omega_2-\omega_3)^2 y_2\right)^4}} \\
&\qquad - \frac{(\omega_1-\omega_2)^2 y_1 y_2 \left((\omega_1-\omega_3)^2 y_1 + (\omega_2-\omega_3)^2 y_2\right)}{\left((\omega_1-\omega_3)^2 y_1 - (\omega_2-\omega_3)^2 y_2\right)^2}.
\end{aligned}
$$

As a consequence, the Lorentzian parameters can directly be calculated by providing any triplet of points, as long as the conditions are met. Note that the middle value of a given triplet of points needs to be larger than the side ones (4.2), but does not necessarily equal the maximum of the underlying Lorentz function due to the effects of discretization.

4.2 Proportional Approximation I

In order to directly calculate the parameters based on the solutions of the previous section, position triplets are found by the occurrence of local maxima in a given spectrum. For a peak j, they are in the following denoted as $\{w_{j,left},\ w_{j,max},\ w_{j,right}\}$, and the respective positions are – as an initial solution – found with $w_{j,max}$ as the maximum position and $w_{j,left}$ and $w_{j,right}$ as the respective next left and right neighboring positions of a spectral component j. In this way, the maximum constraint (4.2) is automatically preserved.

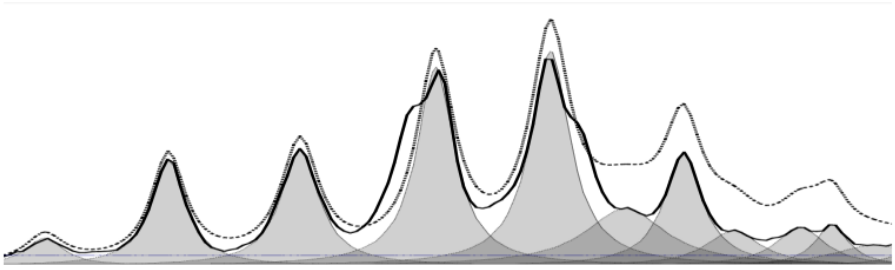


Figure 4.2: Initially found Peaks by direct calculation. Shown are the spectrum (solid curve), the initially found Lorentz functions (grey areas), a predefined height threshold (horizontal line) and the corresponding superposition (dotted line).

A real-world spectrum is commonly given as a superposition of multiple Lorentz functions (2.12), and the actual intensity values of each Lorentz-function remain hidden. The assumption of $Y_j(w) = S(w)\ \forall j \in J$ results in a parameter set, for which each $Y_j(w)$ on its own matches the given spectrum at each of the chosen positions $w_{j,left}$, $w_{j,max}$ and $w_{j,right}$, but for which the sum of all components in (2.12) will not necessarily match the spectrum $S(w)$. Figures 4.2 and 4.3(a) elucidate this situation. The spectrum is displayed by the black solid line, each component $Y_j(w)$ is marked in grey, and the dashed line shows the resulting superimposed signal of each component. It can be observed, that this has almost no effect if the peaks are far enough from each other (left side of figure 4.2), but also leads to substantial errors for peaks found close to one another (right side of figure 4.2, figure 4.3(a)).

Nevertheless, this initial guess can be used as a starting point. The basic idea to improve the initial parameter set is to iteratively adjust the values $Y_j(w)$ by rule of proportion. Let $\widehat{Y}_j^{[i]}$ denote the model of Lorentz function Y_j and let $\widehat{Y}^{[i]}$ be the current model spectrum, namely the sum of all modeled Lorentz functions (the dashed line in figures 4.2, 4.3(a)) at iteration step i, and let $w_{j,x}$ denote any of the chosen position triplet $\{w_{j,left},\ w_{j,max},\ w_{j,right}\}$. The key idea is to take the ratio between the original spectrum $S(w_{j,x})$ and the current model $\widehat{Y}(w_{j,x})$ as the ratio, by which the current intensity value $Y_j^{[i]}(w_{j,x})$ has to be decreased or increased (figure 4.3(b)) in order to produce a more accurate fit (figure 4.3(c)). The corresponding formula is written as

$$\frac{\widehat{Y}_j^{[i]}(w_{j,x})}{\widehat{Y}_j^{[i-1]}(w_{j,x})} = \frac{S(w_{j,x})}{\widehat{Y}^{[i-1]}(w_{j,x})} \Leftrightarrow \widehat{Y}_j^{[i]}(w_{j,x}) = \widehat{Y}_j^{[i-1]}(w_{j,x}) \frac{S(w_{j,x})}{\sum_{l \in J} \widehat{Y}_l^{[i-1]}(w_{j,x})} \tag{4.3}$$

4.3 Shoulder Detection

For a real-world spectrum, not only the actual parameters of the Lorentz functions are unknown, but also the number of functions to start with. A trivial way to solve this problem is to consider each local maximum exceeding a certain intensity threshold as a potential peak. However, *hidden* peaks as defined in Definition 3.4 are neglected (see Definition 3.4 in Section 3.3.2 of the previous chapter). A straight forward solution to also detect "shoulders" is within this chapter given by iteratively subtracting the model from the original spectrum after approximation of the current parameter set, and to extend the model by the remaining local maxima (figures 4.3(d) - 4.3(f)). This is achieved by the iteration of the following three steps:

1. Find all local maxima exceeding intensity threshold r, calculate the corresponding lorentzian parameters given the analytical solutions of equation-system (4.1), and add them to the model.
2. Iteratively approximate the parameter set by the rule of proportion as given by (4.3).
3. Subtract the resulting model from the original spectrum, and proceed with Step 1.

Figure 4.3 provides a schematic illustration of the approximation.

4.4 The Algorithm

The corresponding algorithm *Lorentzian Peak Reconstruction* (LPR) is given by algorithm 2. The *best* model $\widehat{Y}^{best}$ is hereby defined as the one which minimizes the mean squared error at the peak points (MSEPP), defined as

$$\begin{aligned}\mathrm{MSEPP}(S,\widehat{Y}) = \frac{1}{3\,|J|}\sum_{j}^{|J|}\Big(\quad & (S(w_{j,left}) - \widehat{Y}(w_{j,left}))^2 \\ & + (S(w_{j,max}) - \widehat{Y}(w_{j,max}))^2 \\ & + (S(w_{j,right}) - \widehat{Y}(w_{j,right}))^2 \Big) \end{aligned} \tag{4.4}$$

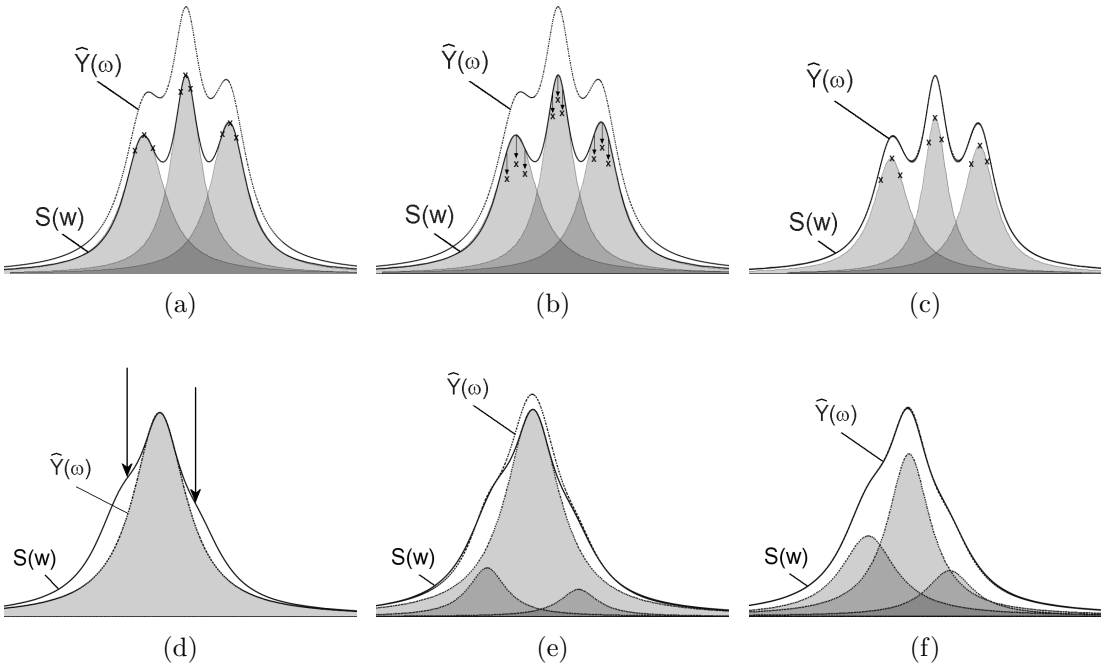


Figure 4.3: Parameter approximation and Shoulder Detection by rule of proportion. (a): Initial guess exceeds a superimposed spectrum. (b): Iteratively adjust the heights by rule of proportion (4.3) and recalculate the parameters. (c): Result after three iterates. (d): *Hidden* Peaks are found by iteratively subtracting the model from the spectrum. (e): Incorporating the new "shoulders" to the model by repeating the proportional approximation. (f): The final result.

and can be calculated in time $O(|J|^2)$. MSEPP considers the squared distance for the chosen three points of a peak only.

In the following, we will denote lines 4 - 13 as the inner loop, and lines 1 - 16 as the outer loop of algorithm 2. Each iteration of the outer loop needs time $O(n+K_2\,n\,|J|+|J|)$ with $|J|$ denoting the number of model peaks, K_2 denoting the predefined number of inner loop iterations, and n denoting the number of spectral datapoints of the spectrum. This runtime estimate concludes by observing that finding local maxima and calculating the corresponding lorentzian parameters in lines 2-3 takes time $O(n)$, that the inner loop can be performed in time $O(n|J|)$, and that line 15 can be executed in time $O(n|J|)$ as well.

In summary, lines 2 - 3 detect single peaks based on the occurrence of local maxima and a user-specified height threshold r, and lines 4 - 13 approximate the parameters in accordance with the spectrum. Hidden peaks are detected by iteratively subtracting the currently best model from the spectrum (line 15).

Algorithm 2 Lorentzian Peak Reconstruction (LPR)

`Input:` Spectrum S containing n data points, height threshold r, parameters K_1, K_2
`Output:` Spectrum model $\widehat{Y}$

1: **for** $a = 1$ to K_1 **do**
2: Find all local maximum positions $w_{j,max}$, for which $S(w_{j,max})$ >r holds.
3: Calculate lorentzian parameters using the solutions of equation system (4.1), given the point triplets $\{(w_{j,left}, S(w_{j,left})), (w_{j,max}, S(w_{j,max})), (w_{j,right}, S(w_{j,right}))\}$ and add to J.
4: **for** $b = 1$ to K_2 **do**
5: Calculate the model $\widehat{Y}^{[b]}$ for all $\{w_1, \ldots, w_n\} \in \mathbf{w}$.
6: **for all** peaks j in J **do**
7: Calculate new heights $\widehat{Y}_j^{[b]}(w_{j,x})$ by (4.3)
8: **if** $w_{j,max}$ is not local maximum in $\widehat{Y}_j^{[b]}$ anymore **then**
9: Find new maximum position $w_{j,max}$ within the range $[w_{j-1,right}, w_{j+1,left}]$ in $\widehat{Y}_j^{[b]}$
10: **end if**
11: Calculate new parameters ω_j, λ_j, A_j, given the new heights $\widehat{Y}^{[b]}(w_{j,x})$
12: **end for**
13: **end for**
14: Remove all peaks with $Y_j^{[b]}(w_{j,max}) \leq r$
15: Subtract the currently best model $\widehat{Y}^{best}$ from S
16: **end for**
17: **return** $\widehat{Y}^{best}$

4.5 Results

The performance of algorithm 2 is evaluated on four different spectra, one real-world spectrum and three simulated spectra. The simulated spectra contain each 100 Lorentz functions $Y_1, \ldots, Y_{100}$, with parameters A_j, λ_j and ω_j given as

$$\begin{aligned} A_j &= u_A, & u_A &\approx U(50, 100),\ j \in \{1, \ldots, 100\} \\ \lambda_j &= u_\lambda, & u_\lambda &\approx U(0.002, 0.005),\ j \in \{1, \ldots, 100\} \\ \omega_j &= \omega_{j-1} + u_\omega \cdot max(\lambda_{j-1}, \lambda_j), & u_\omega &\approx U(1.5, 5.0),\ j \in \{2, \ldots, 100\} \end{aligned}$$

u_A, u_λ and u_ω are uniformly distributed random variables. The position parameter of the first Lorentz function ω_1 has been chosen as 0.0. The spectral resolution is given as $w_{i+1} - w_i = 0.0005\ \forall i \in \{1, \ldots, n-1\}$, where n denotes the number of datapoints. The distance between two consecutive peak pairs Y_{j-1}, Y_j is given

by multiplying the greater of their respective HWHH parameters λ_{j-1}, λ_j with a random number, resulting in the first simulated spectrum Sim_A.

For noise simulation, two more spectra Sim_B and Sim_C are similarly generated, but with adding to each spectral datapoint an equally distributed random number v as

$$S(w) = \sum_{j=1}^{100} A_j \frac{\lambda_j}{\lambda_j^2 + (w - \omega_j)^2} + v, \tag{4.5}$$

$$\text{with} \quad v \approx U(0, v_{\max}) \quad \text{and} \quad v_{\max} = \frac{\max\limits_{i \in \{1,\ldots,n\}} S(w_i)}{\rho}.$$

ρ is reciprocal to the maximal distortion level $v_{\max}$, with a high value for ρ implying a low level of noise. In the following, ρ is referred to as the *signal-to-distortion ratio* (SDR). It holds $\rho = 1000$ for the second spectrum Sim_B, and $\rho = 100$ for the third spectrum Sim_C. Figure 4.4 shows an example simulated spectrum.

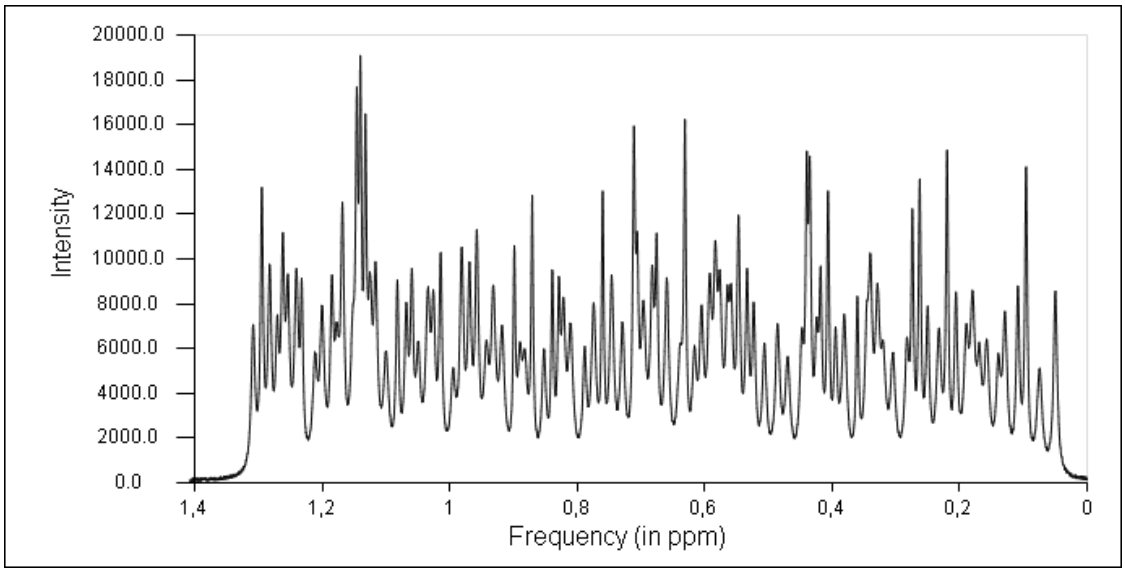


Figure 4.4: Example of a simulated spectrum.

All runs of algorithm 2 have been performed with the parameter settings $K_1 = 3$ (number of outer loops), $K_2 = 50$ (number of inner loops) and height thresholds $r = 0.02 \cdot \max$ for the simulations, and $r = 0.0005 \cdot \max$ for the real-world spectrum, where max denotes the maximal value of the respective data vector. For evaluation purposes, two distinct measures are used: *Mean squared error* (MSE) and *mean*

squared error at the peak points (MSEPP). MSE measures the mean squared error between the given spectrum S and the model $\widehat{Y}$ as

$$\mathrm{MSE}(S, \widehat{Y}) = \frac{1}{n} \sum_{w=1}^{n} (S(w) - \widehat{Y}(w))^2$$

MSEPP refers to the measurement function used in algorithm 2 (4.4). In order to monitor the squared error between the current model $\widehat{Y}$ and the actual superposition of Lorentz functions Y regardless to additional noise in the spectra, two analogous measures MSE Orig and MSEPP Orig are introduced as

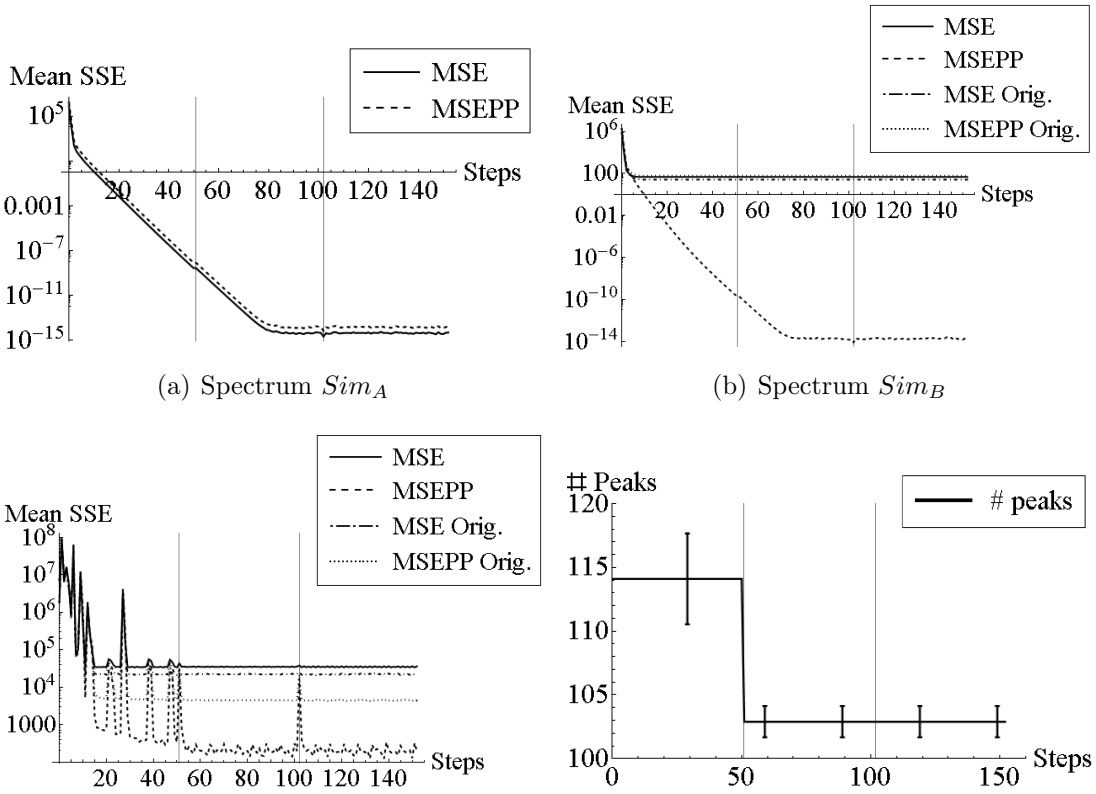
$$\mathrm{MSE\ Orig}(Y, \widehat{Y}) = \frac{1}{n} \sum_{w=1}^{n} (Y(w) - \widehat{Y}(w))^2,$$

$$\begin{aligned} \mathrm{MSEPP\ Orig}(Y, \widehat{Y}) = \frac{1}{3\,|J|} \sum_{j}^{|J|} \big(& (Y(w_{j,left}) - \widehat{Y}(w_{j,left}))^2 \\ & + (Y(w_{j,max}) - \widehat{Y}(w_{j,max}))^2 \\ & + (Y(w_{j,right}) - \widehat{Y}(w_{j,right}))^2 \big) \end{aligned}$$

Figure 4.5 shows the performance of algorithm 2 for both error functions MSE and MSEPP. Thereby, each iteration of the outer loop is marked by a vertical line, and a step on the horizontal axis reflects the accuracy of the current model after executing either lines 3 or 12 of algorithm 2. For spectrum Sim_A, remarkably both MSE and MSEPP decrease exponentially in the number of iterations from 10^8 straight down to 10^{-14} (figure 4.5(a)). This shows, that the proposed algorithm is capable of reconstructing a pure sum of Lorentz functions.

Figure 4.5(b) shows the performance for the simulated spectrum Sim_B. Again substantial error decrease can be observed within the first iterations, but only MSEPP shows similar behavior to that observed in the case of Sim_A. This can be reasoned by the fact, that the proportional approximation based on (4.3) relies on three peak-specific points only.

Increasing the level of noise has indeed a dramatic effect on the performance of algorithm 2, as can be observed in figure 4.5(c). Within the first 51 steps, namely the first iteration of the outer loop, highly unstable behavior is found. As can be observed in figure 4.5(d), a feasible explanation is given by observing, that the average number of 114 detected peaks highly exceeds the original number of 100 Lorentz functions. Although filtering the model by discarding peaks with smaller



(a) Spectrum Sim_A

(b) Spectrum Sim_B

(c) Spectrum Sim_C

(d) Number of peaks for spectrum Sim_C. The length of the errorbars equal twice the standard deviation out of ten spectra.

Figure 4.5: Simulation results of Algorithm 2

maximum values than the threshold r (line 14 of algorithm 2) results in a more stable development for all error measures, the original number of peaks is still exceeded on average by 3 - 4 peaks. Note, that for Sim_A and Sim_B the number of peaks was constantly 100 and is therefore not shown. The error at the peak points given by MSEPP and MSEPP Orig is below their respective counterparts MSE and MSE Orig, which shows the algorithm's ability to focus on the important parts of a spectrum, namely the peaks themselves. For all simulated spectra, the execution of all 152 steps needed between 30 and 40 seconds on average for ten spectra.

For demonstration purposes, the performance of algorithm 2 is also shown for an example of a real-world spectrum, namely a metabolite spectrum of human colon cell lines containing 32768 (2^{15}) datapoints. The spectrum is obtained on a VARIAN INOVA 800 spectrometer operating at 799.77 MHz. The data was acquired with a 13 KHz spectral width, 22114 data points, and 1.7 second acquisition time. Zero filling was performed resulting in 32768 (2^{15}) data points, and the FFT

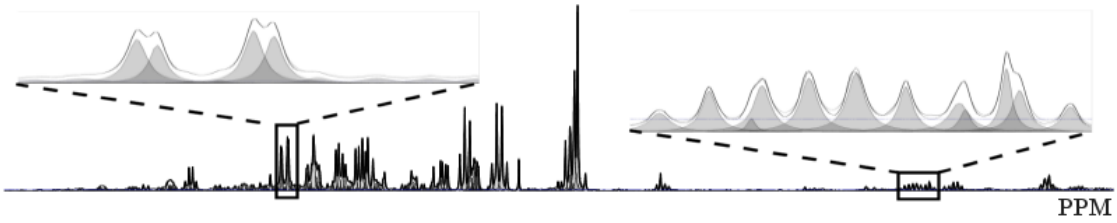


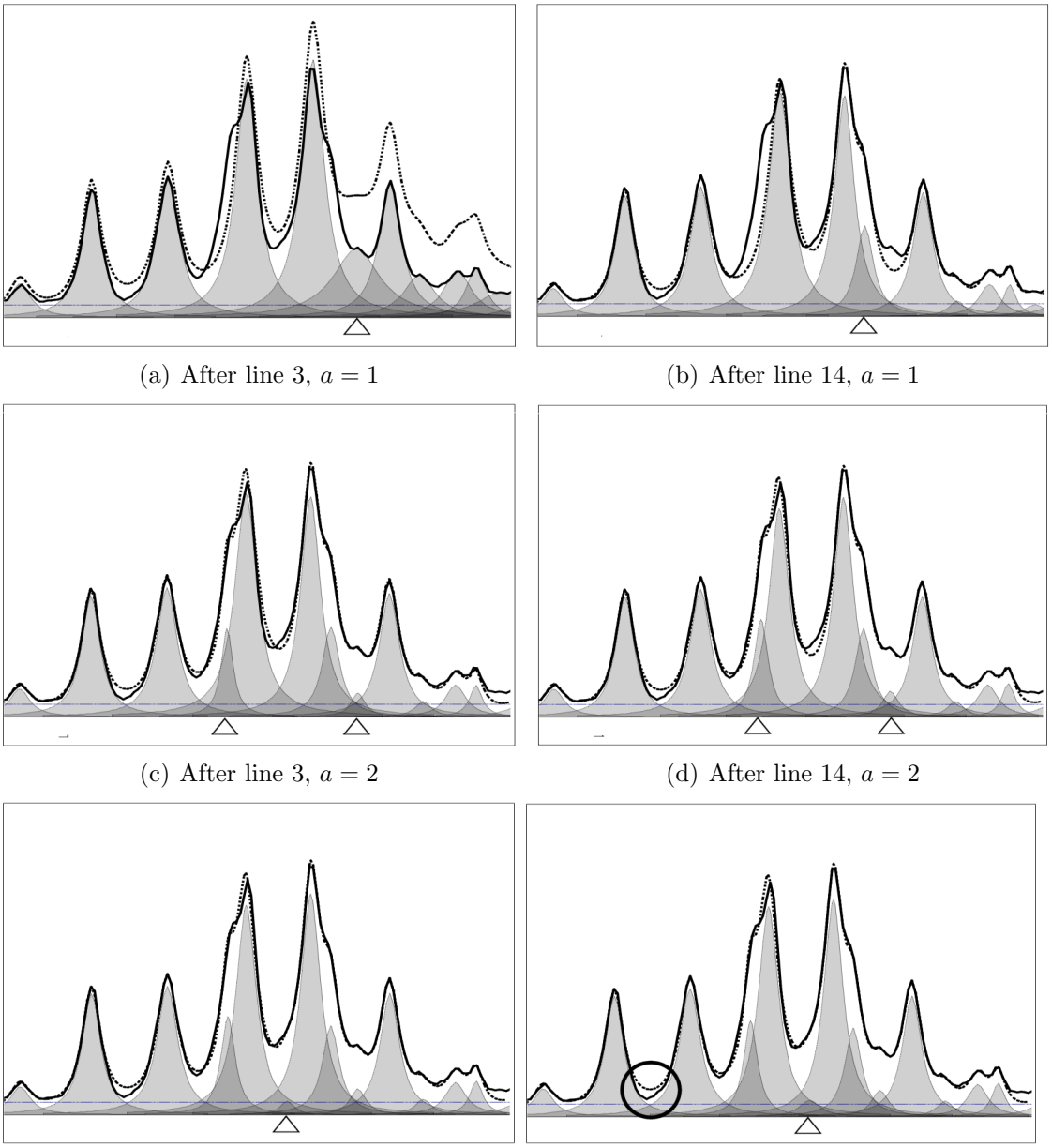
Figure 4.6: A subregion of the example "real-world" spectrum S_{real}. The modeled Lorentz functions are shown in grey color.

algorithm was applied without any line broadening. Baseline and phase correction were performed by the software ACD/SpecManager 6.0. This spectrum is in the following denoted as Sim_{real}. Figure 4.6 shows spectrum S_{real} in the spectral window ranging from 1.9 ppm to 4.8 ppm after peak reconstruction. In total, 292 peaks are found for the parameter setting $K_1 = 3$, $K_2 = 50$ and $r = 0.0005\,max$ (maximal spectrum intensity).

For reasons of simplicity, the reconstruction performance is shown in detail only for the subregion of the spectrum between 2.3 ppm and 2.4 ppm. As shown in figure 4.7, this region maintains both local maxima and *shoulders*. The reconstructed Lorentz functions are shown after the first (top row), second (center row) and third (bottom row) outer loop iteration of algorithm 2. The left column shows the current model before entering the inner loop (line 4), while the right column shows the current best model $\widehat{Y}$ after finishing the inner loop (line 13 of algorithm 2). Interestingly, the peak marked with a triangle in the top row moves to the left during the parameter approximation (compare figures 4.7(a) and 4.7(b)). A reason is given by the fact that the initial parameters are highly inaccurate (see dashed line in figure 4.7(a)), leading to substantial changes of the modeled height values $\widehat{Y}_j$ after applying the rule of proportion (4.3), i.e. the right side of that peak in figure 4.7(a) is more decreased than the left side, and by proportionally calculating the new values by (4.3), the local maximum moves to the left.

The detection of two more peaks in the second iteration of the outer loop is shown in figure 4.7(c) (triangles in the center row). As can be observed in figure 4.7(d), the new peaks fill in the gaps, and the subsequent approximation has no greater effect on the model.

The bottom row shows a similar situation: First, one more peak is detected (figure 4.7(d)), and then the peak set is approximated (figure 4.7(e)). In some regions, the model lies above the original spectrum (e.g. the region marked by



(a) After line 3, $a = 1$

(b) After line 14, $a = 1$

(c) After line 3, $a = 2$

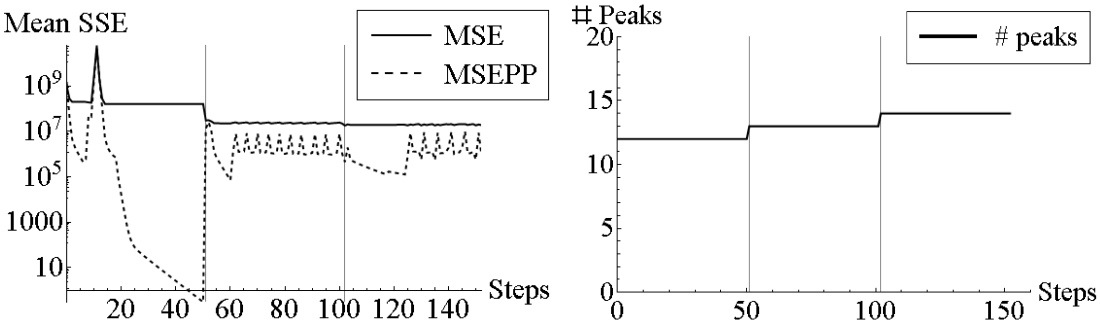
(d) After line 14, $a = 2$

(e) After line 3, $a = 3$

(f) After line 14, $a = 3$

Figure 4.7: Peak reconstruction of algorithm 2 on a subregion of S_{real}. Shown are the spectrum (solid curve), the set of Lorentz functions (grey areas), the predefined height threshold (horizontal line) and the corresponding superposition (dotted line) for intermediate states of the algorithm as indicated.

the circle in the bottom left of figure 4.7(e)). Since the algorithm considers only the distinct three positions $w_{j,l}, w_{j,m}$ and $w_{j,r}$ of each Lorentz function Y_j, such inaccurate regions may occur as a result of noise, distortions or yet hidden Lorentz functions.

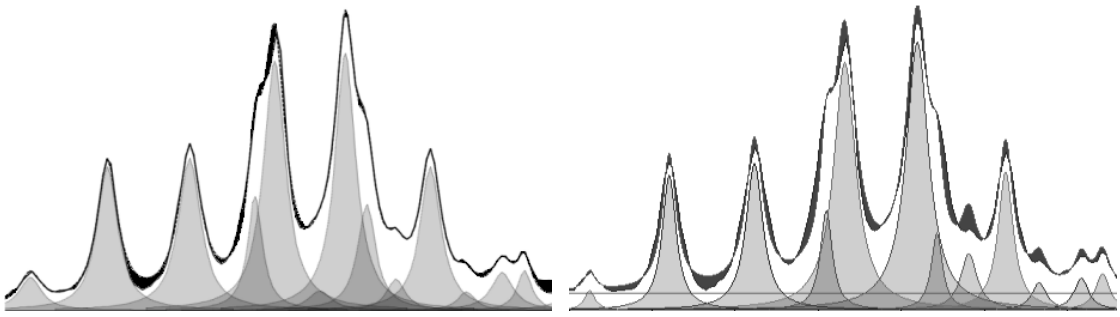


(a) Reconstruction error (b) Number of peaks

Figure 4.8: Error measures and number of Lorentz functions during the reconstruction of the subregion of S_{real}

Figure 4.8 shows the approximation error in each iteration step for the subregion of S_{real}. The error-function MSE almost remains unchanged during the execution of the inner loops, and even decreases directly after including the additionally found peaks (filling the gap). Generally, the measure MSEPP decreases during the first 50 steps, but between steps 5 to 20, MSEPP rises by several orders of magnitude before falling again. The previously mentioned process of the "moving" peak observed in figure 4.7 gives a reasonable explanation for this observation. The unstable development of MSEPP from step 60 may again be explained by either an incomplete or overrepresented model, or the occurrence of noise and other distortions. However, in general an error decay over magnitudes of order can be observed again.

At last, Figure 4.9 shows the final result of Algorithm 2 in comparison to an example outcome of the Levenberg-Marquardt Algorithm (Algorithm 1 of the previous chapter). The latter has been achieved by the commercial software *PeakFit 4.0*. Thereby, the peaks were found automatically as local maxima exceeding a height threshold (indicated by the horizontal line), and two peaks have been added manually to the model to also account for the shoulders of the two biggest peaks as well. One clearly observes, that the former Algorithm 2 yields a lesser amount of error than the latter in terms of the absolute difference (colored in black) between the model and the spectrum.



(a) Algorithm 2 (b) Levenberg-Marquardt

Figure 4.9: Final approximation result for S_{Real} by algorithm 2 (left) and by the Levenberg-Marquardt algorithm (right). The absolute difference between the model and the spectrum is marked in black color.

4.6 Discussion and Conclusion

In this chapter, an algorithm for reconstructing an NMR spectrum into a set of Lorentz functions is proposed. Empirical studies on simulated spectra show an exponential error decrease. The studies show that the achieved models maintain an especially low error at the chosen peak points of the spectra. Thus, the results indicate, that the approach of determining the analytical solutions for the parameters of a single Lorentz function in conjunction with the rule of proportion has much potential in fulfilling the task of spectrum modeling and thus feature extraction of NMR spectra.

Unfortunately, several drawbacks exist for Algorithm 2 and are pointed out as follows: As seen in the results section of this chapter, the proposed algorithm has a rather suboptimal performance on noisy data for the task of peak selection (see Section 3.1.3 of the previous chapter). High fluctuations of the error functions are observed for a considerably low noise amplitude of 1% of a spectrum's maximal intensity value. Moreover, further test runs with varying values for the height threshold r indicated that a clear separation of signaling peaks and noise distortions only based on the height information of an observed maximum is hard to achieve. In conclusion, the reliability of Algorithm 2 concerning an accurate identification of the actually underlying set of Lorentz functions of a given spectrum is considered to be rather unsatisfying. As mentioned before in the previous chapter, yet finding the set of Lorentz functions correctly is crucial to any approximation scheme if it is to accurately model a spectrum. For this reason, the focus of further investigations lies in improving the *peak selection* part of the approach, and further empirical

studies including for instance a detailed comparison to the Levenberg-Marquardt algorithm are yet omitted.

Mitigating the effects of noise by applying smoothing filters constitutes another reasonable strategy to help in preventing incorrect peak selection. However, a more general observation calls the whole approach into question: The cyclic methodology of finding local maxima, approximating the set of parameters, subtracting the model found so far from the spectrum, and finding local maxima again inherently leads to computational overhead, since at the beginning of each cycle the whole parameter set is adjusted all over again, and all achievements of the previous approximation, namely the inner loops of Algorithm 2, become obsolete. In other words, computational power is spent in each approximation cycle in order to detect new shoulders, which themselves change the whole model, and consequently, the efforts made so far for parameter approximation become useless.

In the following chapter, an approach is proposed to tackle the mentioned drawbacks by explicitly separating the tasks of peak selection and parameter approximation into two distinct problems. As will be shown theoretically and empirically, this leads to an improvement in both time consumption and quality of the results.

Chapter 5

Approach II: Lorentzian Spectrum Reconstruction

In Chapter 4, a first approach for reconstructing a spectrum into its distinct set of peaks was proposed, based on the analytical solution for the parameters of a Lorentz function and proportional approximation in accordance with a given spectrum. Although the results are shown to be promising, a major drawback is observed for the task of *model selection* (see Section 3.1.3). Further investigations on changing the threshold led to the conclusion that this approach is rather impractical to automatically separate signaling peaks from noise distortions. With the aim of improving the peak selection procedure, this chapter proposes an extension of the *Lorentzian Peak Reconstruction* approach. More specifically, the two tasks of peak selection and parameter approximation are solved sequentially by first simultaneously detecting *maximum* and *hidden* peaks of a spectrum, then approximating the corresponding parameters at a stroke.

5.1 Curvature-Based Peak Selection

5.1.1 Initial Considerations

A reasonable way to find all peaks simultaneously, even if they are overlapped, is to take into account the changes in the curvature of the spectrum, as already mentioned in Section 3.3.2 of Chapter 3. As a matter of fact, each local maximum of the spectrum has a corresponding root in its first derivative, and the roots of the

second derivative are the *inflection points*, those points at which the curvature of the original function changes its direction. A positive or negative second derivative value corresponds to a curvature of the original function in counter-clockwise or clockwise direction, respectively. Moreover, a local minimum in the second derivative stands for a locally maximal turn in clockwise direction of the original function, and hence gives rise to the curvature property of interest.

For the *Lorentzian Peak Reconstruction* approach of the previous chapter, separating the signal from noise and other distortions is based on a user-specified height threshold r. However, it turned out that in this way the outcome is inaccurate in the presence of noise. Here, each position w_i maintaining a negative local minimum in the second discrete derivative S'', i.e. $S''(w_i) < 0$ and $S''(w_{i+1}) > S''(w_i) \leq S''(w_{i-1})$, is considered to be a potential peak position, in order to identify both *maximum* and *hidden* Lorentz functions in a spectrum at once. In this way, peak identification can be seen as based on searching for little "bumps" in the spectrum instead of local maxima only.

More formally, a second derivative minimum w_m is assigned a surrounding interval $[w_l, w_r]$ with $w_l, w_r \in \mathbf{w}$ as the closest zero crossings, local maxima or plateaus in S'', either of which is closer positioned to w_m, more formally defined as follows:

Definition 5.1 (Peak Triplet)
A peak triplet is given as a triplet of positions $\{w_l, w_m, w_r\} \subset \boldsymbol{w}$*, for which the following holds:*

$$
\begin{aligned}
&1. \quad S''(w_m) < 0 \quad \wedge \quad S''(w_{m+1}) > S''(w_m) \leq S''(w_{m-1}), \\
&2. \quad l > m \quad \wedge \quad S''(w_{j+1}) > S''(w_j) \qquad \forall j \in \{l-1, \ldots, m\} \\
&\qquad\qquad \wedge \quad \big(\underbrace{S''(w_{l+1}) \leq S''(w_l) > S''(w_{l-1})}_{\textit{maximum or plateau}} \\
&\qquad\qquad\qquad \vee \underbrace{(S''(w_{l+1}) \geq 0 \ \wedge \ S''(w_l) < 0)}_{\textit{zero crossing}} \big), \\
&3. \quad r < m \quad \wedge \quad S''(w_{j+1}) < S''(w_j) \qquad \forall j \in \{m-1, \ldots, r\} \\
&\qquad\qquad \wedge \quad \big(\underbrace{S''(w_{r+1}) < S''(w_r) \geq S''(w_{r-1})}_{\textit{maximum or plateau}} \\
&\qquad\qquad\qquad \vee \underbrace{(S''(w_r) < 0 \ \wedge \ S''(w_{r-1}) \geq 0)}_{\textit{zero crossing}} \big).
\end{aligned}
$$

Note that, as mentioned in Chapter 3, the positions are given in descending order, i.e. $w_i > w_{i+1} \ \forall i \in \{1, \ldots, n-1\}$. In opposition to the previous chapter, where a point triplet was defined as a local maximum and the next neighbouring points of the spectrum, a peak triplet is here defined by properties of a spectrum's curvature, i.e. as a local minimum of the second derivative S'' and its nearest zero crossings, local maxima or plateaus.

Definition 5.2 (Triplet Score)
With $\boldsymbol{p} = (w_l, w_m, w_r)$ denoting a peak triplet, a score to distinguish between signal and noise is defined as:

$$score(\boldsymbol{p}) = \min\left(\sum_{k=l}^{m} |S''(w_k)|, \sum_{k=m}^{r} |S''(w_k)| \right). \tag{5.1}$$

In cases where the resolution of a spectrum is low, it is suggested to calculate the respective area by the gaussian trapezoidal formula instead. The main idea of (5.1) is to account for both the overall negativity of the second derivative and the corresponding interval width, namely for the degree and the length of a consecutive, clockwise-rotating curvature of the spectrum, assuming that noise distortions result in high fluctuations but over a smaller number of datapoints. The minimum of both sides is chosen in order to suppress an overrating of triplets due to the possible occurrence of asymmetric second derivative shapes.

Separation then takes place by discarding those peak triplets whose corresponding score falls below the mean plus δ times the standard deviation of scores out of a presumed signal-free region R. In most metabolite experiments, this region can be found below -0.5 ppm or above 10 ppm.

The occurrence of noise and other distortions in the spectrum leads to the occurrence of additional local maxima and minima in the second derivative and thus to a wrong selection of peak triplets. In order to suppress these effects, an *ad hoc* solution within this thesis is given by repeatedly applying the *mean filter*, also known as *smoothing filter, averaging filter, lowpass filter* or *box filter* (Gonzales & Woods, 2002; Davies, 2005). In basic description, each datapoint is replaced by the average value of its neighborhood, thus leading to a *smoothing* of the spectrum. A more formal description of the *mean filter* is given by algorithm 3.

Algorithm 3 (a,b-Mean Filter)

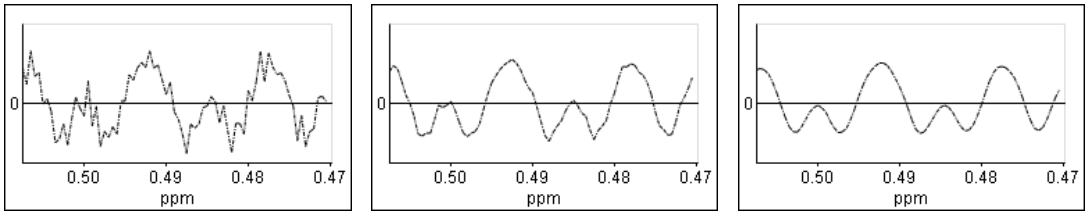
Input: Spectrum $S = \{S(w_1), \ldots, S(w_n)\}$, parameters $a, b \in \mathbb{N}$
Output: Filtered spectrum S^*

1: **for** $j = 1$ to b **do**
2: **for** $i = 1$ to n **do**
3: $$S^*(w_i) = \frac{1}{a} \sum_{k=i-\lfloor \frac{a}{2} \rfloor}^{i+\lceil \frac{a}{2} \rceil} S(w_{k'}) \quad \text{with} \quad k' = \begin{cases} |k| + 1, & \text{for } k \leq 0, \\ 2n - k, & \text{for } k > n, \\ k, & \text{else.} \end{cases}$$
4: **end for**
5: $S \leftarrow S^*$
6: **end for**
7: **return** S^*

Parameter b stands for the number of repeats, and parameter a denotes the filter width. Note that the spectrum at the upper and lower bounds is mirrored, as indicated by the case differentiation. With the observation that the sum in line 3 between two consecutive iteration steps only differs by $S(w_{i-\lfloor \frac{a}{2} \rfloor})$ and $S(w_{i+1+\lceil \frac{a}{2} \rceil})$, line 3 can be executed in constant time $O(1)$, resulting in an overall runtime of $O\big(b(n+a)\big)$. In the remainder of the thesis, applying an a, b-mean filter is also denoted as a, b-*filtering*. Figure 5.1 shows an example of a distorted second derivative before and after 2,2- and 3,3-filtering.

5.1.2 The Algorithm

The peak selection approach as a result of the previous section is described more formally by the *Curvature-Based Peak Selection* algorithm (CBPS, algorithm 4).



(a) Raw data (b) 2,2-filtering (c) 3,3-filtering

Figure 5.1: Examples of the second derivative after applying a mean filter as indicated.

In summary, after identifying the peak triplets of a spectrum, the scores of each of them is calculated. Filtering the signal from noise and other distortions takes place by comparing the found peak triplets with those out of a predefined region R of the spectrum.

Algorithm 4 Curvature-Based Peak Selection (CBPS)

Input: Spectrum S, second derivative S'', a signal-free region $R \subset S$, threshold parameter δ

Output: Filtered list of peak triplets L

1: $L, L' = \emptyset$
2: Find all peak triplets, given S'' and add to L'
3: Compute scores of each triplet $\mathbf{p}_i$ of L'
4: Compute $mean_{score}$ and sd_{score} given L', R
5: **for** $j = 1$ to $|L'|$ **do**
6: **if** $score(\mathbf{p}_j) \geq mean_{score} + \delta\, sd_{score}$ **then**
7: Add $\mathbf{p}_j$ to L
8: **end if**
9: **end for**
10: **return** L

With the observation that a peak triplet $\mathbf{p}_i$ as defined in Definition 5.1 can only have an overlap with neighbouring triplets in their boundary positions l_i, r_i, the worst-case runtime complexity of algorithm 4 lies in $O(n + |L|)$. This results from finding the peak triplets in line 2, which takes at most $O(n+|L|)$, calculating the discrete areas in line 3 needs $O(n + |L|)$ at most. Calculating the mean and standard deviation of the scores takes $O(|L|)$, and the *for*-loop in lines $5-9$ takes $O(|L|)$. Algorithm 4 is linear to the sum of spectral datapoints plus the number of second derivative minima. Considering that the maximal number of second derivative minima equals $\frac{n-1}{2}$ (each consecutive triplet consists of three points and overlap in their boundary position ($r_i = l_{i+1}$), algorithm 4 is linear to the number of datapoints, i.e. $O(n)$.

5.1.3 Results and Discussion

The performance is tested on 20 simulated spectra, each given as a sum of 100 Lorentz functions with a global HWHH parameter $\lambda = 0.005$, a global amplitude parameter $A = 1.0$, and with varying positions ω_i of a peak i given as

$$\omega_i = \omega_{i-1} + \lambda + u_\omega \, \lambda, \quad u_\omega \approx U(0,1), \tag{5.2}$$

beginning with $\omega_1 = 20\lambda$ and u_ω as a uniformly distributed random number. Spectrum distortions have been simulated by adding a uniformly distributed random number[1] v to each spectral datapoint as

$$S(w) = \sum_{j=1}^{100} A \frac{\lambda}{\lambda^2 + (w - \omega_j)^2} + v, \tag{5.3}$$

$$\text{with} \quad v_{\max} = \frac{(\frac{A}{\lambda})}{\rho} \quad \text{and} \quad v \in U(0, v_{\max}). \tag{5.4}$$

ρ is again the Signal-to-Distortion ratio (SDR) and reciprocal to the maximal distortion level v_{max}. Here, the SDR is expressed relative to the common peak height $\frac{A}{\lambda}$ instead of the maximal value of the entire spectrum (compare (4.5) of the previous chapter). In this way, the distortion level can be expressed proportional to the maximal height of a single Lorentz-function, and independent of respective overlapping effects.

The spectral range is given as $[0.0, \omega_{100} + 20\lambda]$ with resolution $|w_{i+1} - w_i| = \frac{\lambda}{10} = 0.0005 \; \forall i \in \{1, \ldots, n-1\}$ (n denotes the number of datapoints). The region of no signal R (see algorithm 4) was set to the ranges $[0, 10\lambda]$ and $[\omega_{100} + 10\lambda, \omega_{100} + 20\lambda]$.

The reasons for choosing uniform line width and amplitude parameters are both to maintain an equal level of distortion for each peak, and to control the resulting degree of peak overlap. Hidden peaks occur for the distance $d(Y_i, Y_j) \leq \frac{2}{\sqrt{3}} \lambda$ indeed only for the scenario of two equally scaled and shaped Lorentz functions. However, the first derivative of a Lorentz function decreases roughly to the power of three for increasing distances to its maximum (see Proposition 3.1, Figure 3.5 and (3.13) of Chapter 3). Thus, in combination with the positioning of the Lorentz functions as specified by (5.2), the contribution of further Lorentz functions on the overlapping

[1] Only positive distortions are considered to guarantee positive height values and therefore positive width and area parameters of a Lorentz function (compare (2.12)). In case of a real-world spectrum, only peak triplets containing positive values are further considered by now.

of a consecutive pair of peaks can be assumed to be constant, and is therefore neglected. On the basis of this pair-wise simplification, a rough estimate for the expected number of hidden peaks $E(\#hidden)$ in a sum of 100 Lorentz functions follows as

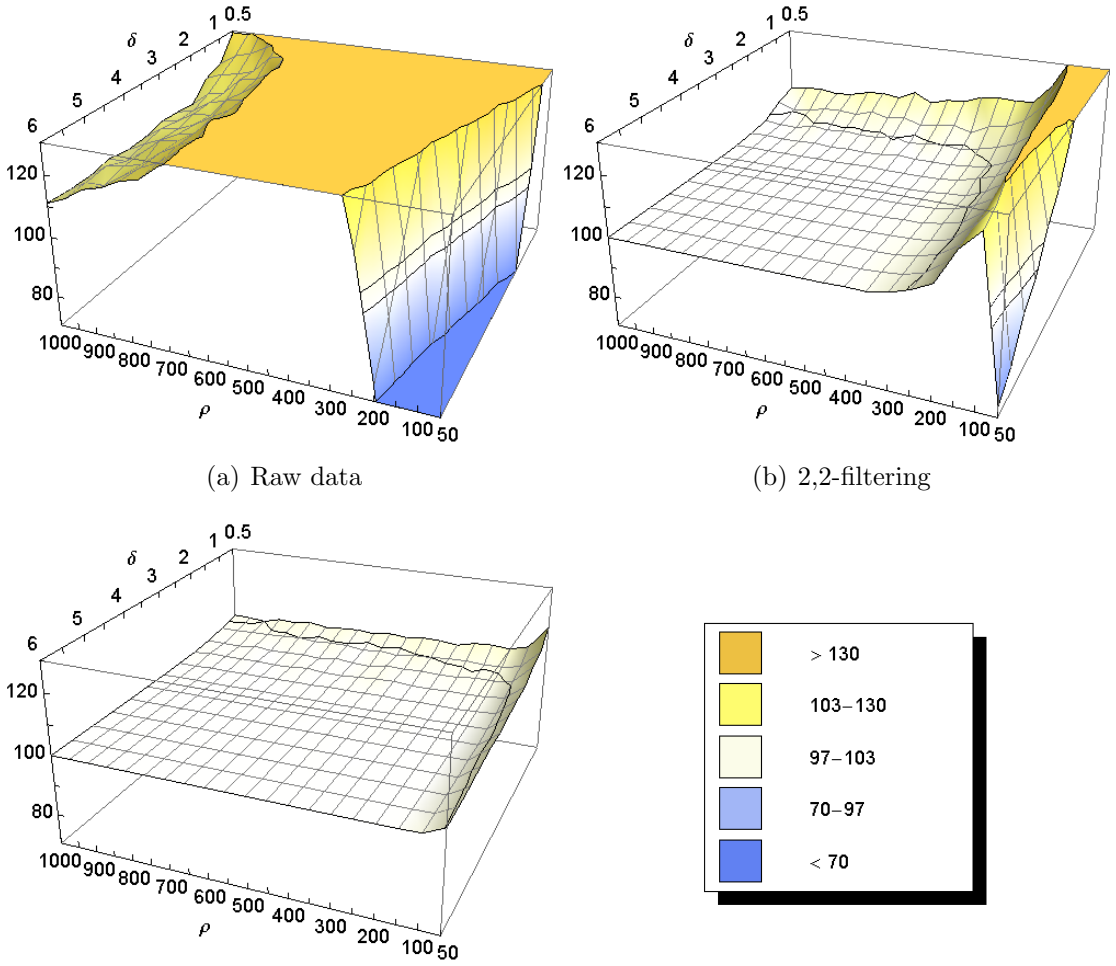
$$E(\#hidden) \approx 99 \left(\frac{2}{\sqrt{3}} - 1 \right) \approx 15, \tag{5.5}$$

by noting that the probability for each peak i to become a hidden peak then simplifies to the probability for u to become less equal than $\frac{2}{\sqrt{3}} - 1$. The correspondingly estimated probability for achieving a spectrum with all 100 peaks as *maximum* peaks is $(2 - \frac{2}{\sqrt{3}})^{99} \approx 6 \cdot 10^{-8}$.

Figure 5.2 shows the peak selection result of the CBPS algorithm (algorithm 4) for varying SDRs ρ (5.4) and varying picking thresholds δ (see line 6 in algorithm 4). For the unfiltered scenario, the number of found peaks exceeds the correct number considerably, as can be observed in figure 5.2(a). The reason lies in the occurrence of additional minima and maxima of the second derivative due to the incorporation of simulated noise (see for example figure 5.1(a)). In order to mitigate these distortion effects, the signal is smoothened by either applying a 2,2-mean filter (figure 5.1(b)) or a 3,3-mean filter (figure 5.1(c)).

Figure 5.3 provides the triplet score distribution along the position of the respective peaks in an example spectrum of figure 5.2 with SDR $\rho = 200$. The threshold score for $\delta = 3.0$ (line 6 of algorithm 4) is indicated by the solid horizontal line. The peak scores of the unfiltered scenario (figure 5.3(a)) are unfavourably distributed in terms of separating the signal peaks based on δ. After smoothing the spectra, a clear separation of peak and noise triplets can be observed in figures 5.3(b) and 5.3(c), leading to a drastic improvement in the selected number of peaks (compare figures 5.1(b) and 5.1(c)). Figure 5.3(c), for example, also shows that the range of the picking threshold δ may actually be much higher (here up to 50) than the considered maximal value of 6 for the mesh plots (figures 5.2(a) - 5.2(c)). This indicates that the triplet score as defined in (5.1) seems to be a reasonable measure for discriminating Lorentz functions from noise.

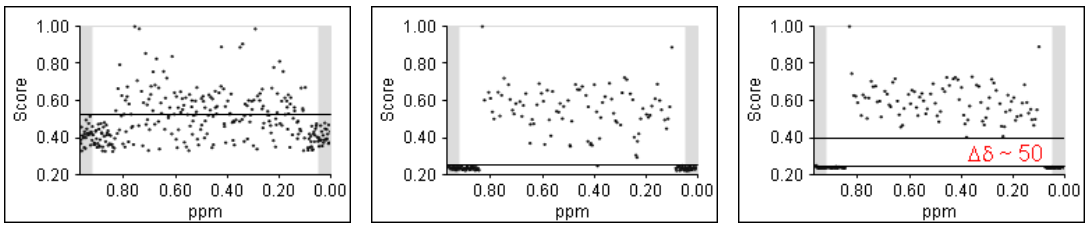
Figure 5.4 provides corresponding contour plots of the CBPS algorithm. The colour indicates the selected number of peaks, and each crossing of the gridlines indicates an evaluation point. The same results can be observed as before: In the case of raw data, CBPS results in an over-selection of peaks, even for higher SDRs (lower distortion amounts), as shown in figure 5.4(a). Again smoothing the data



(a) Raw data (b) 2,2-filtering

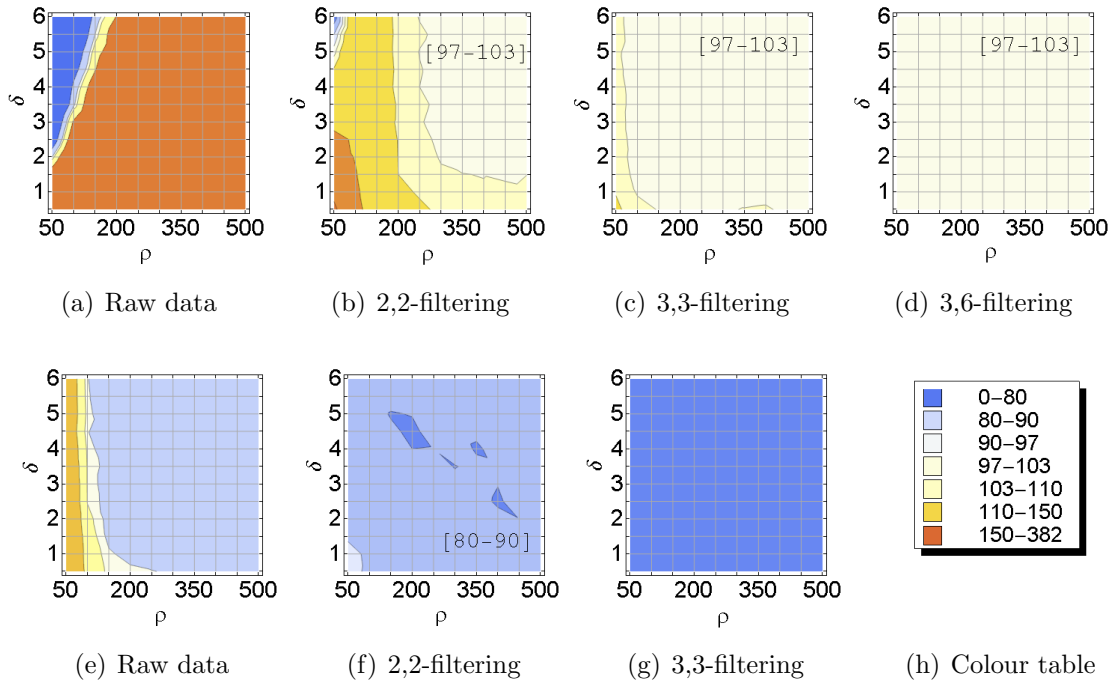
(c) 3,3-filtering (d) Colour table

Figure 5.2: Peak selection results of algorithm 4 with mean-filtering as indicated. Shown are the number of selected peaks on average out of 20 spectra on the $z-$axis for varying SDRs ρ along the $x-$axis and varying picking thresholds δ along the $y-$axis.



(a) Raw data (b) 2,2-filtering (c) 3,3-filtering

Figure 5.3: Example triplet scores for SDR $\rho = 200$ in correspondence to figure 5.2. The grey areas denote the chosen signal-free region R, and the solid horizontal line indicates the selection score for threshold $\delta = 3.0$.



(a) Raw data (b) 2,2-filtering (c) 3,3-filtering (d) 3,6-filtering

(e) Raw data (f) 2,2-filtering (g) 3,3-filtering (h) Colour table

Figure 5.4: Peak selection results for algorithm 4 (top row) and algorithm 2 on average out of 20 spectra for varying SDRs ρ along the horizontal axis and varying selection threshold δ on the vertical axis. The number of selected peaks is indicated by the color. The data is smoothened by a mean filter as indicated.

leads to essentially improved picking results, as indicated by figures 5.4(b) - 5.4(d). In case of the 3,3- and 3,6-mean filter, a deviation of only 3% (97-103) can be observed for a SDR of only 100:1 and 50:1, respectively.

In comparison to the previous chapter, figures 5.4(e) - 5.4(g) show the outcome of executing line 2 of algorithm 2 on the same set of spectra with height threshold $r = 0.02$ max. As expected, the predominantly occurring empirical result of 80-90 selected local maxima in figures 5.4(e) and 5.4(f) is in accordance with the estimation of about 15 hidden peaks (5.5). The plot after applying a 3,6-mean filter does not essentially differ from figure 5.4(g) and is therefore omitted here.

5.1.4 Conclusions

In summary, the CBPS algorithm yields an overall deviation in the number of selected peaks of only 3% $(97 - 103)$ for the majority of parameter settings shown, indicating the ability to simultaneously detect *maximum* and *hidden* peaks, and thus representing an alternative method for solving the task of peak selection.

In comparison with the *Lorentzian Peak Reconstruction* approach of the previous chapter, the *outer loop* of algorithm 2 is omitted, and the parameter approximation scheme is applied on the final model. Although the improved results are observed only in conjunction with smoothing the data, namely after removing distortions in the second derivatives, the only user-specified picking threshold δ can be chosen from a quite broad range of values, indicating high applicability and robustness of the approach in general. However, the impact of smoothing the data by mean filtering is presumably not neglectable in the context of subsequent parameter approximation. It should be also noted that a correct number of selected peaks does not necessarily imply a correct selection of peaks. Both will be investigated and discussed further in the following section.

5.2 Proportional Approximation II

Once the set of peak triplets is known the corresponding parameters need to be fitted in accordance with the spectrum. By the use of the CBPS algorithm (4), peak selection does not need to be taken into account during the approximation scheme anymore, as it was the case for the *Lorentzian Peak Reconstruction* approach of the previous chapter. Thus, there is no need of additional "outer loops", and the approximation may directly yield the final result.

Reconsidering, that a given spectrum is ideally expressed as

$$S(\omega) = \sum_{j}^{|J|} A_j \frac{\lambda_j}{\lambda_j^2 + (\omega - \omega_j)^2} = \sum_{j}^{|J|} Y_j(\omega),$$

with $\widehat{Y}_j^{(i)}$ denoting the model of Lorentz function Y_j, $\widehat{Y}^i$ denoting the current model spectrum, i.e. the sum of all calculated components at iteration step i, and $w_{j,x}$ denoting the chosen positions $\{w_{j,left}, w_{j,max}, w_{j,right}\}$ of a triplet $\mathbf{p}_j$ (i.e. $x \in \{left, max, right\}$), the approximation again takes place by adjusting the corresponding height values as given by (4.3):

$$\frac{\widehat{Y}_j^{(i)}(w_{j,x})}{\widehat{Y}_j^{(i-1)}(w_{j,x})} = \frac{S(w_{j,x})}{\widehat{Y}^{i-1}(w_{j,x})} \Leftrightarrow \widehat{Y}_j^{(i)}(w_{j,x}) = \widehat{Y}_j^{(i-1)}(w_{j,x}) \frac{S(w_{j,x})}{\sum_{l \in J} \widehat{Y}_l^{(i-1)}(w_{j,x})}$$

and calculating the corresponding parameters given the exact solution.

As stated in Section 4.1 of the previous chapter, a triplet of positions needs to preserve the condition of forming a local maximum by their corresponding intensity values in order to directly calculate the corresponding lorentzian parameters (4.2). Considering that the CBPS algorithm also selects triplets corresponding to shoulders (see definition 3.4), additional care needs to be taken to allow direct parameter calculation, since in such cases the local maximum condition is not necessarily met. An intuitive way to re-establish this condition is given by mirroring the point with the smallest value to the vertical axis at the position of the second derivative minimum, and then taking the resulting point triplet for calculating the parameters. More formally, this leads to the following definition:

Definition 5.3 (Mirrored Point)
Given a spectrum S and a peak triplet $\boldsymbol{p} = \{w_l, w_m, w_r\}$ with indices $l > m > r$, for which the corresponding intensity values form an ascending shoulder, e.g. given as

$$S(w_l) < S(w_m) < S(w_r).$$

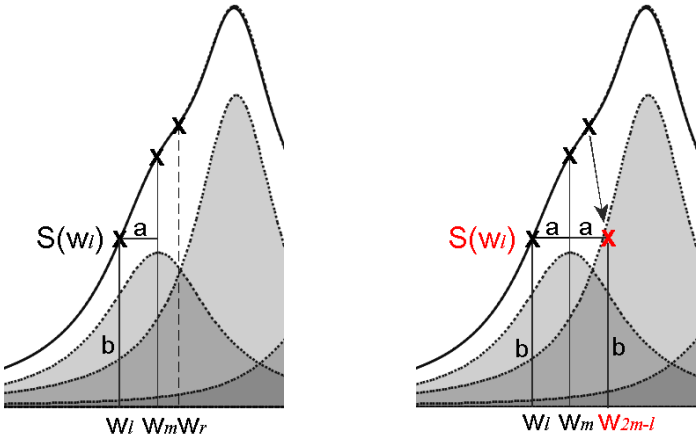
Then, the mirrored point of $\boldsymbol{p}$ is defined as

$$(w_{2m-l}, S(w_l)).$$

The mirrored point of a descending shoulder is defined analogously as

$$(w_{2m-r}, S(w_r)).$$

A corresponding illustration is provided by figure 5.5. This methodology is applied in all situations, in which the maximum constraint is not preserved, and has the beneficial effect that the search for local maximum point triplets along the corrected spectrum (lines 8-10 of algorithm 2) becomes obsolete. The corresponding algorithm *Proportional Approximation* (PA) is given by algorithm 5.



(a) Before Mirroring (b) After Mirroring

Figure 5.5: The mirror point for a peak triplet of an ascending shoulder in order to establish the maximum constraint of (4.2).

5.2.1 The Algorithm

Algorithm 5 Proportional Approximation (PA)

Input: Spectrum S, List L of peak triplets

Output: Approximated Parameter Set J

1: **Initial guess:** $\widehat{Y}_j^{(0)}(w_{j,x}) = S(w_{j,x})$ for all $j \in L$, calculate initial parameters
2: **for** $b = 1$ to K **do**
3: **for** $j = 1$ to $|L|$ **do**
4: Calculate the sum $\sum_{l \in J} \widehat{Y}_l^{(b-1)}(w_{j,x})$
5: Calculate new heights $\widehat{Y}_j^{(b)}(w_{j,x})$ by (4.3), and determine mirrored points, if needed
6: Calculate new parameters $\omega_j^{(b)}, \lambda_j^{(b)}, A_j^{(b)}$ by solutions of equation system (4.1)
7: **end for**
8: **end for**
9: **return** J

By observing, that the number of peak triplets equals the number of resulting Lorentz functions, and with K denoting the number of iterations for the outer *for*-loop of lines $2-8$ of algorithm 5, the worst-case runtime in terms of counting the number of essential comparisons is given as $O(K\,|J|^2)$. Interestingly, the runtime is independent of the number of spectral datapoints, since calculating the sum of

all Lorentz functions in line 5 takes time $O(|J|)$ for each peak triplet on its own, and the time needed to calculate the new height values and the new parameters in lines 6 lies in $O(1)$.

In summary, a spectrum can be translated into its distinct set of lorentzian parameters by the sequential execution of algorithms 4 and 5, as described in the algorithm *Lorentzian Spectrum Reconstruction* (LSR, algorithm 6).

Algorithm 6 Lorentzian Spectrum Reconstruction (LSR)

Input: Spectrum S, signal-free region $R \subset S$, threshold parameter δ, maximal iteration number K

Output: List J of peaks containing the approximated parameters

1: Find the list L' of peak triplets (algorithm 4) given S
2: Filter L' given R and parameter δ (algorithm 4) to receive L
3: Approximate parameter set of the Lorentz functions (algorithm 5) to receive J,
   given spectrum S, filtered peak triplet list L and parameter K
4: **return** J

The worst-case runtime of algorithm 6 lies in $O(n + K\,|J|^2)$, where n again denotes the number of spectral datapoints.

5.2.2 Results

In the following, the results of the proposed Proportional Approximation (PA) algorithm (Algorithm 5) are shown and discussed in comparison to the Levenberg-Marquardt algorithm (see Section 3.1.3.3 in Chapter 3), denoted henceforth as LM. The former has been implemented in the programming language C-Sharp (C#), and for the latter algorithm the software *Mathematica 6.0, Wolfram Research* was used (Weisstein, 1999). The evaluation is based on 20 spectra again, but with each containing only 20 Lorentz functions for reasons of time consumption. The parameters are given as

$$
\begin{aligned}
A_j &= u_A, & u_A &\approx U(50, 100),\ j \in \{1, \ldots, 100\},\\
\lambda_j &= u_\lambda, & u_\lambda &\approx U(0.002, 0.005),\ j \in \{1, \ldots, 100\},\\
\omega_i &= \omega_{i-1} + u_\omega \max(\lambda_i, \lambda_{i-1}), & u_\omega &\approx U(1.5, 2.0),\ j \in \{2, \ldots, 100\},\ \omega_1 = 0.0.
\end{aligned}
$$

u_A, u_λ and u_ω are uniformly distributed random variables. In opposition to the spectra generated for the picking evaluation, the pairwise distances between consecutive Lorentz functions now differ to a lesser degree of freedom, accounting for overlapping effects additionally introduced by varying the shape and amplitude parameters as described (5.2). Distortions are introduced similar to the picking evaluation (5.3), except that they are now considered relative to the maximal peak height by a uniformly distributed random number v, given as

$$S(w) = \sum_{j=1}^{20} A_j \frac{\lambda_j}{\lambda_j^2 + (w - \omega_j)^2} + v$$

$$\text{with} \quad v \approx U(0, v_{\max}) \quad \text{and} \quad v_{\max} = \frac{\max\limits_{j \in J} \left(\dfrac{A_j}{\lambda_j} \right)}{\rho}. \tag{5.6}$$

Parameter ρ again specifies the Signal-to-Distortions ratio (SDR) (5.4). For evaluation purposes, three different SDRs ρ are considered: $\rho = 1000$, $\rho = 500$ and $\rho = 200$, and a 5,3-mean filter (algorithm 3) is applied to smoothen the spectra. Subsequently, the peaks are found by the *Curvature-Based Peak Selection* (algorithm 4) with threshold parameter $\delta = 3.0$ and a noise region chosen as mentioned above, resulting in the selection of 20 peaks for all spectra.

Given spectrum S containing n datapoints and $|J|$ Lorentz functions with parameters ω_j, λ_j and A_j, and with $\widehat{Y} = \sum_i \widehat{Y}_i$ denoting the model with model parameters $\widehat{\omega}_j$, $\widehat{\lambda}_j$ and $\widehat{A}_j$, the following measures are used for evaluation purposes:

1. *Mean Squared Error* (MSE) as the standard error function of the discrete spectrum, given as

$$\frac{1}{n} \sum_{i=1}^{n} \big(S(w_i) - \widehat{Y}(w_i) \big)^2 \tag{5.7}$$

2. *Mean Squared Error at the Peak Hills* (MSE-PH) accounting for the mean squared error within the peak intervals $[w_{l_j}, w_{r_j}]$ (Definition 5.1), given as

$$\frac{1}{|J|} \sum_{j=1}^{|J|} \left(\frac{1}{r_j - l_j} \sum_{i=l_j}^{r_j} \big(S(w_i) - \widehat{Y}(w_i) \big)^2 \right) \tag{5.8}$$

3. *Mean Squared Error at the Peak Maxima* (MSE-PM) accounting for the squared error at each discrete peak maximum position w_{m_j}, given as

$$\frac{1}{|J|}\sum_{j=1}^{|J|}\big(S(w_{m_j}) - \widehat{Y}(w_{m_j})\big)^2 \tag{5.9}$$

4. *Mean Percentage Error of the Position Parameters* (MPE-Pos) accounting for the percentage error of the position parameters $\widehat{\omega}_j$, relatively to the original HWHH parameters λ_j of peak Y_j, given as

$$\frac{100}{|J|}\sum_{j=1}^{|J|}\Big|\frac{(\widehat{\omega}_j - \omega_j)}{\lambda_j}\Big| \tag{5.10}$$

5. *Mean Percentage Error of the HWHH Parameters* (MPE-HWHH) accounting for the percentage error of each HWHH $\widehat{\lambda}_j$, given as

$$\frac{100}{|J|}\sum_{j=1}^{|J|}\Big|1 - \frac{\widehat{\lambda}_j}{\lambda_j}\Big| \tag{5.11}$$

6. *Mean Percentage Error of the Area Parameters* (MPE-Area) accounting for the percentage error of the parameters $\widehat{A}_j$, given as

$$\frac{100}{|J|}\sum_{j=1}^{|J|}\Big|1 - \frac{\widehat{A}_j}{A_j}\Big| \tag{5.12}$$

The former three functions MSE, MSE-PH and MSE-PM account for the average squared error between the model and the spectrum in all datapoints, within the boundaries of the corresponding triplets, and at the minimum point of the second derivative, respectively, while the latter three functions MPE-Pos, MPE-HWHH and MPE-Area are used to describe the relative deviation in the particular parameters.

For LM, the parameters have been initialized with the initial parameters $\widehat{\omega}_j^{[0]}$, $\widehat{\lambda}_j^{[0]}$ and $\widehat{A}_j^{[0]}$, found in line 1 of algorithm 5. Since the outcome of LM is highly

dependent on the parameter initialization, five additional runs with varying parameters $\widehat{\omega}_j$, $\widehat{\lambda}_j$ and $\widehat{A}_j$ as

$$\widehat{\omega}_j = \widehat{\omega}_j^{[0]} + u_\omega, \qquad u_\omega \approx U(-0.001, 0.001) \tag{5.13}$$

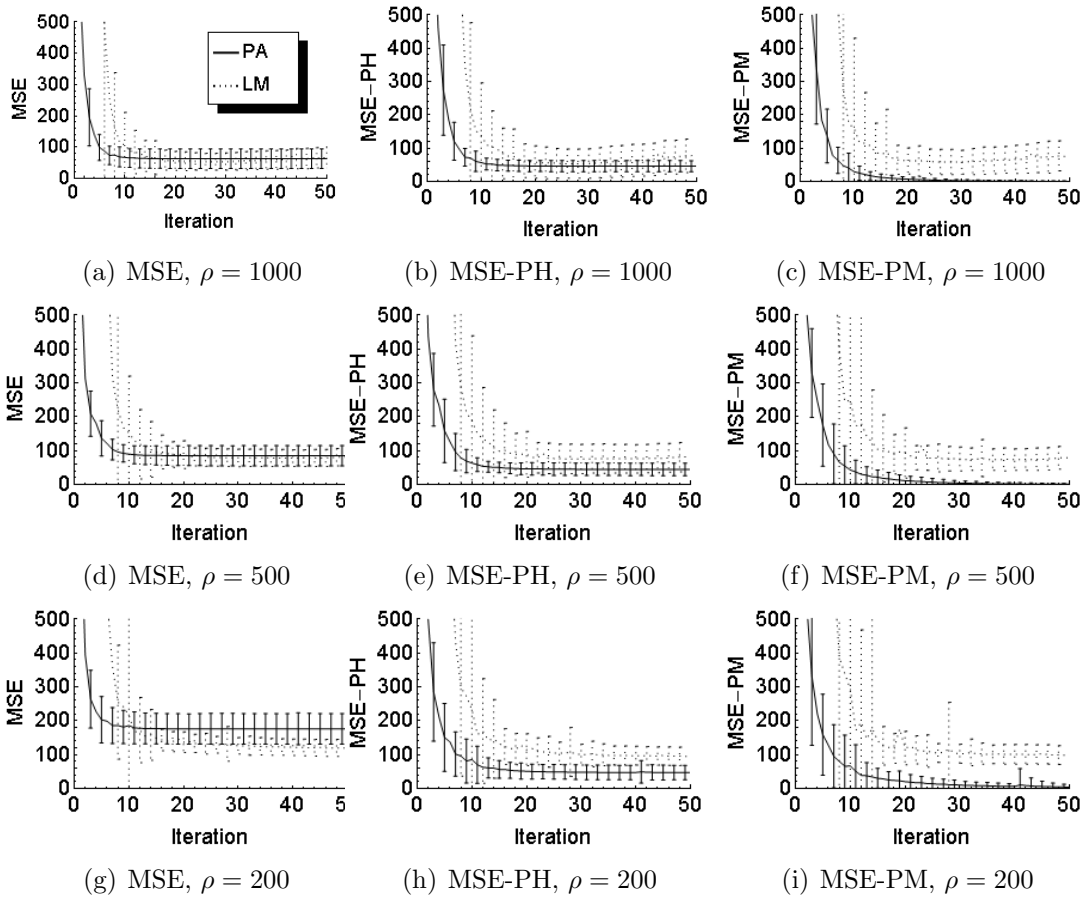
$$\widehat{\lambda}_j = \widehat{\lambda}_j^{[0]} \cdot u_\lambda, \qquad u_\lambda \approx U(0.5, 1.0) \tag{5.14}$$

$$\widehat{A}_i = \widehat{A}_j^{[0]} \cdot u_A, \qquad u_A \approx U(0.5, 1.0) \tag{5.15}$$

using uniformly distributed random variables u_ω, u_λ and u_A are considered as well. The parameters are decreased to address the fact that the spectrum is always exceeded by the initial guess $\widehat{Y}_j^{[0]}(w_{j,x}) = S(w_{j,x})$. For each spectrum, the fit which minimizes MSE (5.7) out of the initial and the additional five runs with decreased parameters as described are further considered for evaluation.

Figure 5.6 shows the mean squared error performance for SDRs $\rho = 1000$, 500 and 200 (5.6) and 50 iterations. It can be observed, that PA leads to a faster error-decrease than LM for all considered mean squared error functions, and outperforms LM especially for the error function MSE-PM, as shown in figure 5.6(c). The reason lies in the fact, that PA is based on the selected peak specific point triplets only, whereas LM is based on decreasing the mean error considering all data points. This observation is also in accordance with the results of the previous chapter in which the mean squared error at the peak maxima is shown to be essentially decreased in comparison to the average of all datapoints. Interestingly, PA even outperforms LM in terms of MSE-PH, the error at the *peak hills*, as can be observed in figures 5.6(e), 5.6(b) and 5.6(h), which again underlines its ability to focus on a spectrum's regions of interests, namely the peaks. Furthermore, PA yields are more robust performance than LM, as can be observed by comparing the length of the error bars of the two approaches.

Figure 5.7 shows the mean percentage error in the lorentzian parameters of PA and LM for different SDRs $\rho = 1000, 500$ and 200. PA clearly outperforms LM in both accuracy and especially robustness for all of the three lorentzian parameters position, HWHH and area of a peak. With a percentage error of less than 10%, PA yields peak position parameters with promisingly low deviation, as shown in figures 5.7(a) - 5.7(c). It can also be observed, that the error is already relatively small from the very beginning of the approximation, indicating that the proposed picking procedure (algorithm 4) is indeed capable of not only selecting the number of peaks correctly, but also identifying the set of peaks with high accuracy. With an average standard deviation of ca. $1\% - 3\%$ $(10\% - 50\%)$ in the position, ca.



(a) MSE, $\rho = 1000$ (b) MSE-PH, $\rho = 1000$ (c) MSE-PM, $\rho = 1000$

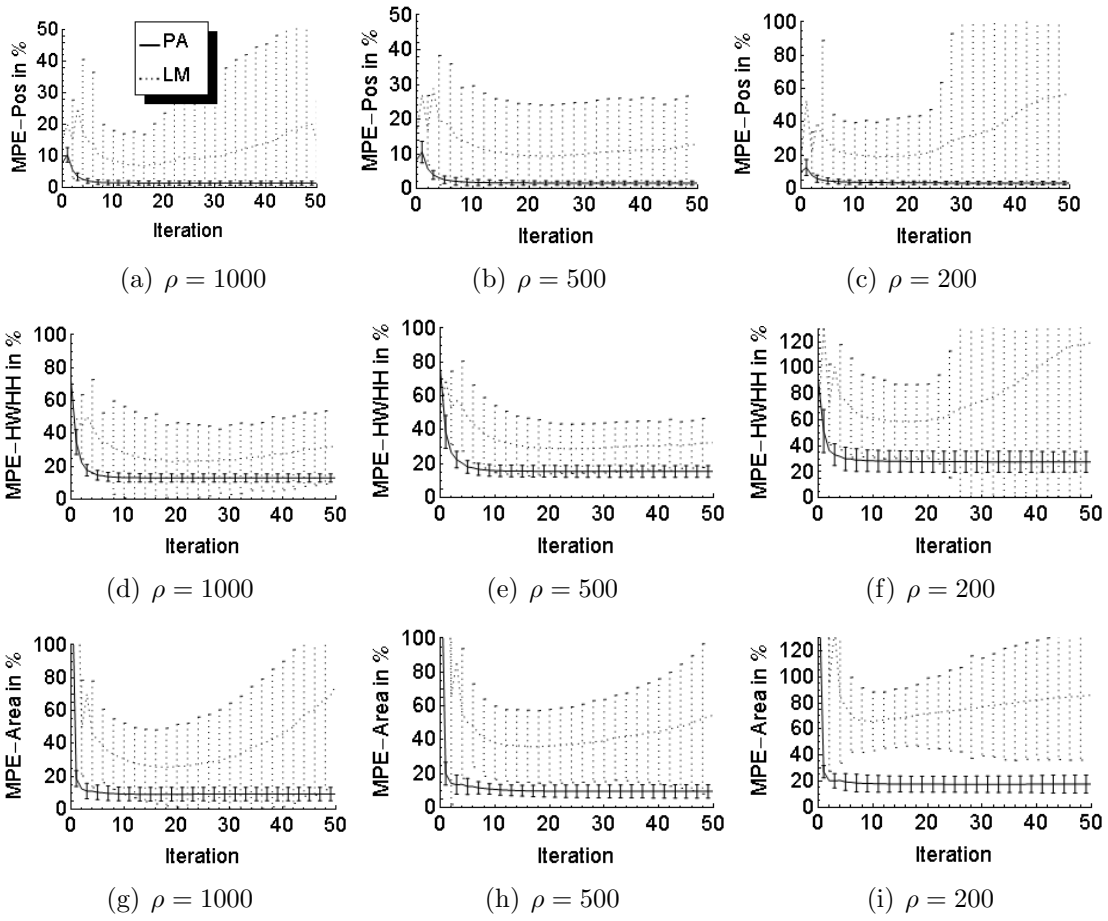
(d) MSE, $\rho = 500$ (e) MSE-PH, $\rho = 500$ (f) MSE-PM, $\rho = 500$

(g) MSE, $\rho = 200$ (h) MSE-PH, $\rho = 200$ (i) MSE-PM, $\rho = 200$

Figure 5.6: Mean squared error of the proposed method PA (solid line) and for Levenberg-Marquardt (dotted line) for SDRs $\rho = 1000$ (top), $\rho = 500$ (center) and $\rho = 200$ (bottom). The length of the error bars equals two times the standard deviation.

$5\% - 10\%$ ($20\% - 50\%$) in the HWHH, and ca. $5\% - 10\%$ ($20\% - 50\%$) in the area parameter for PA (LM), the proposed approach also shows much more robust behaviour than LM.

Peak fitting examples at different iteration steps are shown in figure 5.8. Here, the initial peak set for both LM and PA is provided by the CBPS algorithm (see figures 5.8(a), 5.8(b)). An interesting observation is given for LM by the peak marked in yellow. Resulting from the greedy nature of LM, the peak is steadily decreased during the whole fitting procedure, until its contribution to the sum becomes negligibly small, as shown in figures 5.8(c), 5.8(e) and 5.8(g). However, the global peak sum (red line) still fits the spectrum (black dotted line) considerably well. Figure 5.8(d) shows the peak set after one iteration of the proportional approximation algorithm. All peaks are significantly decreased, each in proportion



(a) $\rho = 1000$ (b) $\rho = 500$ (c) $\rho = 200$

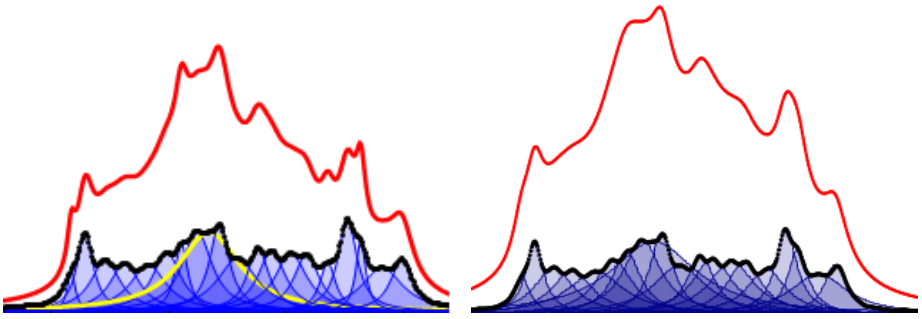
(d) $\rho = 1000$ (e) $\rho = 500$ (f) $\rho = 200$

(g) $\rho = 1000$ (h) $\rho = 500$ (i) $\rho = 200$

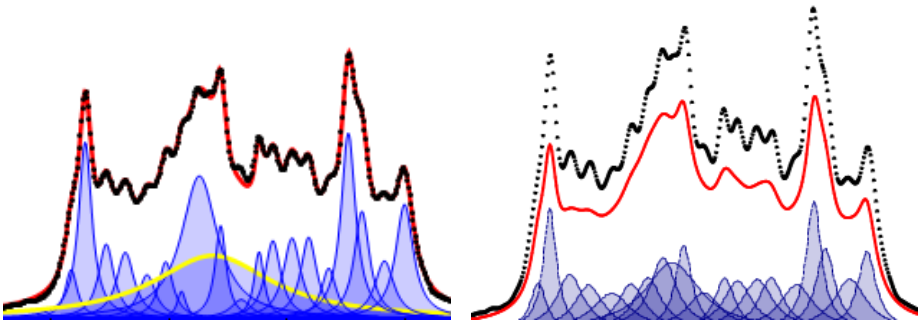
Figure 5.7: Mean percentage error of the parameters for the proposed method PA (solid line) and for Levenberg-Marquardt (dotted line) along 50 iterations, top: position (MPE-Pos), center: HWHH (MPE-HWHH), bottom: area (MPE-Area). The length of the error bars equals two times the standard deviation.

to the initial sum and the spectrum, and after 10 iterations an acceptable solution is already found.

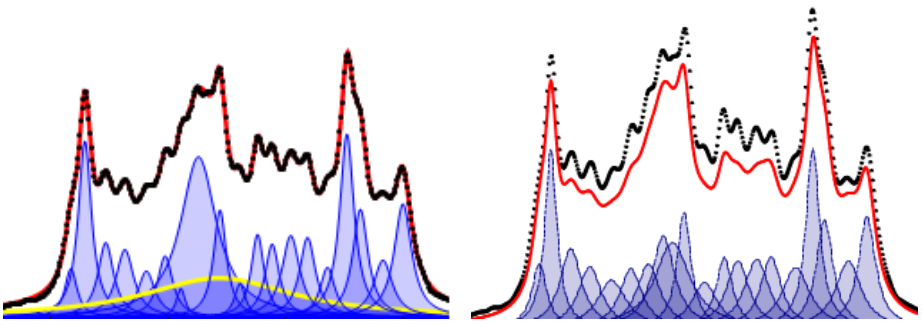
Figure 5.9 shows some regions of the real-world metabolite NMR spectrum S_{real} of the previous chapter after applying the proposed automated reconstruction approach. With smoothing the spectrum by a 3,3-mean filter and selecting the noise region R to the ranges $[12.8, 10.0]$ and $[-1, -3.4]$ (in ppm), 531 peaks were selected by the CBPS algorithm (algorithm 4) with picking threshold $\delta = 6.0$. The execution of the proportional approximation algorithm (algorithm 5) was finished after ca 9 seconds.



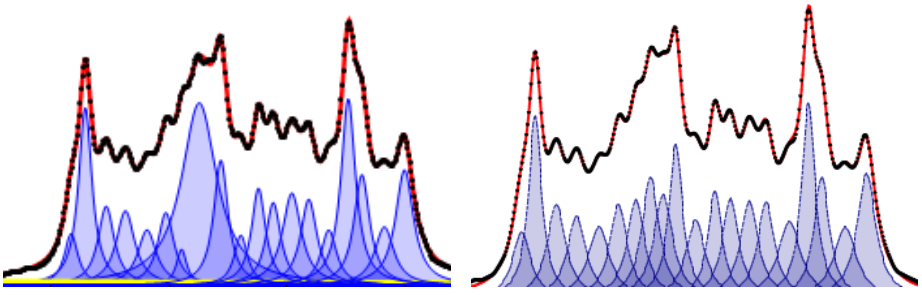
(a) LM, 0 steps (b) PA, 0 steps



(c) LM, 10 steps (d) PA, 1 steps



(e) LM, 20 steps (f) PA, 2 steps



(g) LM, 50 steps (h) PA, 10 steps

Figure 5.8: Example stages of the approximation for Levenberg-Marquardt (left) and the Proportional Approximation (PA). Shown are the spectrum (black dotted line), the model (red solid line) and the distinct Lorentz functions (blue areas)

.

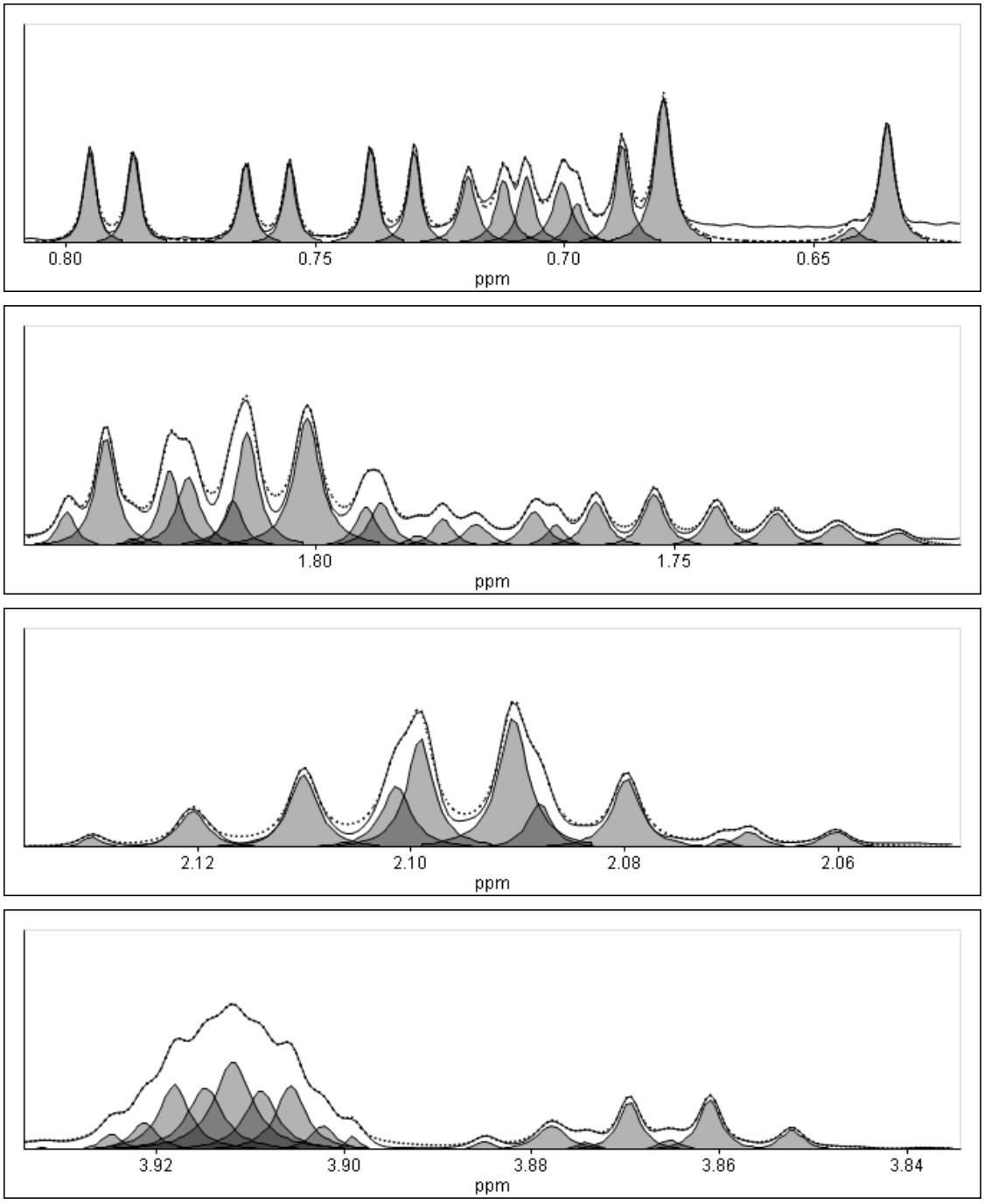


Figure 5.9: Portions of an example real-world spectrum (solid line), the fitted Lorentz functions by algorithm 5 (grey areas) and the corresponding sum (dotted line) after 10 iteration steps.

5.3 Discussion and Conclusion

In this chapter, a two-step approach for automated feature extraction is proposed, solving sequentially the tasks of peak selection and parameter approximation. Based on theoretical aspects concerning the overlap of two equally shaped and scaled Lorentzian functions, a runtime-efficient selection procedure (algorithm 4) capable of simultaneously detecting hidden and unhidden peaks is proposed. Simulations empirically demonstrate, that the proposed approach in conjunction with mean-filtering is able to find the set of signaling Lorentzian functions properly for a broad range of varying noise amplitudes and picking thresholds. A subsequent parameter approximation scheme (algorithm 5) is proposed, exploiting the analytical solution of a single Lorentzian function and adjusting the parameters in each step of the iteration by the rule of proportion, similar to the previous Chapter 4 except that the set of Lorentzian functions remains unchanged during the approximation.

Empirical studies show, that the proposed approach highly outperforms the Levenberg-Marquardt algorithm in terms of minimal error and robustness of the found model parameters. In particular, the results for the position and area parameters are highly promising. In comparison of figures 5.6 and 5.7, one clearly observes in case of the Levenberg-Marquardt algorithm that a low amount of mean squared error does not necessarily imply sufficiently accurate approximation of the model parameters. Especially the greedy hill-climbing nature of the algorithm, namely the combination of the *steepest descent* method and the local approximation of the fitness function by a second order polynomial (*Gauss-Newton*), is likely to result in only locally optimal solutions, in which the distinct parameters themselves might actually be highly falsified. In extreme cases, distinctive Lorentz functions even practically disappear from the model, as for example shown in figure 5.8. On the other hand, considering the parameters in a more compact way, for example by expressing the parameters of a single Lorentz function by polynomial expressions of a corresponding point triplet as proposed, allows to focus on the important regions of a spectrum, and the model is approximated faster, more accurately and more reliably in terms of the O-notation and the shown error measures.

As a final remark, this study is carried out for a sum of Lorentz functions only. However, the basic characteristics of feature extraction in general does not essentially change for datasets containing similar unimodal basis functions, and likewise, similar results and observations are expected to be made, provided that

the corresponding closed-form representation of the respective parameters can be achieved.

Chapter 6

Identification of Peak Palindromes

6.1 Motivation

In case of bio-molecular NMR experiments, especially those of samples containing a highly heterogeneous mixture of various substrate molecules, reliable automated analysis of large spectral series is a challenging task. In particular, the occurrence of chemical shifts prohibit an *ad hoc* point-by-point spectrum comparison, since it usually cannot be guaranteed that peaks originating from identical substrates are equally positioned along the spectrum series (see Section 2.6 in Chapter 2). Various methods for spectrum comparison have been proposed, ranging from early work, for example that based on non-linear least-squares approaches (Diehl *et al.*, 1975), to more recent approaches based on spectral binning (Chang *et al.*, 2007) or spectrum and peak alignment (Torgrip *et al.*, 2003; Yu *et al.*, 2006; Wong *et al.*, 2005), or both (Stoyanova *et al.*, 2004; Forshed *et al.*, 2005). However, an optimal solution has not yet been found.

In Chapters 4 and 5, two approaches for decomposing a spectrum into its distinct set of lorentzian parameters (2.12) have been proposed, for which empirical studies showed promising results in terms of accuracy, runtime efficiency and robustness. A peak-wise representation essentially reduces the data amount of a spectrum while simultaneously preserving the information content, but unfortunately an *ad hoc* peak-by-peak comparison of multiple spectra is still prohibited for the aforementioned reasons of peak shifts. A possible way to overcome the problem of concentration and pH dependence of the chemical shifts is to focus on the molecule-specific peak patterns of the spectrum. Based on the phenomenon

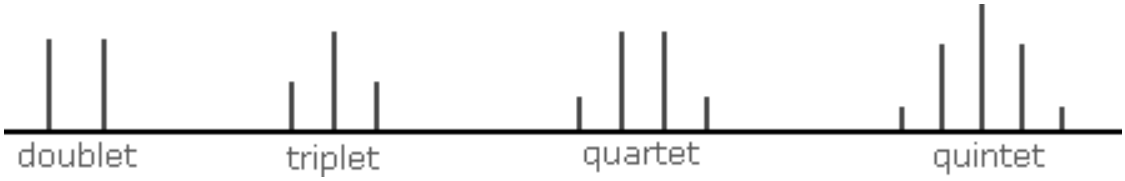


Figure 6.1: Example splitting patterns of multiplets, following Reusch (1999).

of *spin-spin coupling effects* during an NMR experiment, each molecule maintains its own set of specific spectral peak patterns called *singlets* (single peak) and *multiplets* (multiple peaks) (see Section 2.9). Figure 6.1 shows some basic example multiplets with the corresponding peaks displayed in a stick-wise manner.

In general, the majority of known bio-molecules exhibit multiplet patterns rather than resulting in a single peak of a spectrum. As an interesting property, the peak-to-peak distances within a multiplet are solely dependent on the magnetic moments of the particular nuclei, and spectrum-specific shifts therefore occur for a multiplet as a whole (see Section 2.9 in Chapter 2). Amongst these multiplets, quite a lot occur as peak sets with mirror-symmetric position parameters, mirror-symmetric heights, and mirror-symmetric peak shapes (Hoye *et al.*, 1994), in other words as patterns of peaks with a *palindromic* structure.

The problem of palindrome recognition on strings has been tackled for a long time (Manacher, 1975; Knuth *et al.*, 1977), with current applications in DNA and protein sequence analysis, as e.g. in Gupta *et al.* (2004). Thus, in the nature of strings being defined as a list of letters, the position distance between each consecutive pair of letters is equal and plays no role in the detection of palindromes on strings. However, this does not hold in general for palindromic peak sets. Furthermore, given an alphabet A and a string $s = \{s_1, \ldots, s_n\}$ with $s_i \in A$, palindromes on s have commonly been defined as a pair of two consecutive substrings wGw^R with w^R being w reversed and $G \in \{\emptyset, A^*\}$ as the potential gap in the center only (Kolpakov & Kucherov, 2008), whereas palindromic peak patterns indeed may overlap with each other. This thesis proposes an approach to identify mirror-symmetric peak patterns with respect to both the parameters of a Lorentz function and the possible occurrence of pattern overlap.

6.2 Definitions

For the remainder of this chapter, $Y = \{Y_1, \ldots, Y_n\}$ denotes the set of Lorentz functions representing a spectrum in descending order of the positions $\{\omega_1 > \cdots > \omega_n\}$. The following definitions are given in order to characterize *mirror-symmetric* or *palindromic* properties of a set of Lorentz functions.

Definition 6.1 (Peak Palindrome)
A Peak Palindrome is a set of peaks $M \subseteq Y$ of length $|M| = m$, for which the following symmetry scores are particularly defined as follows:

$$MPSE(M,c) = \frac{1}{k}\sum_{i=1}^{k} d_\omega(Y_i, c, Y_{m-i}) = \left| c - \frac{1}{m}\sum_{i=1}^{m}\omega_i \right| \quad \text{with } d_\omega(Y_l, c, Y_r) = |(\omega_l - c) - (c - \omega_r)|, \quad (6.1)$$

$$MSSE(M) = \frac{1}{k}\sum_{i=1}^{k} d_\lambda(Y_i, Y_{m-i}) \quad \text{with } d_\lambda(Y_l, Y_r) = |\lambda_r - \lambda_l|, \quad (6.2)$$

$$MASE(M) = \frac{1}{k}\sum_{i=1}^{k} d_A(Y_i, Y_{m-i}) \quad \text{with } d_A(Y_l, Y_r) = \frac{\max(A_l, A_r)}{\min(A_l, A_r)}, \quad (6.3)$$

$$\text{and where } k = \left\lfloor \frac{m}{2} \right\rfloor \text{ and } c \in \mathbb{R}.$$

M is called a δ-palindrome to a position $c \in \mathbb{R}$, if $MPSE(M,c) \leq \delta_\omega$, $MSSE(M) \leq \delta_\lambda$ and $MASE(M) \leq \delta_A$ holds, and M is called a perfect palindrome to c for $MPSE(M,c) = 0$, $MSSE(M) = 0$ and $MASE(M) = 1$. M is further called odd or even with respect to the number of peaks m. Accordingly, a pair of peaks (Y_l, Y_r) is called a δ-pair to c, if $d_\omega(Y_i, c, Y_j) \leq \delta_\omega$, $d_\lambda(Y_i, Y_j) \leq \delta_\lambda$ and $d_A(Y_i, Y_j) \leq \delta_A$ holds, and (Y_l, Y_r) is called a perfect pair for $d_\omega(Y_l, c, Y_r) = 0$, $d_\lambda(Y_l, Y_r) = 0$ and $d_A(Y_l, Y_r) = 1$. Finally, M is called a δ-pair palindrome, if each addend in (6.1) is a δ-pair, i.e. $d_\omega(Y_i, c, Y_{m-i}) \leq \delta_\omega$, $d_\lambda(Y_i, Y_{m-i}) \leq \delta_\lambda$ and $d_A(Y_i, Y_{m-i}) \leq \delta_A$ for all $i \in \{1, \ldots, k\}$.

Regarding the mirror-symmetry property, a δ-pair palindrome is more rigorously defined than a δ-palindrome. Moreover, the degree of symmetry in the

positions, areas and shapes of the peaks in M are expressed by the *Mean Position Symmetry Error* (MPSE), *Mean Shape Symmetry Error* (MSSE) and *Mean Area Symmetry Error* (MASE), respectively. Note that each addend in MPSE is dependent on a given position c, whilst MSSE and MASE are defined position-independently, which will become important in the remainder of this chapter. Also note that by definition it holds $d_A(Y_l, Y_r) \geq 1$. Moreover, the error in the areas of a peak pair (MASE) is expressed in a *scale-invariant* manner due to the assumption that distortions predominately occur proportionally to the absolute area of a multiplet, namely proportionally to the number of signaling nuclei (Friebolin, 1999).

Definition 6.2 (Best Matching Peak)
Given Y and a peak $Y_i \in Y$, and given symmetry weights α_λ, α_A and symmetry thresholds δ_λ, δ_A, the best matching peak Y_j^* to Y_i is defined by holding the following equations

$$d_\lambda(Y_i, Y_{j^*}) \leq \delta_\lambda \quad \wedge \quad d_A(Y_i, Y_{j^*}) \leq \delta_A,$$
$$\text{with} \quad j^* := \min_{j \in Y} \left(\alpha_\lambda \frac{d_\lambda(Y_i, Y_j)}{\delta_\lambda} + \alpha_A \frac{d_A(Y_i, Y_j)}{\delta_A} \right)$$

The best matching ancestor Y_{j^*} to Y_i is defined as the best matching peak Y_{j^*} with $j^* < i$. The best matching successor is defined analogously as the best matching peak Y_{j^*} for all $j^* > i$.

In other words, Y_{j^*} corresponds to the Lorentz function, which both fulfills the δ-pair constraints and minimizes the weighted normalized symmetry. Note, if for a peak a *best matching peak* exists, given thresholds δ_λ, δ_A and weights α_λ, α_A, then by definition they form a δ-pair palindrome.

Definition 6.3 (Weighted Pair Symmetry Error)
Given two peaks Y_l, Y_r, given a center position $c \in \mathbb{R}$, and given symmetry weights $\alpha_\omega, \alpha_\lambda, \alpha_A$ and symmetry thresholds $\delta_\omega, \delta_\lambda, \delta_A$, the weighted pair symmetry error $WPSE(Y_l, c, Y_r)$ is defined as

$$WPSE(Y_l, c, Y_r) = \alpha_\omega \frac{d_\omega(Y_l, c, Y_r)}{\delta_\omega} + \alpha_\lambda \frac{d_\lambda(Y_l, Y_r)}{\delta_\lambda} + \alpha_A \frac{d_A(Y_l, Y_r)}{\delta_A}$$

Based on the proposed definitions, the following section proposes a greedy algorithm to efficiently detect peak palindromes out of a set of Lorentz functions.

6.3 The Algorithm

In the following, a greedy approach is proposed to find δ-palindromes out of a peak set Y. Roughly speaking, the term *greedy* means that the search space is traversed based on only locally optimal, *greedy* decisions, commonly resulting in a decreased runtime, but to the loss of guaranteed optimal solutions.

In general, it can be presumed a priori that the spectral range covered by a multiplet is typically much smaller than the range of the whole spectrum (Friebolin, 1999), making it reasonable to introduce an additional threshold parameter δ^*_ω and correspondingly to find δ-palindromes of maximal spectral width δ^*_ω in a bottom-up manner. In particular, the problem of δ-palindrome identification is in the following divided into two stages: A. Finding a set of potential palindrome centers, and B. Maximizing for each center the number of δ-pairs to result in a δ-palindrome. Both parts are described in detail in the following.

6.3.1 Center Position Selection

In contrast to palindromes on strings where the maximal number of palindrome positions is linearly proportional to the string length, the position parameters in Y are given as arbitrary real numbers, and the maximal number of palindrome positions results as the number of all subsets of Y, $2^{|Y|}$. Since typically biological NMR experiments result in spectra containing hundreds of peaks, a potential palindrome center c is here considered heuristically to be given either for odd palindromes as the peak position $c = \omega_i$ of a peak Y_i, and for even palindromes as the mean position $c = \frac{\omega_i+\omega_j}{2}$ of a pair of peaks (Y_i, Y_j). This results in the maximal number of n center positions for the former, and $\frac{n^2-n}{2}$ for the latter case. Thus, for even palindromes not all pairs of peaks actually need to be considered because each δ-palindrome of length m would equivalently contribute $\Theta(m^2)$ positions on its own to the number of positions, whereas only one position is actually of interest. On the other hand, the intuitive alternative of considering only consecutive pairs of peaks might probably be too restrictive by means of overlapping palindromes.

As a compromise, for each peak $Y_i \in Y$ only the *best matching successor* (see Definition 6.2) is considered to give a potential palindrome center, resulting in a total number of $n-1$ potential positions of even palindromes. With a spectral range threshold δ^*_ω as mentioned above, all $n-1$ even positions can each be found

in time $O(k)$, where k denotes the average number of peaks within the spectral range $[\omega_i, \omega_i - \delta^*_\omega]$ of peak Y_i.

6.3.2 Palindromic Peak Addition

Provided position c and the next neighbouring peak pair (Y_{l-1}, Y_{l+1}) to a peak Y_l in the odd case, or to a pair of peaks (Y_l, Y_r) in the even case, a corresponding δ-palindrome M is found by iteratively adding or replacing those δ-pairs with lowest *weighted pair symmetry error* WPSE (Definition 6.3), in the range $[c + \frac{\delta^*_\omega}{2}, c - \frac{\delta^*_\omega}{2}]$ (remember, the peaks are given in descending order of their position parameters). Thereby, always the closer of the two peaks surrounding c is further iterated to the next neighbouring peak (bottom-up). Obviously, the runtime is $O(k)$, again determined by k as the number of considered peaks within the range specified by δ^*_ω.

In order to prevent redundant palindrome representation, i.e. given by two δ-palindromes M_1 and M_2 at respective positions c_1 and c_2 with $M_1 \subseteq M_2$ and $\text{MPSE}(M_1, c_2) \leq \delta_\omega$, we shall take a closer look on the symmetry measures of Definition 6.1. It can be observed, that MPSE is defined as the absolute distance between the mean peak position of a palindrome M and a given position c, whilst MSSE and MASE are position-independent, which leads to the following proposition:

Proposition 6.1
Given a δ-pair palindrome $M_1 = \{Y_1, \ldots, Y_{|M_1|}\}$ to position c, given a δ-palindrome $M_2 = \{Z_1, \ldots, Z_{|M_2|}\}$ to some other position, and given a position symmetry threshold δ_ω, the following holds:

$$MPSE(M_2, c) \leq \delta_\omega \;\Rightarrow\; M_1 \cup M_2 \text{ is a } \delta\text{-palindrome to } c$$

Proof *By noting that the symmetry scores MASE and MSSE are both independent from c, and with denoting $k_1 = \left\lfloor \frac{|M_1|}{2} \right\rfloor$ and $k_2 = \left\lfloor \frac{|M_2|}{2} \right\rfloor$, the proof follows as:*

$$M_1 \text{ is a } \delta\text{-pair palindrome } \Leftrightarrow d_\omega(Y_i, c, Y_{|M_1|-i}) \leq \delta_\omega \text{for all } i \in \{1, \ldots, k_1\},$$

$$\wedge \quad M_2 \text{ is a } \delta\text{-palindrome } \Leftrightarrow \frac{1}{k_2} \sum_{j}^{k_2} d_\omega(Z_j, c, Z_{|M_2|-j}) \leq \delta_\omega$$

$$\Rightarrow \frac{\sum_{i}^{k_1} d_\omega(Y_i, c, Y_{|M_1|-i}) + \sum_{j}^{k_2} d_\omega(Z_j, c, Z_{|M_2|-j})}{k_1 + k_2} = MPSE(M_1 \cup M_2, c) \leq \delta_\omega$$

□

In other words, given the δ-pair palindrome M_1 to a position c, the essential meaning of proposition 6.1 is that it suffices to guarantee $\text{MPSE}(M_2, c) \leq \delta_\omega$ for another δ-palindrome M_2 in order to preserve the symmetry constraint after unifying M_1 with M_2. The algorithm to find a δ-palindrome to a position c is given as follows:

Algorithm 7 (Palindromic Peak Addition)

Input: Y, c, Y_i, Y_j with $i \leq j$, $\omega_i \geq c \geq \omega_j$, thresholds $\delta^*_\omega, \delta_\omega, \delta_\lambda, \delta_A$, weights $\alpha_\omega, \alpha_\lambda, \alpha_A$

Output: δ-palindrome M to position c

peak pointer $pL, pR \leftarrow \emptyset$
if $Y_i \neq Y_j$ **then** *// even case*
  Add Y_i, Y_j to M
else *// odd case*
  Add Y_i to M
end if
$l = i + 1; r = j + 1;$
while $l > 0$, $r < |Y|$ **do**
  if $\omega_l - c > 0.5\,\delta^*_\omega$ or $c - \omega_r > 0.5\,\delta^*_\omega$ **then** *// spectral range constraint*
    return M
  else if $d_\omega(Y_l, Y_r) \leq \delta_\omega$ **and** $d_\lambda(Y_l, Y_r) \leq \delta_\lambda$ **and** $d_A(Y_l, Y_r) \leq \delta_A$ **then** *// δ-pair palindrome constraint*
    if δ-palindrome M_1 with inner peaks Y_l, Y_r already exists, and $\mathrm{MPSE}(M_1, c) \leq \delta_\omega$ **then**
      Add peaks of M to M_1 *// proposition 6.1*
      return M_1
    else if both Y_l and Y_r are not contained in M **then**
      Add Y_l, Y_r to M
    else if M contains Y_l already, and $\mathrm{WPSE}(Y_l, c, Y_r) < WPSE(Y_l, c, pR)$ **then**
      Replace pR by Y_r in M
    else if $\mathrm{WPSE}(Y_l, c, Y_r) < WPSE(pL, c, Y_r)$ **then**
      Replace pL by Y_l in M
    end if
    $pL \leftarrow Y_l, pR \leftarrow Y_r$
  end if
  if $\omega_l - c < c - \omega_r$ **then** *// iterate only the nearest peak*
    $l = l - 1$
  else
    $r = r + 1$
  end if
end while
return M

The peak addition takes place in lines $9-22$, and the bottom-up peak iteration is denoted in lines $24-28$. As stated above, the worst-case runtime of algorithm 7 lies in $O(k)$. Proposition 6.1 is applied in lines $12-14$ preventing a blow-up in both runtime and number of solutions. Each considered pair of peaks after line 11 is a δ-pair. Thus, each time line 13 is executed, a δ-pair palindrome M is added to some existing δ-palindrome M_1 and proposition 6.1 can directly be applied.

The following algorithm 8 summarizes the approach of finding a set of δ-palindromes in Y:

Algorithm 8 (Finding a set of δ-palindromes)

Input: Y, threshold parameters $\delta_\omega, \delta_\lambda, \delta_A$, weight parameters $\alpha_\omega, \alpha\lambda, \alpha_A$

Output: Set of δ-palindromes P

1: **for all** peaks $i \in Y$ **do**
2: Find the *best matching successor* Y_{j^*} to peak Y_i within the range $[\omega_i, \omega_i + \delta^*_\omega]$
3: **if** (Y_i, Y_{j^*}) is a δ-pair **then**
4: Find *even* palindrome M by algorithm 7 for $c = \frac{\omega_i + \omega_{j^*}}{2}, Y_i, Y_{j^*}$, and add to P
5: **end if**
6: Find *odd* palindrome M by algorithm 7 for $c = \omega_i$ and Y_i, and add to P if $|M| > 1$
7: **end for**
8: **return** P

Considering that each of the lines 2, 4 and 6 take time $O(k)$ with k as the average number of peaks within the respective ranges specified by c and δ^*_ω, the worst-case runtime of algorithm 8 lies in $O(|Y|\, k)$.

6.4 Results and Discussion

For demonstration purposes, preliminary results are shown and discussed for both simulated metabolite spectra as well as for the real-world example spectrum of the previous chapters, S_{real}. In the case of simulated data, three metabolites are simulated with 700.153 MHz, 0.0001192 ppm digital resolution and with 65536 datapoints. Those simulated are Valine, Methionine and Iso-Leucine. The corresponding Lorentzian functions have been found, and the parameters have been approximated by the *Lorentzian Spectrum Reconstruction* approach of the previous chapter

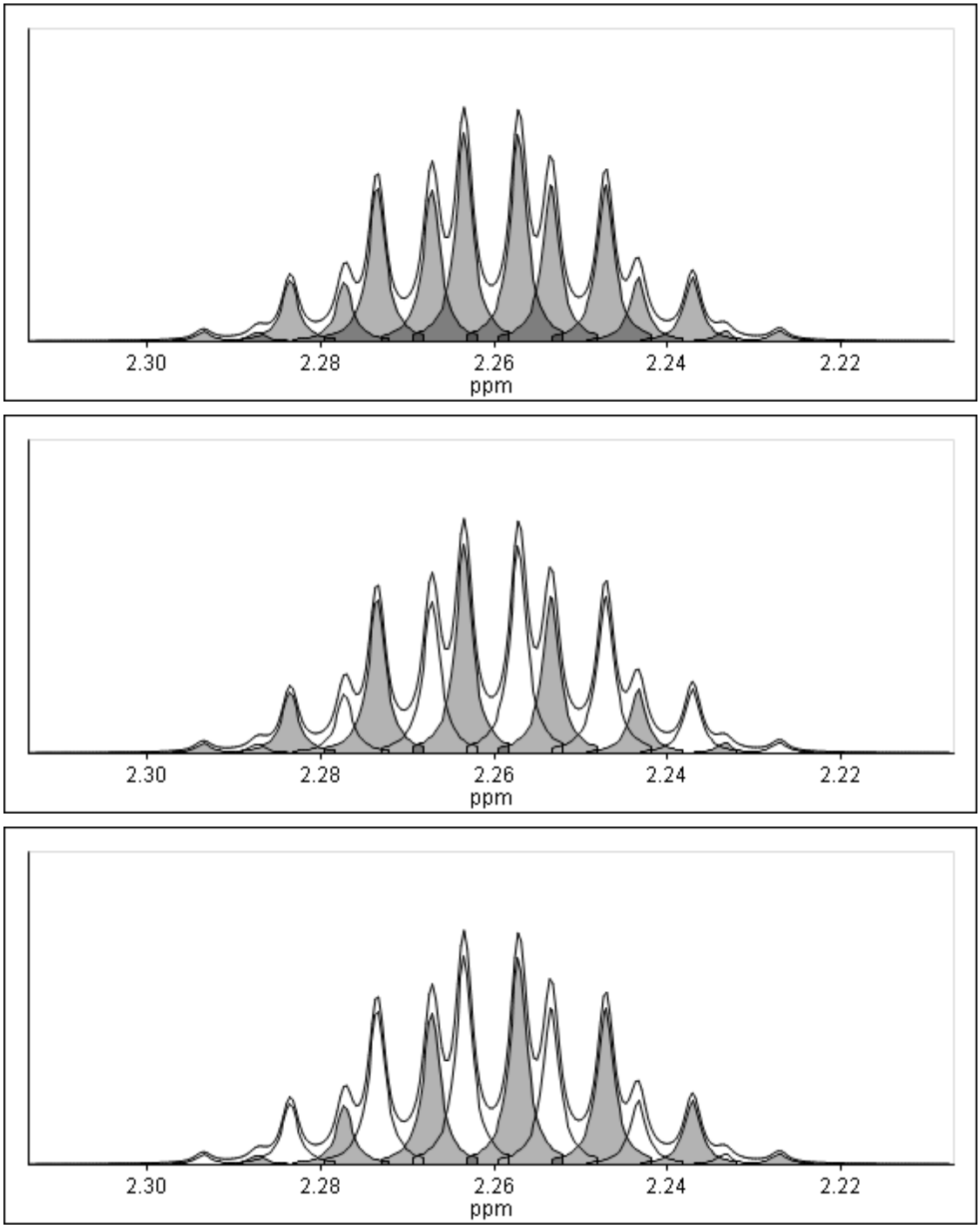


Figure 6.2: Some peak palindromes found for Valin (simulated). The spectrum (upper solid line) is shown for an example region of interest with the fitted Lorentz functions (solid peak lines below) and the found peak palindromes (grey areas).

(algorithm 6). The palindromes have been found with the spectral range threshold $\delta^*_\omega = 0.1$ ppm, the symmetry thresholds $\delta_\omega = 0.001$ ppm, $\delta_\lambda = 0.0002$ ppm, $\delta_A = 1.2$, and the weight parameters have been chosen as $\alpha_\omega = \alpha_\lambda = \alpha_A = 1.0$.

Interestingly, for the example multiplet structure of Valin shown in figure 6.2, a δ-palindrome containing 14 peaks (top) as well as two δ-palindromes containing each 7 peaks were found (center and bottom). Figure 6.3 shows some peak palindromes found as potential triplets and quartets of an example peak region of Methionine. Again, several mirror-symmetric multiplet structures give multiple solutions for explaining the whole peak pattern. Each palindrome shown in the first two rows represents a feasible solution in combination with any of the bottom two rows. A similar situation can be observerd for Iso-Leucine in figure 6.4. Within a cluster of peaks, again several mirror-symmetric sub-patterns are found and underline the potential of the proposed approach.

The real-world spectrum S_{real} was smoothened by applying a 3,3-mean filter (algorithm 3). Peak picking and peak fitting by the *Lorentzian Spectrum Reconstruction* approach of the previous chapter with the parameter setting $\delta = 6.0, K = 20$ resulted in 584 approximated Lorentz functions. The spectral range threshold was chosen as $\delta^*_\omega = 0.08$ ppm, the threshold parameters were chosen as $\delta_\omega = 0.005$ ppm, $\delta_\lambda = 0.0005$ ppm, $\delta_A = 1.5$, and the weight parameters were chosen as $\alpha_\omega = 1.0$, $\alpha_\lambda = 0.5$, $\alpha_A = 0.5$, to emphasize mirror symmetry in the position parameters.

Figure 6.5 shows some example peak palindromes found by algorithm 8. Next to the predominant occurrence of doublets (two peaks) and triplets (three peaks), quartets and higher order multiplets were found as well, some of which are shown.

6.5 Conclusions

In summary, this chapter proposed three measures to express the degree of multiplet symmetry for an arbitrary set of Lorentz functions, and introduced a parameterized greedy algorithm to automatically detect mirror-symmetric peak patterns. It is shown, that the proposed approach is indeed able to identify peak palindromes out of a set of overlapping peak patterns. However, the approach does not result in a single optimal solution of peak palindromes in terms of unambiguously representing the actual set of multiplets contained in a sample. Rather, a set of patterns is found with the potential to support the identification of overlapped multiplet patterns contained in a spectrum. In general, the approach has the potential to facilitate automated spectrum analysis in terms of bridging the gap between the lorentzian peak representation of a spectrum and a distinct mixture of multiplet patterns of the molecules contained in a sample. The prospective field of

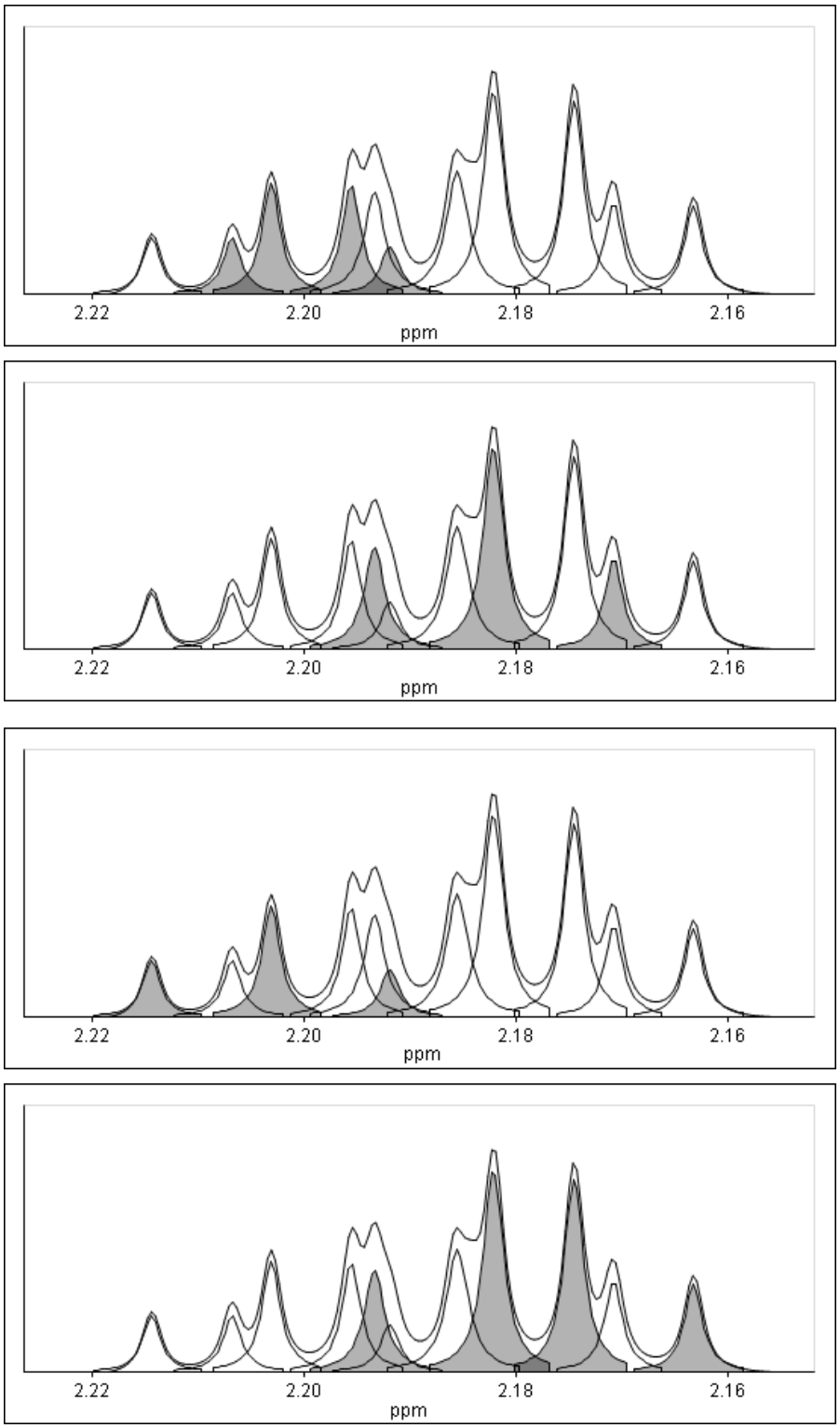


Figure 6.3: Some peak palindromes found for Methionine (simulated). The spectrum (upper solid line) is shown for an example region of interest with the fitted Lorentz functions (solid peak lines below) and the found peak palindromes (grey areas).

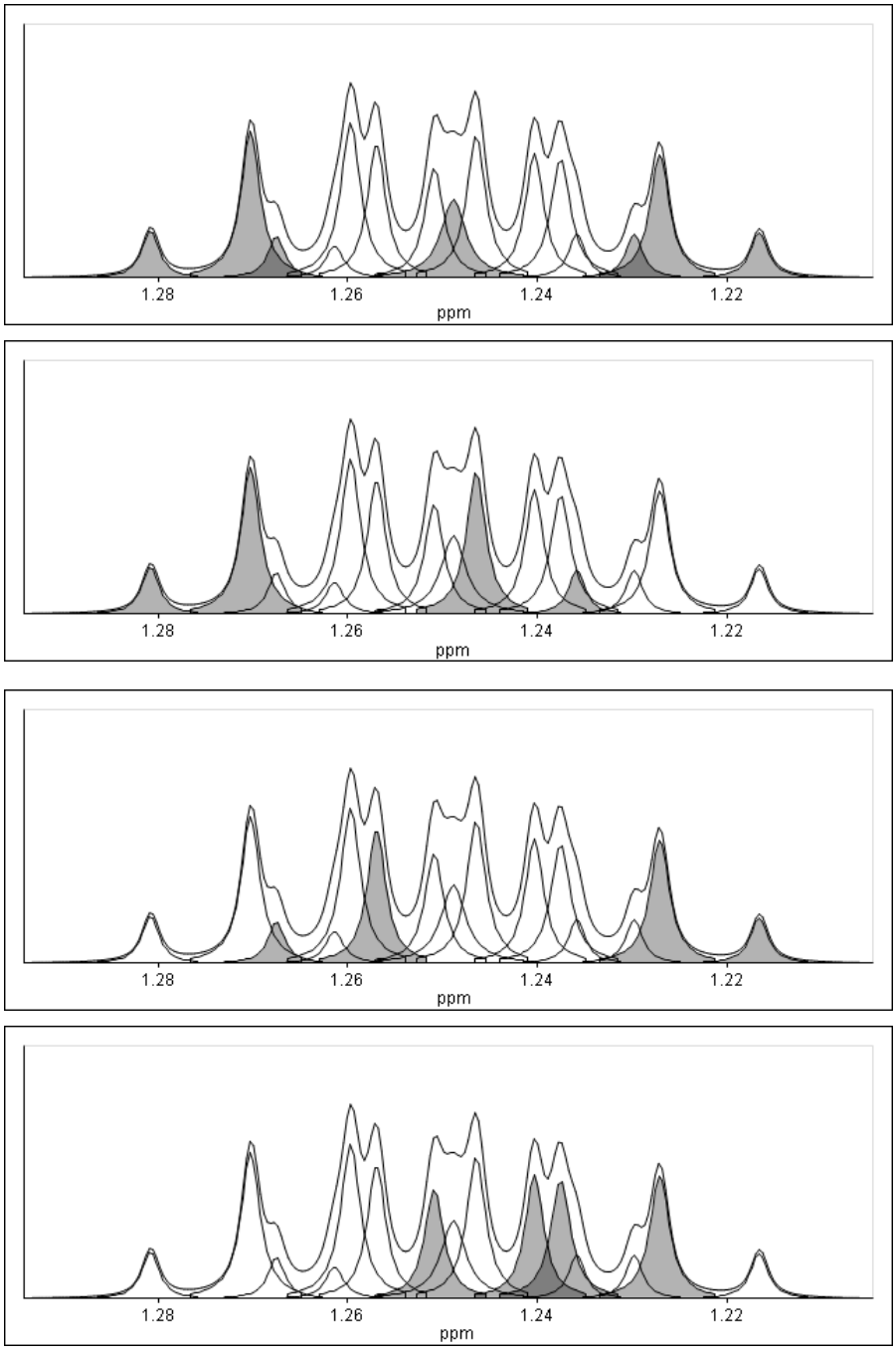


Figure 6.4: Some peak palindromes found for Iso-Leucine (simulated). The spectrum (upper solid line) is shown for an example region of interest with the fitted Lorentz functions (solid peak lines below) and the found peak palindromes (grey areas).

application ranges from automated database construction and extension to incorporation into more sophisticated multiplet and spectrum matching and alignment approaches.

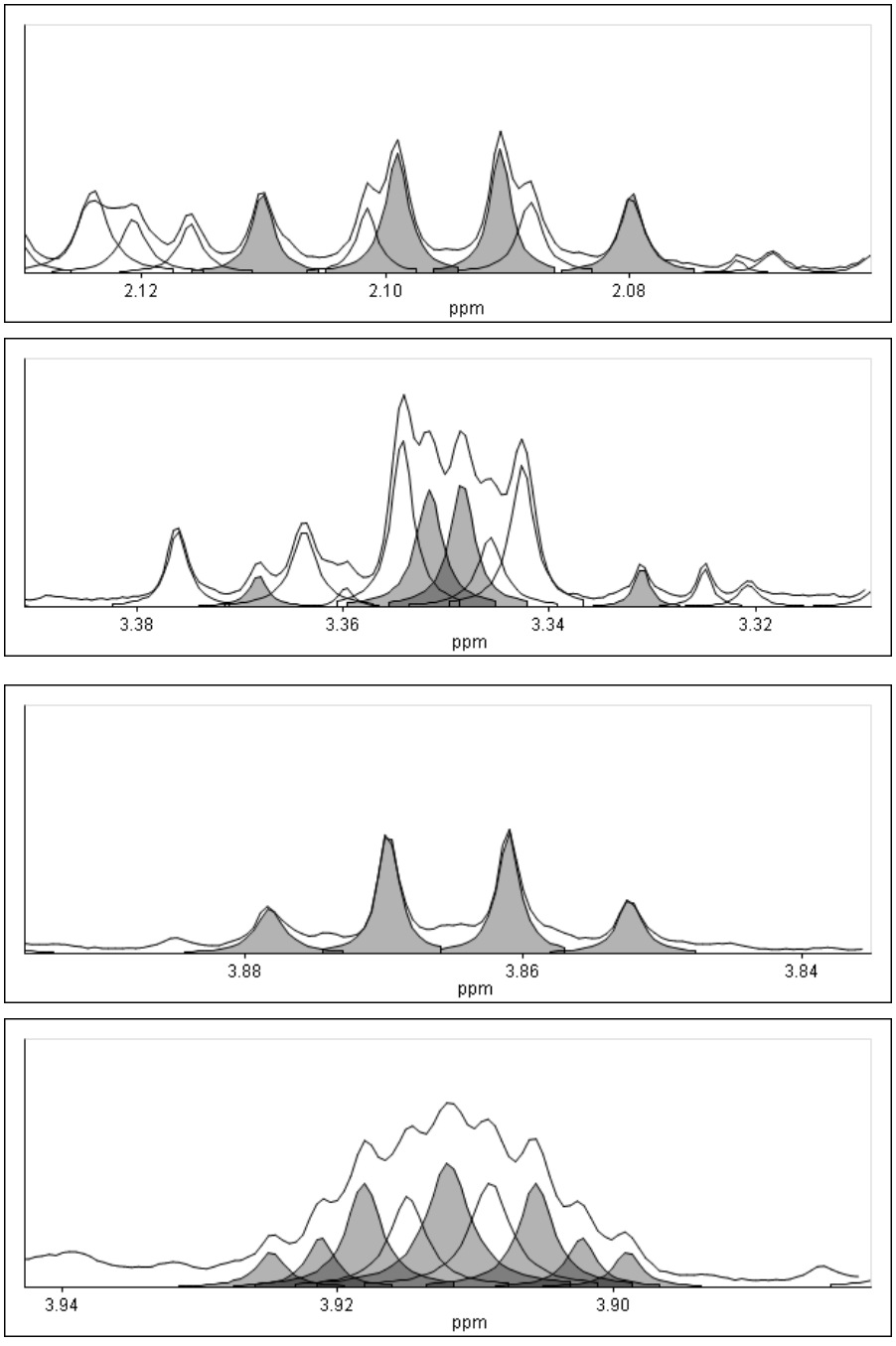


Figure 6.5: Some Palindromic peak sets found for the example real-world spectrum. The spectrum (upper solid line) is shown for an example region of interest with the fitted Lorentz functions (solid peak lines below) and the found peak palindromes (grey areas).

Chapter 7

Summary, Conclusions and Future Work

7.1 Summary and Conclusion

In this thesis an alternative approach for the task of feature extraction in 1D NMR data analysis has been proposed and is summarized as follows: In the absence of noise, magnetic field inhomogeneities, preprocessing artifacts and phasing errors, the pure signal can be modeled as a sum of Lorentz functions. Based on the analytical solution for the corresponding parameters of a single Lorentz function and on a proportional height adjustment procedure throughout the whole iteration scheme, an initial approximation algorithm called *Lorentzian Peak Reconstruction* is proposed in the first attempt in Chapter 4. The method fulfills both the tasks of *model selection* and *parameter approximation* simultaneously and empirical studies show promising results for both simulated datasets and an example real-world sub-spectrum. The key difference to conventional approaches is the inherent focus on the informative regions of the spectrum, namely the peaks instead of the average error along all spectral datapoints under consideration. As a drawback, the proposed method suffers from inaccuracies in model selection where noise and other distortions are present, leading to an over-estimation of the number of peaks, and thus to high fluctuations in the *mean squared error* within the peak regions during the approximation.

An improved method with a more stable and accurate peak selection, *Lorentzian Spectrum Reconstruction*, is proposed in Chapter 5. This solves the tasks of *model selection* and *parameter approximation* in a sequential manner by the algorithms

Curvature-Based Peak Selection and *Proportional Approximation II*. A peak in a given spectrum is identified by properties of its curvature, namely by the occurrence of roots, local minima and local maxima of the second derivative, instead of searching for local maxima of the spectrum. In conjunction with mean-filtering, empirical studies on simulated datasets show promising results regarding the number and position accuracy of the selected peaks. Subsequently, the approximation scheme is extended in order to cope with point triplets that are not given as local maxima, allowing even the parameters of shoulders to be proportionally approximated. Empirical studies on simulated datasets show moderately improved results in comparison to the common Levenberg-Marquardt algorithm in terms of minimizing the squared distance of all datapoints, and, as the most interesting results of this thesis, dramatically improved results regarding the accuracy of the parameters in the respective model.

At least for NMR spectral data, the observed results indicate that the methodology of focusing on few but potentially most significant points of a spectrum has much more potential than considering all points of the spectrum simultaneously. In addition, chopping the whole problem of NMR feature extraction into several smaller problems, i.e. data smoothing, peak selection and parameter approximation, and solving them in a sequential manner has empirically shown to be highly effective in terms of runtime complexity and quality of the resulting solutions. Nevertheless, the Levenberg-Marquardt algorithm is by all means an elegant method to search for local optima of the fitness function based on partial derivatives and the steepest descent. However, the greedy hill-climbing nature inherently limits the outcome of the algorithm. In cases where the corresponding fitness landscape seem to contain a high density of local optima, approaching the problem in a completely alternative way can indeed lead to essential improvements, as shown within this thesis.

Several interesting questions are still waiting to be answered, for example the minimal spectral distance of a pair of Lorentz functions to occur as a distinct second derivative minimum, or convergence properties of the proportional fitting procedure. In general, it can be concluded, that the proposed approach is highly suitable for solving the task of automated NMR feature extraction in terms of model selection and parameter approximation. It is worth noting that the proposed approach is in general not restricted to NMR spectroscopic data, but applicable to all spectra given as superpositions of known functions, for which the analytical solution of the respective parameters can be determined a priori. Thus, further

investigations of the proposed methods concerning the task of quantification in datasets of different basis functions are needed and have the potential to yield similarly interesting results.

To directly enable a reliable multivariate analysis and classification of NMR spectra obtained from a series of metabolite solutions, one has to cope with the sensitivity of the chemical shifts to concentration, temperature and the pH-value of the solutions. In particular, the signal-specific horizontal shifts of peaks in the frequency domain mainly prohibit a peak-by-peak comparison of different spectra. However, assignment is still beneficial for a thorough analysis of a spectrum series since it allows the construction of feature vectors, which contain quantitative information of the same source of origin throughout the whole set of experiments. In this context, given the spectrum model as a set of Lorentz functions, a heuristic algorithm for the identification of palindromic peak sets is proposed in Chapter 6, exploiting characteristic properties of the signal response of distinctive classes of molecules. In general, the algorithm aims at bridging the gap between the peak-wise representation of a spectrum and the *inter-spectrum peak linkage*, namely the linkage of signals originating from same sources of origin but observed in different NMR experiments.

With the introduction of three symmetry measures describing mirror-symmetric properties of a chosen subset of peaks, the algorithm aims at uncovering mirror-symmetric peak patterns with respect to potential overlap. These peak patterns potentially represent molecule-specific multiplet signals and are thus related to each other. As shown in the respective results Section 6.4, several multiplet-like peak patterns are found for several example regions of simulated metabolite spectra and an example real-world spectrum.

7.2 Future Work

The proposed approach of *peak selection*, *parameter approximation* and *palindrome detection* demonstrates a way to transform an NMR spectrum into a set of relevant features, reducing the amount of data while simultaneously preserving the information content. By exploiting characteristic properties of the NMR signal, e.g. the fact that the peak distribution within a multiplet remains constant while chemical shift leads to a shifting in the frequency domain, the proposed methods

show a direction to automatically relate corresponding signals from different NMR spectra with each other. As already mentioned in the respective chapters, several questions arise and as yet answered:

1. How to automatically choose the degree of smoothing a given spectrum?
2. What is the impact of smoothing filters others than the mean filter prior to peak selection?
3. What are the limits of overlap detection by further derivation?

In the following, initial ideas and potential lines of investigation regarding the mentioned questions are further discussed.

7.2.1 Impact of further smoothing filters

Within this thesis, the *mean filter* is the only smoothing method considered, but other methods also exist, such as *order-statistics filters* (*median filter*) or *convolution* approaches. The reason of applying the *mean filter* is to remove local minima and maxima of the second derivative, which are additionally induced by the presence of noise and other distortions of the spectrum, on the background of equally distributed (white) noise. An inherent drawback of smoothing is, that there is no particular differentiation between signal and noise. As a result, shoulders may disappear after smoothing the spectrum. Based on the assumption that the corresponding peak triplets yield similar triplet scores to those of a predefined signal-free region (definition 5.2), an alternative smoothing procedure would be given by executing the following steps:

1. Identify peak triplets $\{w_{j,l}, w_{j,m}, w_{j,r}\}$ on the original spectrum without smoothing, and calculate the corresponding triplet scores.
2. Filter the triplets $\mathbf{p}_j$ by $score(\mathbf{p}_j) < threshold$
3. Calculate the envelopes $env_{\max}$ and $env_{\min}$ on the maxima and minima of the filtered triplets and adjust the resolution by linear interpolation

4. Adjust the original second derivative as

$$S''(w_{j,l}) = \frac{env_{\max}(w_{j,l}) + env_{\min}(w_{j,l})}{2}, \tag{7.1}$$

$$S''(w_{j,m}) = \frac{env_{\max}(w_{j,m}) + env_{\min}(w_{j,m})}{2}, \tag{7.2}$$

$$S''(w_{j,r}) = \frac{env_{\max}(w_{j,r}) + env_{\min}(w_{j,r})}{2}, \tag{7.3}$$

$$\tag{7.4}$$

   namely replace the original values by the average of the envelopes.

5. Calculate the spectrum S as the second order integral of S''.

By this, those parts of the spectrum that seem to maintain a low amount of noisy error are preserved and the distorted parts are smoothed with respect to the relative curvature of the neighboring datapoints.

7.2.2 Automated Smoothing

In extension to chapter 5, we additionally introduce the concept of triplet significance in units of σ as follows:

Definition 7.1 (Peak Triplet Significance θ)

With Q denoting the set of peak triplets found in blank signal, the significance θ of a peak triplet p with score $s(p)$ is defined as

$$s(p) = \bar{s} + \theta\,\sigma \quad \Leftrightarrow \quad \theta = \frac{s(p) - \bar{s}}{\sigma}, \tag{7.5}$$

$$\textit{with} \quad \bar{s} = \frac{1}{|Q|}\sum_{i \in Q} s(p_i) \quad \textit{and} \quad \sigma = \sqrt{\frac{1}{|Q|}\sum_{i \in Q}\big(s(p_i) - \bar{s}\big)^2}$$

as the mean and standard deviation score of all peak triplets in Q, respectively. Further,we call a triplet p to be accepted, if for a given significance threshold δ it holds $\theta \geq \delta$. The set of accepted triplets is in the remainder of this paper denoted as A.

For the purpose of noise reduction, we consider a slight variant of the *mean filter* procedure of chapter 5. In particular, we here consider the repeated averaging with a particular filter window of length 3, given as

$$smd(y_i) = \frac{1}{3}\sum_{k=i-1}^{i+1} y_{k'}, \quad \text{with } k' = \begin{cases} |k| + 2, & \text{for } k \leq 0, \\ 2n - k, & \text{for } k > n, \\ k, & \text{else.} \end{cases} \tag{7.6}$$

Similar to algorithm 3, the runtime complexity is given as $O(n+a)$ with b denoting the number of repeats.

Repeated averaging has already been considered decades ago, and is valued for its computational efficiency and easy-to-implement characteristics (Faes *et al.*, 1994). In the following we will briefly describe the effects of filtering, and refer for example to Cai (1988) for more detailed information.

Repeatedly smoothing by (7.6) steadily changes the coefficients of the filter window. For example, the first execution of (7.6) replaces each value y_i by $smd(y_i) = \frac{1}{3}(y_{i-1} + y_i + y_{i+1})$, after the second execution the value at index i is given as $smd(smd(y_i)) = \frac{1}{9}(y_{i-2} + 2\,y_{i-1} + 3\,y_i + 2\,y_{i+1} + y_{i+2})$, and analogously the triple execution of of (7.6) results in $\frac{1}{27}(y_{i-3} + 3\,y_{i-2} + 6\,y_{i-1} + 7\,y_i + 6\,y_{i+1} + 3\,y_{i+2} + y_{i+3})$. In fact, the weights after b times executing (7.6) equal the *trinomial coefficients*[1] obtained after expansion of $(1 + t + t^2)^b$ (see e.g. Merlini *et al.* (2002)). As a consequence of the *central limit theorem*, the coefficients discretely approximate the probability density function of the Gaussian distribution (Rice, 1995). Mean filtering thus corresponds to discrete convolution with a *Gaussian kernel*.

A heuristic non-parametric approach for automated Gaussian smoothing has been proposed by Lin *et al.* (1996), based on the change in the number and the maximal pair-wise distance of adjacent local maxima. The shown results though indicate that the proposed method seems to be rather unsuitable for datasets maintaining a higher diversity of peak positions as those considered in their paper.

[1]It might be worth mentioning that following Andrews (1990) no less than EULER found them worthy for a 20-page account Euler (1765).

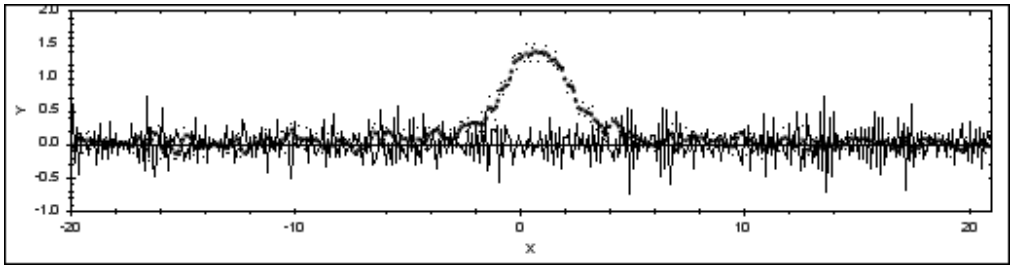
Vivo-Truyols & Schoenmakers (2006) proposed an automated smoothing approach based on the lag-one autocorrelation coefficient ρ_1, given as

$$\rho_1 = 1 - \frac{1}{2}\frac{\sum_{i=2}(y_i - y_{i-1})^2}{\sum_{i=1} y_i^2}\frac{n}{n-1}, \tag{7.7}$$

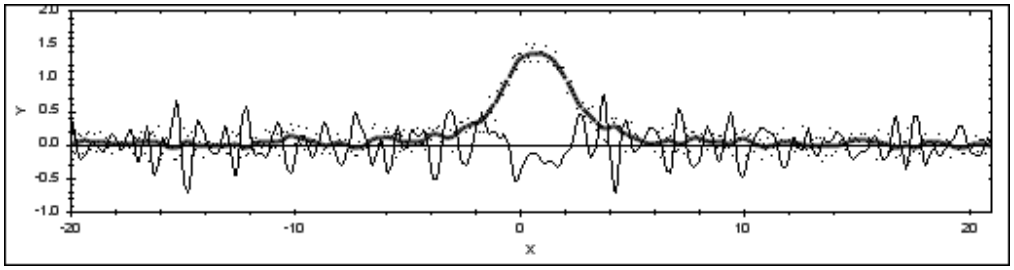
where n stands for the number of real-valued datapoints y_i. In rough summary, the smoothing degree is repeatedly increased, until the lag-one autocorrelation coefficient of the residual, namely the observed spectrum subtracted by the smoothed, is closest to the lag-one autocorrelation coefficient of blank signal. As a drawback, the method tends to excessively smoothen the considered spectra. In the following, an automated method for finding a proper smoothing degree is proposed based on changes in the curvature during the smoothing procedure of a given spectrum.

7.2.2.1 Method

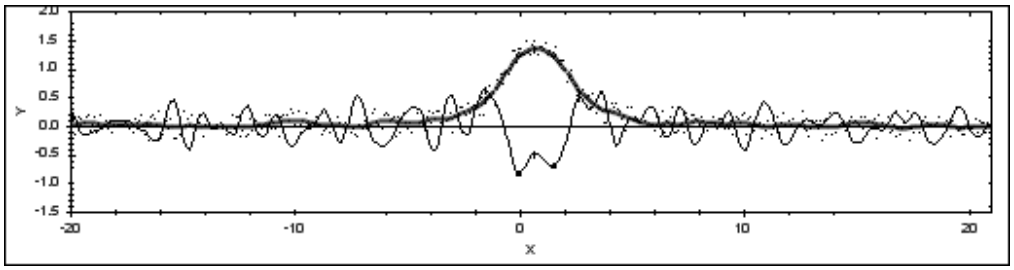
Repeatedly filtering the data by (7.6) allows to mitigate the impact of noise, but at some point also tends to merge curvatures emerging from distinct Lorentz functions. Figure 7.1 for example shows the effects of smoothing (thick dotted line) for varying smoothing repeats b of an example spectrum (dotted line) with uniformly distributed noise $U(-0.3, 0.3)$ and significance threshold $\delta = 6.0$. The spectrum is originally given as a sum of two Lorentz functions with parameters $\lambda = A = 1$, and with distance $d = |\omega_2 - \omega_1| = 1.5$ at a resolution $\frac{1}{\Delta x} = 10$. In the beginning ($b = 1$), the second derivative (thin solid line) is highly distorted due to the effects of noise (Fig. 7.1(a)). Increasing the degree of smoothing by repeatedly executing (7.6) allows to encapsulate two major clockwise-rotating curvatures of the spectrum as two adjacent peak triplets (Figs. 7.1(b) - 7.1(d)). Further smoothing leads to a merge of the two triplets, and only a single peak is observed for $b > 50$ (Fig. 7.1(e)). The aim thus is to find the number of smoothing repeats, which properly balances the trade-off between noise removal and signal preservation.



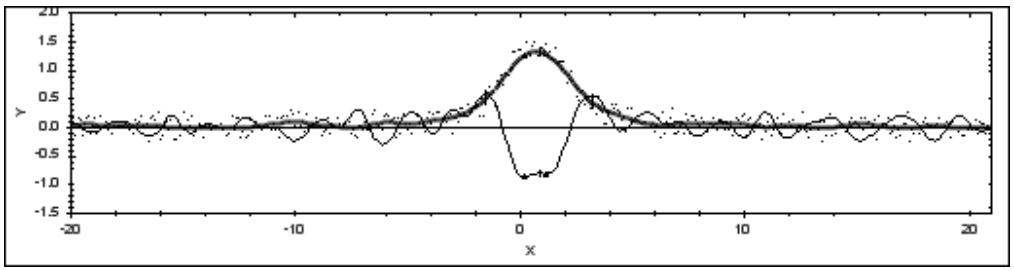
(a) $b = 1$



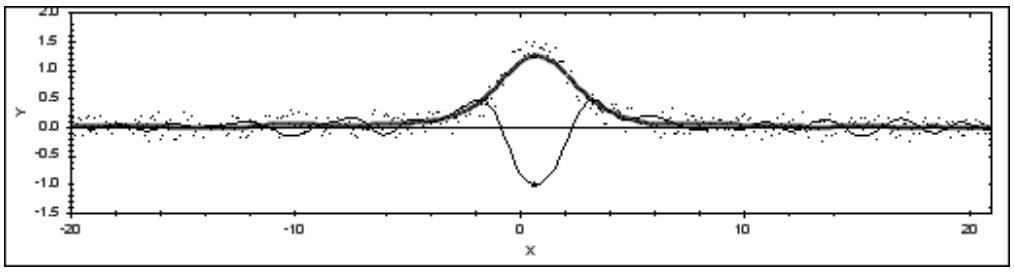
(b) $b = 10$



(c) $b = 25$

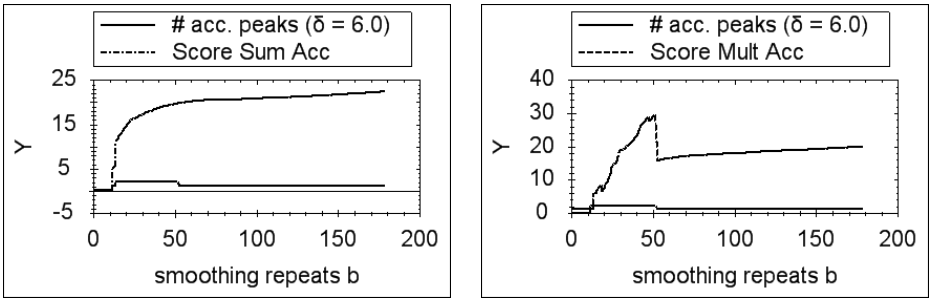


(d) $b = 50$



(e) $b = 100$

Figure 7.1: Smoothing effects on an example spectrum of two standard Lorentz functions. Shown are the original spectrum (dots) after adding white noise at an amplitude of 0.3, the smoothed spectrum (thick solid line) and the corresponding second derivative (thin solid line). b denotes the number of smoothing steps.



(a) f_{sum} (b) f_{mult}

Figure 7.2: The respective selection scores f_{sum} and f_{mult} (top lines) of Fig. 7.1 in comparison with the respective number of selected peaks (bottom line) for increasing smoothing repeats b.

For a given significance threshold δ, an intuitive approach for determining a reasonable degree of smoothing is given by maximizing the sum of scores

$$f_{sum}(A) = \sum_{j \in |A|} s(p_j) \tag{7.8}$$

for the set of accepted triplets A (see Def. 7.1). However, considering that the score of each triplet itself is given as a sum of second derivative values already, we may presume that merging of two *adjacent* triplets p_1 and p_2 has almost no effect on f_{sum} (compare Def. 5.2), written as

$$s(p_1) + s(p_2) \approx s(p_{1,2}) \tag{7.9}$$

with $p_{1,2}$ denoting the triplet received after merging of p_1 and p_2. To circumvent this problem, we may consider

$$s(p_1) \geq 2 \ \wedge \ s(p_2) \geq 2 \quad \Rightarrow \quad s(p_1)\, s(p_2) \ \geq \ s(p_1) + s(p_2) \approx s(p_{1,2}), \tag{7.10}$$

and the degree of smoothing can then be found by maximizing

$$f_{mult}(A) = \log \left(\prod_{i=1}^{|A|} \big(2 + s(p_i) - (\bar{s} + \delta\sigma)\big) \right) = \sum_{i=1}^{|A|} \log\big(2 + \Theta_i - \delta\big). \tag{7.11}$$

δ denotes a predefined significance threshold, and Θ_i denotes the significance of peak triplet p_i (see Def. 7.1). f_{mult} now allows to prevent significant peak triplets

p_i from being merged together, since by definition it holds $\Theta_i \geq \delta$ for all triplets $p_i \in A$. Figure 7.2 shows the corresponding selection scores f_{sum} and f_{mult} for the merging scenario of Fig. 7.1. In agreement with the assumption from above, the sum of scores of accepted triplets (f_{sum}, top line in fig. 7.2(a)) keeps increasing even after the two triplets have merged to a single one ($b > 50$). In contrast, an essential decrease can be observed for f_{mult} (top line in fig. 7.2(b)). A pseudo-code representation for an automated smoothing approach based on f_{mult} is given by Algorithm 9.

Algorithm 9

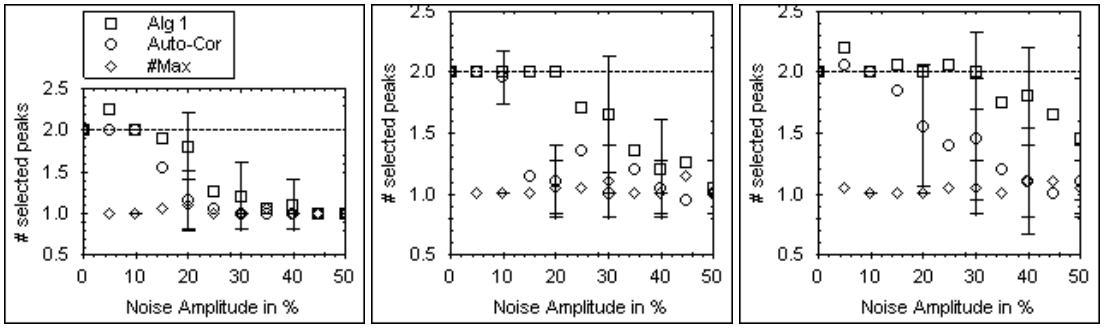
Input: Spectrum $\mathbf{a}$, blank signal $\mathbf{a}_{blank}$, significance threshold δ
Output: List of accepted peak triplets A^*

1: $best \leftarrow 0; last \leftarrow \infty; A^* \leftarrow \emptyset;$
2: **while** ($|Q| > 1$ and $|\rho_{1,blank} - \rho_{1,res}| \leq last$) **do**
3: Find set of triplets Q out of blank signal $\mathbf{a}_{blank}$;
4: $last \leftarrow |\rho_{1,blank} - \rho_{1,res}|$;
5: Find accepted peak triplets A on $\mathbf{a}$, given δ;
6: **if** ($f_{mult}(A) \geq best$) **then**
7: $best \leftarrow f_{mult}(A)$; $A^* \leftarrow A$;
8: **end if**
9: Apply (7.6) on all values in $\mathbf{a}$ and $\mathbf{a}_{blank}$;
10: **end while**
11: **return** A^*;

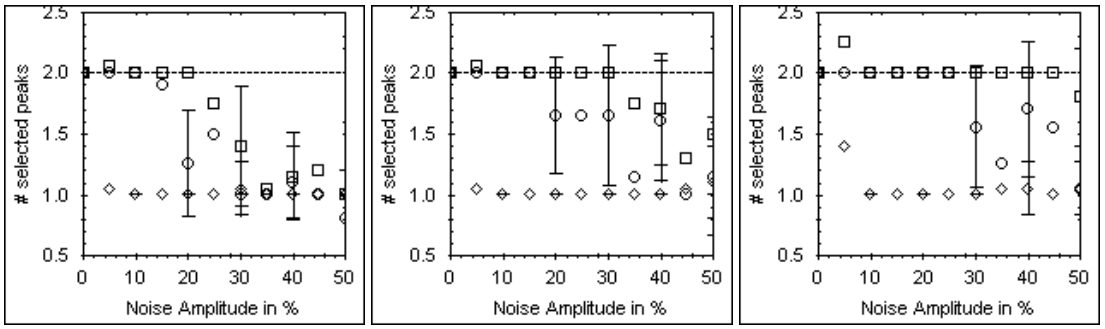
In summary, a given spectrum $\mathbf{a}$ is repeatedly smoothed, as long as blank triplets exist, and as long as the autocorrelation coefficient of the residual, $p_{1,res}$, approaches that of blank signal, $p_{1,blank}$. Note that the chosen degree of smoothing by Algorithm 9 is less or equal to that of Vivo-Truyols & Schoenmakers (2006). With b denoting the number of smoothing repeats needed for the execution of lines 3-11, with n denoting the number of datapoints in $\mathbf{a}$, and with m denoting the number of datapoints in the blank signal, Algorithm 9 has a total worst-case runtime of $O\big(\, b(n + m + |A| + |Q|) \,\big)$.

7.2.2.2 Results

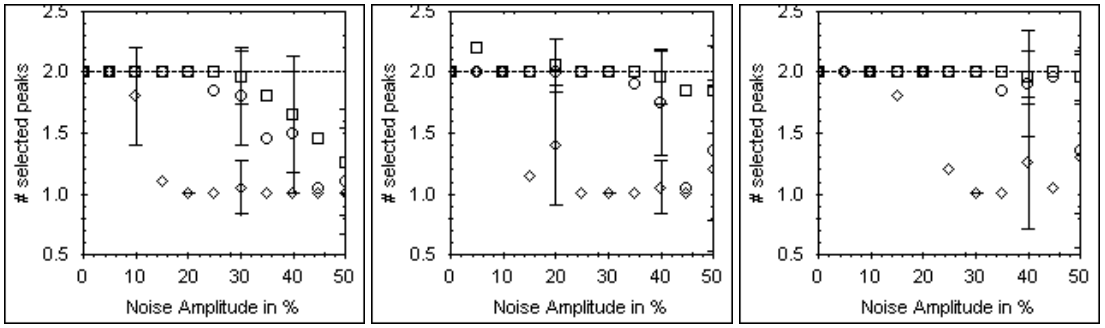
In this section, initial results of Algorithm 9 are presented, based on simulated spectra containing two *standard* Lorentz functions (3.12) with width and scale parameters $A = \lambda = 1$. With $d = |\omega_2 - \omega_1|$ denoting the distance between the two peaks, and with $r = \frac{1}{\Delta x}$ denoting the resolution of the spectrum, three peak



(a) $d = 1.2$, $r = 5$ (b) $d = 1.2$, $r = 10$ (c) $d = 1.2$, $r = 20$



(d) $d = 1.5$, $r = 5$ (e) $d = 1.5$, $r = 10$ (f) $d = 1.5$, $r = 20$



(g) $d = 2.0$, $r = 5$ (h) $d = 2.0$, $r = 10$ (i) $d = 2.0$, $r = 20$

Figure 7.3: Number of accepted peak triplets after the execution of Algorithm 9 for different scenarios and varying noise amplitudes. The length of the error bars equals two times the standard deviation out of 20 runs (see the text for more details).

distances $d \in \{1.2, 1.5, 2.0\}$ and three different spectrum resolutions $r \in \{5, 10, 20\}$ are considered, resulting in a total number of nine different smoothing scenarios. Each scenario is sampled 20 times, and uniformly distributed noise $U(-\frac{v}{100}, \frac{v}{100})$ is added to each datapoint, with noise amplitudes v in the range $0 \leq v \leq 50$. Triplets are found within the range $[\omega_1 - 5, \omega_2 + 5]$ out of a total spectral range of $[\omega_1 - 20, \omega_2 + 20]$. All evaluation runs are based on a significance threshold $\delta = 6.0$.

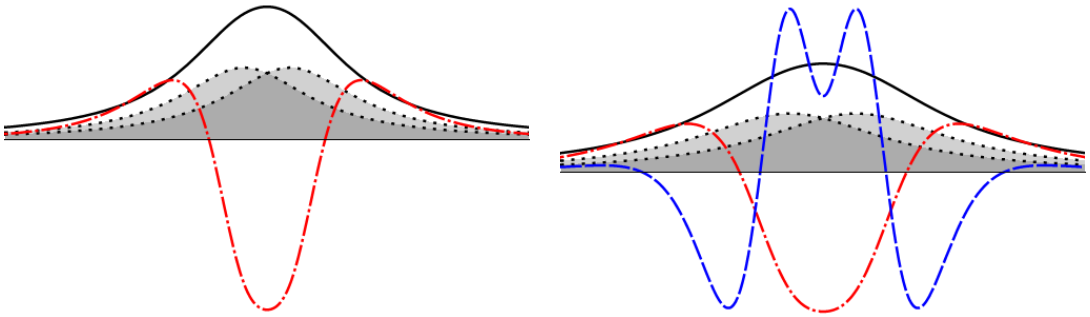
Fig. 7.3 shows the performance of Algorithm 9 on average out of 20 runs for each of the considered scenarios. The number of accepted peak triplets $|A^*|$ with maximal selection score f_{mult} is denoted as Alg 1, and shown as squares. In addition, the number of accepted triplets $|A|$ found after the last execution of line 5 in Algorithm 9 is denoted as Auto-Cor, and shown as circles. In a sense, these results represent the outcome of the lag-one autocorrelation approach of Vivo-Truyols & Schoenmakers (2006), and thus can be seen as baseline results, compared to which the impact of f_{mult} can be determined. In addition, for the degree of smoothing chosen by Algorithm 9, the number of local maxima with a maximal value higher than the spectrum average value is denoted as #Max, and shown as diamonds.

The figures generally show that it can be highly beneficial to identify peaks as curvatures of the spectrum rather than as maxima, since both methods *Auto-Cor* and Algorithm 9 leave the spectra in most cases with exactly one maximum after smoothing. Furthermore, it can also generally be observed that an increase in the peak distance d (from top to bottom in all columns) and also an increase in the resolution r (from left to right in all rows) have both a beneficial impact, i.e. the maximal noise amplitude, for which two peaks can still be identified, increases for increasing d or r or both.

An interesting result is given by the fact that Algorithm 9 is capable of identifying both peaks on average for even higher noise amplitudes and even smaller peak distances than *Auto-Cor* in all considered scenarios. Thus, at least for the considered datasets, maximizing (7.11) apparently comes to a better compromise between noise reduction and signal preservation than minimizing the autocorrelation distance only.

7.2.3 Overlap Detection by further derivation

For the task of peak selection, presented in Chapter 5, the curvature of the spectrum was investigated by observing local, negative minima of the discrete second derivative function. As shown by figure 7.4(a), this approach cannot detect arbitrarily overlapped peaks. In this context, a theoretical investigation concerning the relationship between the distance and number of distinct optima (local minima in the second, local maxima in the 4th, etc) in further derivative functions would be



(a) (b)

Figure 7.4: Two overlapping Lorentz functions (grey areas) with their sum (solid line), the second derivative (dot-dashed red line) and the 4th derivative (dashed blue line). The presence of two odistinct peaks can be observed by local maxima of the 4th derivative function.

beneficial, as for instance initially given by propositions 3.1 and 3.2 for the number of local maxima in the original function.

Figure 7.4 shows the overlap scenario of two Lorentz functions with equal HWHH $\lambda = 1$, area $A = 1$, and distance $\omega_2 - \omega_1 = \frac{\lambda}{\sqrt{3}}$ (compare figure 3.5 in Chapter 3). While this distance leads to only one local minimum in the second derivative (following proposition 3.1), one can clearly observe two local maxima in the 4th derivative (figure 7.4(b)). This suggests the potential of detecting smaller overlaps of peaks by higher order investigations of the curvature. Similar to the fact that a local minimum in the second derivative function indicates a locally maximal turn in clockwise direction of the original function, a maximum in the fourth derivative indicate a locally maximal turn in counter-clockwise direction of the second derivative. Thus, it denotes a potential overlap of second derivative minima. In a similar fashion, further derivative information might have the potential to improve overlap detection in general. One problem to consider in this context is the impact of noise and other distortions, since their occurrence already has a great impact on the second derivative as shown in the results section 5.1.3. Consequently, the interplay of smoothing filters and derivation is worth investigating. In particular, the loss of local optima resulting from smoothing the spectrum can potentially be counteracted by calculating higher order derivatives, in order to improve simultaneous peak and overlap detection.

References

Andrews, G. (1990). Euler's "exemplum memorabile inductionis fallacis" and q-trinomial coefficients. *Journal of the American Mathematical Society*, **3**, 653–669.

Asfour, A., Raoof, K. & Fournier, J.M. (2000). Nonlinear Identification of NMR Spin Systems by Adaptive Filtering. *Journal of Magnetic Resonance*, **145**, 37–51.

Bartholdi, E. & Ernst, R.R. (1973). Fourier spectroscopy and the causality principle. *Journal of Magnetic Resonance (1969)*, **11**, 9–19.

Bloch, F. (1946). Nuclear induction. *Phys. Rev.*, **70**, 460–474.

Bretthorst, G.L., Hutton, W.C., Garbow, J.R. & Ackerman, J. (2005). Exponential model selection (in nmr) using bayesian probability theory. *Concepts in Magnetic Resonance*, **27A**, 64–72.

Cai, L. (1988). Some notes on repeated averaging smoothing. In *Pattern Recognition*, 597–605.

Campbell, D., R.A.Pethrick & White, J. (????). *Polymer Characterization: Physical Techniques*. CRC Press, 2nd edn.

Chang, D., Weljie, A. & Newton, J. (2007). Leveraging Latent Information in NMR Spectra for Robust Predictive Models. In *Pacific Symposium on Biocomputing*, vol. 12, 115–126.

Cormen, T., Leiserson, C., Rivest, R. & Stein, C. (2001). *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2nd edn.

Davies, E. (2005). *Machine Vision: Theory, Algorithms and Practicalities*. Elsevier, 3rd edn.

Diehl, P., Skora, S. & Vogt, J. (1975). Automatic Analysis of NMR Spectra: An Alternative Approach. *J. Magn. Res.*, **19**, 67–82.

Dieterle, F., Ross, A. & Senn, H. (2006). Probabilistic Quotient Normaliza-

tion as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry*, **78**, 4281–4290.

Dijkstra, M., Roelofsen, H., Vonk, R.J. & Jansen, R.C. (2006). Peak quantification in surface-enhanced laser desorption/inization by using mixture models. *Proteomics*, **6**, 5106–5116.

Ebel, A., Dreher, W. & Leibfritz, D. (2006). Effects of zero-filling and apodization on spectral integrals in discrete Fourier-transform spectroscopy of noisy data. *Journal of Magnetic Resonance*, **182**, 330–338.

Ernst, R., Bodenhausen, G. & Wokaun, A. (1987). *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford University Press.

Euler, L. (1765). Observations analyticae. *Novi Commentarii Academiae Scientarum Petropolitanae*, **11**, 124–143, also in Volume 15 of *Opera Omnia*, Series 1, Teubner, p.50-69.

Faes, T.J.C., Govaerts, H.G., Tenvoorde, B.J. & Rompelman, O. (1994). Frequency synthesis of digital filters based on repeatedly applied unweighed moving average operations. *Med. & Biol. Eng. & Comput.*, **32**, 698–701.

Fiehn, O. (2002). Metabolomics the link between genotypes and phenotypes. *Plant Molecular Biology*, **48**, 155–171.

Forshed, J., Torgrip, R., Aberg, K., Karlberg, B., Lindberg, J. & Jacobsson, S. (2005). A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *J. Pharm. Biomed. Anal.*, **38**, 824–832.

Friebolin, H. (1999). *Basic one and two dimensional NMR spectroscopy*. Wiley-VCH, 3rd edn.

Giancaspro, C. & Comisarow, M.B. (1983). Exact Interpolation of Fourier Transform Spectra. *Applied Spectroscopy*, **37**, 153–166(14).

Gonzales, R. & Woods, R. (2002). *Digital Image Processing*. Prentice Hall, 2nd edn.

Goto, Y. (1998). Highly Accurate Frequency Interpolation of Apodized FFT Magnitude-Mode Spectra. *Applied Spectroscopy*, **52**, 134–138(5).

Gupta, R., Mittal, A., Narang, V. & Sung, W. (2004). Detection of palindromes in DNA sequences using periodicity transform. *IEEE International Workshop on Biomedical Circuits and Systems*, S2/7/INV– S2/720–3.

Hoult, D. & Lauterbur, P. (1979). The Sensitivity of the Zeugmatographic Experiment Involving Human Samples. *Journal of Magnetic Resonance*, **34**, 425.

Hoye, T.R., Hanson, P. & J.R., V. (1994). A Practical Guide to First-Order

Multiplet Analysis in lH NMR Spectroscopy. *J. Org. Chem.*, **59**, 4096–4103.

Jansen, J., Hoefsloot, H., van der Greef, J., Timmerman, M. & Smilde, A. (2005). Multilevel component analysis of time-resolved metabolic fingerprinting data. *Analytica Chimica Acta*, **530**, 173–183.

Jarvi, J., Nyman, S., Komu, M. & Forsstrom, J. (1997). A PC program for automatic analysis of NMR spectrum series. *Computer Methods and Programs in Biomedicine*, **52**, 213–222(10).

Karakaplan, M. (2007). Fitting Lorentzian peaks with evolutionary genetic algorithm based on stochastic search procedure. *Analytica Chimica Acta*, **587**, 235–239.

Knuth, D., Morris, J. & Pratt, V. (1977). Fast Pattern Matching in Strings. *SIAM J. Comput.*, **6**, 323–350.

Koh, H., Maddula, S., Lambert, J., Hergenrder, R. & Hildebrand, L. (2008). Feature selection by lorentzian peak reconstruction for ^{1}nmr post-processing. In *CBMS '08: Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems*, 608–613, IEEE Computer Society, Washington, DC, USA.

Koh, H., Maddula, S., Lambert, J., Hergenrder, R. & Hildebrand, L. (2009). An approach to automated frequency-domain feature extraction in nuclear magnetic resonance spectroscopy. *Journal of Magnetic Resonance*, **201**, 146–156.

Koh, H.W. & Hildebrand, L. (2009). A heuristic approach for the identification of palindromic peak sets. In *BIOCOMP*, 632–638.

Kolpakov, R. & Kucherov, G. (2008). Searching for Gapped Palindromes. *LNCS 5029: Combinatorial Pattern Matching*, **59**, 18–30.

Koradi, R., Billeter, M., Engeli, M., Güntert, P. & Wüthrich, K. (1998). Automated Peak Picking and Peak Integration in Macromolecular NMR Spectra Using AUTOPSY. *Journal of Magnetic Resonance*, **135**, 288–297.

Li, Y., Lacey, M.E., Sweedler, J.V. & Webb, A.G. (2003). Spectral restoration from low signal-to-noise, distorted NMR signals: application to hyphenated capillary electrophoresis-NMR. *Journal of Magnetic Resonance*, **162**, 133–140.

Lin, H., Wang, L. & Yang, S. (1996). Automatic determination of the spread parameter in gaussian smoothing. *Pattern Recognition Letters*, **17**, 1247–1252.

Lindon, J.C. & Nicholson, J.K. (2008). Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annual*

Review of Analytical Chemistry, **1**, 45–69.

Manacher, G. (1975). A New Linear-Time "On-Line" Algorithm for Finding the Smallest Initial Palindrome of a String. *J. Assoc. Comput. Mach.*, **22**, 346–351.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, **11**, 431–441.

Massaro, E., Viti, V., Guidoni, L. & Barone, P. (1989). High-resolution numerical filtering of nmr spectra. *Physics in Medicine and Biology*, **34**, 931–938.

Merlini, D., Sprugnoli, R. & Verri, M.C. (2002). Some statistics on dyck paths. *Journal of Statistical Planning and Inference*, **101**, 211–227.

Metz, K.R., Lam, M.M. & Webb, A.G. (2000). Reference deconvolution: a simple and effective method for resolution enhancement in nuclear magnetic resonance spectroscopy. *Magn. Reson.: Educ. J.*, **12**, 21–42.

Metzger, G., Patel, M. & Hu, X. (1996). Application of Genetic Algorithms to Spectral Quantification. *Journal of Magnetic Resonance, Series B*, **110**, 316–320(5).

Mierisov, S. & Ala-Korpela, M. (2001). Mr spectroscopy quantitation: a review of frequency domain methods. *NMR in Biomedicine*, **14**, 247–259.

Miller, M. & Greene, A. (1989). Maximum-likelihood estimation for nuclear magnetic resonance spectroscopy. *Journal of Magnetic Resonance (1969)*, **83**, 525–548.

Morris, G.A., Barjat, H. & Home, T.J. (1997). Reference deconvolution methods. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **31**, 197–257.

Moseley, H., Riaz, N., Aramini, J., Szyperski, T. & Montelione, G. (2004). A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra. *Journal of Magnetic Resonance*, **170**, 263–277.

Neil, J.J. & Bretthorst, G.L. (1993). On the use of bayesian probability theory for analysis of exponential decay date: An example taken from intravoxel incoherent motion experiments. *Magnetic Resonance in Medicine*, **29**, 642–647.

Nguyen, N., Huang, H., Oraintara, S. & Vo, A. (2009). Peak Detection in Mass Spectrometry by Gabor Filters and Envelope Analysis. *Journal of Bioinformatics and Computational Biology*, **7**, 547–569.

Nicholson, J.K., Lindon, J.C. & Holmes, E. (1999). 'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli

via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica*, **29**, 1181–1189.

Pons, J., Malliavin, T. & Delsuc, M. (1996). Gifa V. 4: A complete package for NMR data set processing. *Journal of Biomolecular NMR*, **8**, 445–452.

Popov, M. (1990). *Modern Nmr Techniques and Their Application in Chemistry*. CRC.

Potts, B.C.M., Deese, A.J., Stevens, G.J., Reily, M.D., Robertson, D.G. & Theiss, J. (2001). NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse. *Journal of Pharmaceutical and Biomedical Analysis*, **26**, 463–476.

Press, W., Flannery, B., Teukolsky, S. & Vetterling, W. (2007). *Numerical Recipes in C++*. Cambridge University. Press.

Purcell, E.M., Torrey, H.C. & Pound, R.V. (1946). Resonance absorption by nuclear magnetic moments in a solid. *Phys. Rev.*, **69**, 37–38.

Rabi, I.I., Zacharias, J.R., Millman, S. & Kusch, P. (1938). A new method of measuring nuclear magnetic moment. *Phys. Rev.*, **53**, 318.

Reusch, W. (1999). Introduction to Spectroscopy - Proton NMR Spectroscopy. Webmaterial of the Michigan State University, Department of Chemistry.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, 2nd edn.

Schorn, C. (2001). *NMR spectroscopy: data acquisition*. Wiley-VCH, Weinheim.

Spielman, D., Webb, P. & Macovski, A. (1988). A statistical framework for in vivo spectroscopic imaging. *Journal of Magnetic Resonance*, **79**, 66–77.

Stoyanova, R., Nicholls, A.W., Nicholson, J.K., Lindon, J.C. & Brown, T.R. (2004). Automatic alignment of individual peaks in large high-resolution spectral data sets. *Journal of Magnetic Resonance*, **170**, 329–335.

Torgrip, R.J.O., berg, M., Karlberg, B. & Jacobsson, S. (2003). Peak alignment using reduced set mapping. *J. Chemometrics*, **17**, 573–582.

van Vaals, J.J. & van Gerwen, P.H.J. (1990). Novel methods for automatic phase correction of NMR spectra. *Journal of Magnetic Resonance (1969)*, **86**, 127–147.

Vanhamme, L., Sundin, T., Hecke, P.V. & Huffel, S.V. (2000a). Mr spectroscopic quantitation: a review of time-domain methods. In *NMR in Biomedicine*, 14–233.

Vanhamme, L., Sundin, T., Van Hecke, P. & Van Huffel, S. (2000b). Mr spectroscopic quantitation: a review of time-domain methods. In *NMR in Biomedicine*, 14–233.

Verdun, F.R., Giancaspro, C. & Marshall, A.G. (1988). Effects of Noise, Time-Domain Damping, Zero-Filling and the FFT Algorithm on the Exact Interpolation of Fast Fourier Transform Spectra. *Applied Spectroscopy*, **42**, 715–721(7).

Viant, M.R. (2003). Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and Biophysical Research Communications*, **310**, 943–948.

Vivo-Truyols, G. & Schoenmakers, P.J. (2006). Automatic selection of optimal savitzky-golay smoothing. *Analytical Chemistry*, **78**, 4598–4608.

Wang, Y., Bollard, M.E., Keun, H., Antti, H., Beckonert, O., Ebbels, T.M., Lindon, J.C., Holmes, E., Tang, H. & Nicholson, J.K. (2003). Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning 1H nuclear magnetic resonance spectroscopy of liver tissues. *Analytical Biochemistry*, **323**, 26–32.

Weisstein, E. (1999). Levenberg-Marquardt Method, URL: "http://mathworld.wolfram.com/Levenberg-MarquardtMethod.html". MathWorld–A Wolfram Web Resource.

Weljie, A., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. (2006). Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. *Proteomics*, **78**, 4430–4442.

Wong, J.W.H., Cagney, G. & Hugh, M. (2005). SpecAlign–processing and alignment of mass spectra datasets. *Bioinformatics*, **21**, 2088–2090.

Yasui, Y., Pepe, M., Thompson, M., Adam, B., Wright, G., Qu, Y., Potter, J., Winget, M., Thornquist, M. & Feng, Z. (2003). A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostat*, **4**, 449–463.

Yu, W., Li, X., Liu, J., Wu, B., Williams, K.R. & Zhao, H. (2006). Multiple Peak Alignment in Sequential Data Analysis: A Scale-Space-Based Approach. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **3**, 208–219.