Journal of Magnetic Resonance 201 (2009) 146–156

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

An approach to automated frequency-domain feature extraction in nuclear magnetic resonance spectroscopy

Hyung-Won Koh [a,b,*], Sasidhar Maddula [a], Jörg Lambert [a], Roland Hergenröder [a], Lars Hildebrand [b]

[a] ISAS – Institute for Analytical Sciences, Bunsen-Kirchhoff-Str. 11, 44139 Dortmund, Germany
[b] Dortmund University of Technology, Computer Science Department, Chair 1, Otto-Hahn-Str. 16, 44221 Dortmund, Germany

ARTICLE INFO

Article history:
Received 8 May 2009
Revised 4 August 2009
Available online 16 September 2009

Keywords:
NMR quantitation
Peak selection
Metabolomics
Metabonomics
Systems biology

ABSTRACT

For the analysis of metabolite systems, nuclear magnetic resonance (NMR) spectroscopy has become an important quantitative monitoring technology. Automated quantitation methods are highly desired and mainly characterized by the tasks of model selection and parameter approximation. This paper proposes a promising automated two stage approach in the frequency-domain, in which signaling peaks are first identified and filtered from noise based on curvature properties of the spectrum, and then proportionally approximated based on the analytical solution of a Lorentz-function. Remarkably, in opposition to common least-squares approaches, the proposed approximation scheme does not rely on partial derivatives, and furthermore, the runtime is independent to the number of spectral datapoints. Simulations provide promising empirical evidence for successful peak selection and parameter approximation, with the results for the latter highly outperforming the LEVENBERG–MARQUARDT algorithm in terms of error minimization and robustness.

1. Introduction and related work

Regarding the complexity on the molecular level of living systems in general, high-throughput devices are essentially needed towards a comprehensive understanding of the underlying mechanisms and principles. In this context, the NMR technology has much potential due to its quantitative, non-invasive and non-destructive monitoring characteristics. However, analyzing and interpreting a series of NMR spectra is still a challenging problem, and satisfying automated approaches are still missing. In theory [1], an NMR spectrum can be described as a superposition of analytically known functions but with unknown parameters. Over decades, various quantitation methods have been proposed, aiming at the identification and approximation of these parameters either in the time-domain [2] or after Fourier-Transform or both [3]. However, a golden standard does not exist yet.

In practice, an NMR signal is distorted in many ways, ranging from inhomogeneous magnetization [4] and noise arising from different sources during the measurement [5] through artifacts of the Fourier-Transform [6,7] 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 [7,9,10], spectral noise filtering [11,12], phase correction and interpolation [6,13,14] and reference deconvolution [15–17], to provide a coarse overview.

However, regarding the analysis of multiple NMR spectra in e.g. high-throughput experiments, a different kind of problems arises as well. Especially high-field 1D NMR experiments of heterogeneous metabolite samples typically yield in more than tens of thousands of datapoints (dps) per spectrum, representing the superposition of hundreds of spectral components and more. Hence, data reduction methods are highly desirable to reduce costs of computation for further analysis. Various methods for the extraction of relevant features for multivariate analysis and classification have been proposed, and are in the following grouped into three classes:

(1) Spectral binning
(2) Targeted profiling
(3) Peak selection and parameter approximation

Spectral binning is often applied prior to principal component analysis and extensions of it [18–23]. It is considered to potentially mitigate effects of peak shifts and other variations by averaging over a certain number of datapoints [24]. 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

* Corresponding author. Address: ISAS – Institute for Analytical Sciences, Bunsen-Kirchhoff-Str. 11, 44139 Dortmund, NRW, Germany.
E-mail addresses: hyung-won.koh@udo.edu, koh@isas.de (H.-W. Koh).
URL: http://lsl-www.cs.uni-dortmund.de/cms/lang-de/koh-contact.html (H.-W. Koh).

1090-7807/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2009.09.003

source of origin at all, even if the differences are small. The results are dramatic loss of spectral resolution, obscured feature vectors, and hence potential misinterpretation of the data [20,25].

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* [26]. 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 spectrum. Subsequently, multivariate methods can be applied based on the achieved metabolite concentrations. The crucial part of this approach is the pattern assignment itself, which e.g. in [26] has been performed manually. 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.

Peak Selection and *Parameter Approximation*, also known as *Quantitation*, constitutes the third class of feature extraction methods. 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, either by directly operating on the time signal [2,4,27–29,35], or after Fourier-Transform [3], e.g. by exact interpolation [6,14], by the LEVENBERG–MARQUARDT algorithm [30–32], or based on genetic algorithms [33,34]. Thereby, a correct identification and separation of signal peaks from noise and artifacts, plays a key-role to successful approximation. Predominantly, methods concerning this task in the frequency-domain focus on the occurrence of local maxima [36,37]. By this, the identification of overlapped Lorentz-functions called "shoulders" is not well supported, as shown in previous work of Koh et al. [38]. This paper proposes an essentially extended approach called *Lorentzian Spectrum Reconstruction* for both peak identification and parameter approximation in order to automatically model an NMR spectrum as a superposition of single Lorentz-functions.

2. Definitions

After applying the Fourier Transformation (FT) [8], an NMR spectrum can ideally be written as [1]

$$S(\omega)=\int_0^\infty s(t)e^{-\mathbb{i}\omega t}dt=\sum_j^{|J|}e^{\mathbb{i}\varphi j}\left(a_j(\omega)+\mathbb{i}d_j(\omega)\right) \tag{1}$$

$$\text{with } a_j(\omega)=A_j\frac{T_{2_j}^*}{1+(\omega-\omega_j)^2T_{2_j}^{*2}} \text{ as the absorption signal} \tag{2}$$

$$\text{and } d_j(\omega)=A_j\frac{T_{2_j}^*(\omega-\omega_j)}{1+(\omega-\omega_j)^2T_{2_j}^{*2}} \text{ as the absorption signal} \tag{3}$$

with ω_j denoting the frequency (spectral position of the maximum), T_{2_j} denoting the decay rate, A_j, denoting the amplitude and φ_j denoting the phase of spectral component j. $\mathbb{i}$ stands for the complex number with $\mathbb{i}^2=-1$.

By assuming that phase correction has been applied properly, the dispersive part of the spectrum (Eq. (3)) can be neglected, resulting in the *absorption mode* signal (Eq. (2)) as a sum of Lorentz-functions:

$$S(\omega)=\sum_j^{|J|}Y_j(\omega),\ \text{with } Y_j(\omega)=\sum_j^{|J|}A_j\frac{\lambda_j}{\lambda_j^2+(\omega-\omega_j)^2} \tag{4}$$

where $\lambda_j=1/T_{2_j}^*$ stands for the half width at half height (HWHH). The height at the maximum of a Lorentz-function Y_j is given as $Y_j(\omega_j)=\frac{A_j}{\lambda_j}$, and the area of Y_j equals $A_j\pi$ (see Appendix A). In the remainder of this work, only positive values for λ and A are considered, i.e. $\lambda,A\geqslant 0$. Fig. 1(a) shows an example Lorentz-function.

Even in the ideal case of Eq. (4), the number of local maxima does not necessarily equal to the number of single Lorentz-functions due to effects of overlapping. In the following, a Lorentz-function is called *hidden* or simply a "shoulder", if there exists another Lorentz-function with the same nearest local maximum in the spectrum but with lesser distance. Fig. 1(b) illustrates an example peak overlap resulting in two *hidden* peaks.

With respect to the following sections, $\mathbf{w}=\{w_1,\ldots,w_n\}$ finally denotes the discrete list of frequencies of a spectrum containing n datapoints in descending order $w_1>\cdots>w_n$ according to the convention in NMR spectroscopy [39]. The corresponding intensities are denoted as $\{S(w_1),\ldots,S(w_n)\}$, and the second discrete derivatives of a spectrum is denoted as S'' (see Appendix B for more details).

3. Methods

In extension to Koh et al. [38], this paper proposes an approach for automated quantitation method called *Lorentzian*

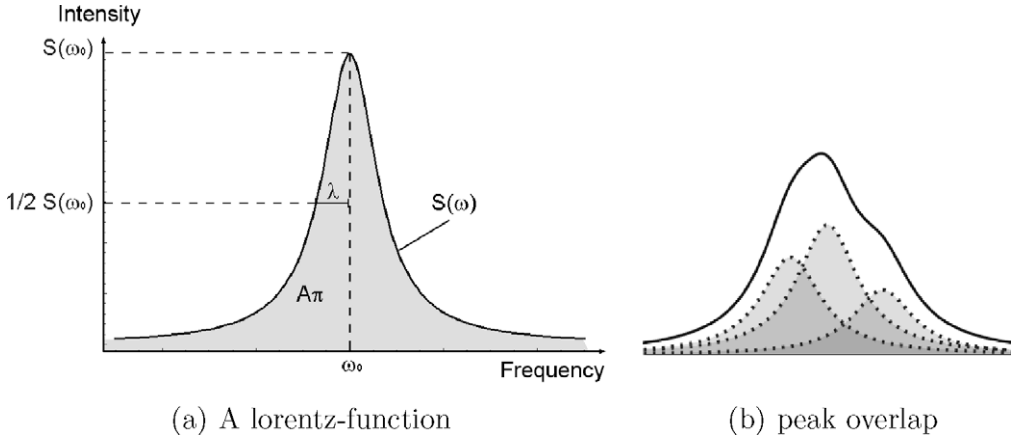

Fig. 1. The Lorentz-function from Eq. (4) as a spectral component of an NMR spectrum. (a) A single Lorentzian (solid line) at position ω_0 with (HWHH) λ and area $A\pi$ (grey area). (b) An example overlapping of three Lorentzians, resulting in a single maximum and two "shoulders" of the spectrum.

Spectrum Reconstruction, addressing both the problem of *peak selection*, namely the identification and selection of Lorentz-functions contained in an NMR spectrum, and *parameter approximation*, the approximation of the corresponding parameters. For clearness, we will discuss both parts separately.

3.1. Peak Selection

A trivial way to find peaks is to search for local maxima, a strategy, which to some extent has been followed by e.g. [36–38] and is commonly available as a basic feature in common NMR software. As mentioned in the previous section, the major drawback of this approach is that *hidden* peaks are detected only by investing additional computational effort, or even neglected at all.

A reasonable way to find all peaks instantly, even if they are overlapped, is to take into account the changes in the curvature of the spectrum by observing local minima of the second derivative S'', which correspond to locally maximal turns in clockwise direction of the original function, hence giving rise to the curvature property of interest. Fig. 2(a–c) show an example overlapping of Lorentz-functions of Eq. (4), the resulting spectrum and its derivatives. For clearness, the Lorentzians are equally scaled and shaped. One clearly observes, that the minima of the second derivative function (dashed line) are much better preserved than the maxima of the spectrum (solid line) (compare Fig. 2(a–c)). For the scenario shown, namely given two equally shaped and scaled Lorentz-functions Y_1 and Y_2, peak overlap resulting in the loss of a local maximum in their sum Y occurs for the distance $d(Y_1,Y_2) \leqslant \frac{2}{\sqrt{3}}\lambda$. Furthermore, the number of roots in the second derivative Y'' of the sum Y equals three for $d(Y_1,Y_2) = \frac{2}{\sqrt{3}}\lambda$, and two for $d(Y_1,Y_2) < \frac{2}{\sqrt{3}}\lambda$ (for the proofs, see Appendix C). In the following, each position w_i holding a negative local minimum of the second discrete derivative S'' is considered as a potential peak position only. Peak identification is based on searching for little "bumps" in the spectrum instead of climates only, intuitively speaking.

Due to spectrum distortions as mentioned in the previous section, a separation step is needed in order to distinguish real signal from noise and other artifacts. For this purpose, 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 root or local maximum in S'', either of which is closer positioned to w_m, more formally written as

$$w_m \Rightarrow S''(w_m) < 0 \wedge S''(w_{m-1}) > S''(w_m) < S''(w_{m+1}) \tag{5}$$

$$\wedge\ w_l = \min_{\substack{j \in \mathbf{w} \\ j<m}} \left(w_j - w_m \left| \left(\overbrace{S''(w_{j-1}) \geqslant 0 \wedge S''(w_j) < 0}^{\text{nearest root}} \right.\right.\right.$$
$$\left.\left.\left. \vee \underbrace{S''(w_{j-1}) \leqslant S''(w_j) > S''(w_{j+1})}_{\text{nearest local maximum}} \right) \right) \tag{6}$$

$$\wedge\ w_r = \min_{\substack{j \in \mathbf{w} \\ j<m}} \left(w_m - w_j \left| \left(\overbrace{S''(w_{j+1}) \geqslant 0 \wedge S''(w_j) < 0}^{\text{nearest root}} \right.\right.\right.$$
$$\left.\left.\left. \vee \underbrace{S''(w_{j-1}) < S''(w_j) \geqslant S''(w_{j+1})}_{\text{nearest local maximum}} \right) \right) \tag{7}$$

With $\mathbf{p} = (w_l, w_m, w_r)$ denoting a position triplet to a second derivative minimum at w_i, a *score* to distinguish between signal and noise is defined as

$$score(\mathbf{p}) = \frac{\min\left(\sum_{k=l}^{m} |S''(w_k)|, \sum_{k=m}^{r} |S''(w_k)| \right)}{\max\ score} \tag{8}$$

with max *score* as a normalizing term. The main idea of Eq. (8) 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 in 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 for frequencies below −0.5 ppm or above 10 ppm (parts per million). The following algorithm describes the peak selection approach:

Algorithm 1. (Curvature-Based Peak Selection)

Input: Spectrum S, second derivative S'', a signal-free region $R \subset S$, threshold parameter δ
Output: Filtered list of peak triplets L

```
L, L' = ∅
Find all peak triplets, given S'' and add to L'
Compute scores of each triplet p of L'
Compute mean_score and sd_score given L', R
for j = 1 to |L'| do
    if score (p_j) ⩾ mean_score + δ × sd_score then
        Add p_j to L
    end if
end for
return L
```

Algorithm 1 has a runtime of $O(n)$ with n denoting the number of spectral datapoints, and can therefore be considered as *runtime-efficient* (for more details, see Appendix D).

3.2. Parameter Approximation

Once the set of peak triplets is known, the corresponding parameter set needs to be fitted in accordance with the spectrum. By making use of the analytical solutions of the parameters, the

(a) $d = \frac{2.5}{\sqrt{3}}\lambda$ (b) $d = \frac{2}{\sqrt{3}}\lambda$ (c) $d = \frac{1.5}{\sqrt{3}}\lambda$ (d) $d = \frac{1}{\sqrt{3}}\lambda$

Fig. 2. Two overlapping Lorentz-functions (grey area) with equal HWHH λ and area A, the corresponding sum (solid line), the resulting 1st (dotted line) and second derivative (dashed line), with distances varying as indicated.

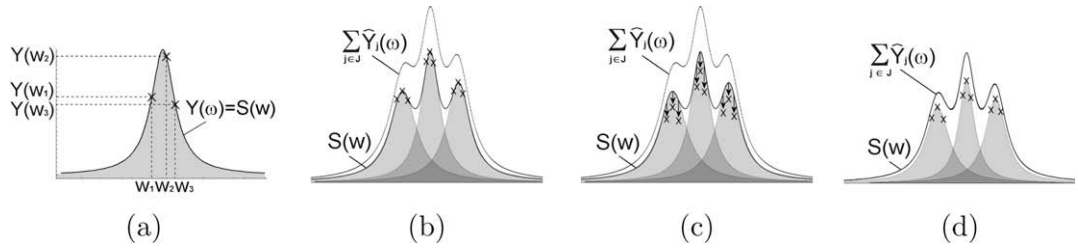

Fig. 3. Parameter approximation by rule of proportion. (a) Three points for a peak are chosen to directly calculate the Lorentzian parameters. (b) Initial guess exceeds a superimposed spectrum. (c) Iteratively adjust the heights by rule of proportion (Eq. (10)) and recalculate the parameters. (d) Result after three iterates.

approximation takes place iteratively by the rule of proportion, as described in the following:

For a spectrum S containing a single peak ($|J| = 1$), it holds $Y(w) = S(w)$, and the parameters can be achieved directly by solving the quadratic equation system

$$\begin{aligned} & S(w_1) = Y(w_1) = A\frac{\lambda}{\lambda^2 + (w_1 - \omega_0)^2} \\ \wedge\ & S(w_2) = Y(w_2) = A\frac{\lambda}{\lambda^2 + (w_2 - \omega_0)^2} \\ \wedge\ & S(w_3) = Y(w_3) = A\frac{\lambda}{\lambda^2 + (w_3 - \omega_0)^2} \end{aligned} \tag{9}$$

for the parameters A, λ and ω_0. This results in polynomial expressions for the parameters, depending on the positions $\{w_1, w_2, w_3\}$ and the corresponding heights $\{Y(w_1), Y(w_2), Y(w_3)\}$ only (see Appendix E for supporting information and [6] for a similar approach). Fig. 3(a) elucidates this situation. Note, that these expressions need to be calculated only once, and can be applied for arbitrary point triplets of a Lorentz-function.

Commonly a real-world spectrum contains more than only one spectral component, and all peak-specific intensity values $Y_j(w_k)$, $w_k \in \mathbf{w}$ of a peak j are unknown a priori (Eq. (4), Fig. 3(b)). Let in the following $\widehat{Y}_j^{(i)}$ denote the model of peak j at iteration step i, and let x_j stand for the elements of the jth peak triplet $\{w_{j,l}, w_{j,m}, w_{j,r}\}$. With the initial guess $\widehat{Y}_j^{(0)}(x_j) = S(x_j)$, the corresponding peak parameters can then be approximated based on the assumption, that the proportion between the new value $\widehat{Y}_j^{(i)}(x_j)$ and the old value $\widehat{Y}_j^{(i-1)}(x_j)$ is similar to the proportion between the actual spectrum and the current sum, more formally written as

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

which directly leads to the new value $\widehat{Y}_j^{(i)}(x_j)$ as

$$\widehat{Y}_j^{(i)}(x_j) = \widehat{Y}_j^{(i-1)}(x_j) \times \frac{S(x_j)}{\sum_{l\in J}\widehat{Y}_l^{(i-1)}(x_j)} \tag{10}$$

(see Fig. 3(c and d)). The corresponding approximation algorithm is described in Algorithm 2.

Algorithm 2 (Proportional Approximation)

Input: Spectrum S, List L of peak triplets
Output: Approximated Parameter Set J
Initial guess: $\widehat{Y}_j^0(x_j) = S(x_j)$ for all $j \in L$
for $i = 1$ to K **do**
  for $j = 1$ to $|L|$ **do**
    if $\widehat{Y}_l^{(i-1)}(x_j) \leqslant S(x_j)$ for all $l \in J$ **then**
      Calculate the sums $\sum_{l\in J}\widehat{Y}_l^{(i-1)}(x_j)$
      Calculate new heights $\widehat{Y}_j^{i}(x_j)$ by Eq. (10)
      Calculate new parameters $\omega_j^{(i)}$, $\lambda_j^{(i)}$, $A_j^{(i)}$ by solutions of Eq. System 9
    end if
  end for
end for
return J

The corresponding worst-case runtime lies in $O(K \times |J|^2)$, and remarkably, it is independent of the number of spectral datapoints of a spectrum (for more details, see Appendix F).

In summary, a spectrum can be translated into its distinct set of Lorentzian parameters by the sequential execution of Algorithms 1 and 2, as described in the following algorithm:

Algorithm 3 (Lorentzian Spectrum Reconstruction II)

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 1) given S
2: Filter L' given R and parameter δ (Algorithm 1) to receive L
3: Approximate Lorentzian parameter set to receive J (Algorithm 2), given S, L and K
4: **return** J

The worst-case runtime of Algorithm 3 lies in $O(n + K \times |J|^2)$, where n again denotes the number of spectral datapoints.

4. Results and discussion

Since the outcome of Algorithm 3 is highly dependent on the outcome of the peak selection step, the results of Algorithm 1 are shown and discussed first.

4.1. Peak Selection results

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

$$\omega_i = \omega_{i-1} + \lambda + u\lambda \tag{11}$$

beginning with $\omega_0 = 20\ \lambda$ and $u \sim [0,1]$ as a uniformly distributed random number. Spectrum distortions have been simulated by adding a uniformly distributed random number v out of the range $v \sim [0, r]$[1] to each spectral datapoint as

$$r = \frac{\left(\frac{A}{\lambda}\right)}{\rho} \tag{12}$$

ρ specifies the ratio between the common peak height $\frac{A}{\lambda}$ and maximal distortion amplitude r, with a high value implying a low level of noise. The noise region (compare Section 3.1) was set to the ranges $[0, 10\lambda]$ and $[\omega_{99} + 10_\lambda, \omega_{99} + 20\lambda]$.

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. In-

[1] Only positive distortions are considered to guarantee positive height values and therefore positive width and area parameters of a Lorentz-function (compare Eq. (4)). In case of a real-world spectrum, only peak triplets containing positive values are further considered by now.

deed, resulting in a hidden peak for a distance $d(Y_1, Y_2) \leqslant \frac{2}{\sqrt{3}}\lambda$, as shown in Appendix C, applies only for the scenario of two Lorentz-functions as shown in Fig. 2, but with observing that the gradient decreases roughly to the power of three for increasing distances to the maximum (Eq. (A.1) in Appendix A), the contribution of further Lorentzians to the overlapping of a consecutive pair of peaks can be assumed to be more or less constant, and is therefore neglected. On the basis of this pairwise simplification, a rough estimate for the expected number of hidden peaks $E(\#hidden)$ in a sum of 100 Lorentz-functions follows as

$$E(\#hidden) \approx 100\left(\frac{2}{\sqrt{3}} - 1\right) \approx 15 \tag{13}$$

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 $\frac{2}{\sqrt{3}} - 1$. At the same time, the generated spectra are very likely to contain both *maximum* and *hidden* peaks as well, since the estimated probability for achieving a spectrum with 100 *maximum* peaks only is then also given as $\left(2 - \frac{2}{\sqrt{3}}\right)^{100} \approx 5 * 10^{-8}$.

Fig. 4 shows the selection result for the simulated spectra. For the unfiltered scenario, the number of found peaks exceeds the correct number considerably, as can be observed in Fig. 4(a). The reason lies in the occurrence of additional minima and maxima of the second derivative due to the artificial distortions, as shown in Fig. 4(d) for an example subregion of the simulated spectra. In order to mitigate these distortion effects, the signal is smoothened by applying a mean filter, as shown in Fig. 4(e) for a two times consecutively executed (second order) two-point mean filter (2,2-filter), and in Fig. 4(f) for a three times consecutively executed (third order) three-point mean filter (3,3-filter). Fig. 4(g) shows, that the peak scores of the unfiltered scenario are unfavorably distributed in terms of separating the signal peaks based on δ. After smoothing the spectra, a clear separation of peak and noise triplet scores can be observed in Fig. 4(h and i), leading to a drastic improvement in the selected number of peaks, as can be seen in Fig. 4(e and f), respectively. It shall be noted, that a correct number of selected peaks does not necessarily imply a correct selection of peaks, and this will be discussed further in the following section.

4.2. Parameter Approximation results

In the following, the results of the proposed Proportional Approximation (*PA*) algorithm (Algorithm 2) are shown and discussed in comparison to the LEVENBERG–MARQUARDT algorithm, in the remainder denoted as *LM*. The former has been implemented by the first author in the programming language *C#*, and for the latter algorithm the software *Mathematica 6.0* was used [40]. The evaluation is based on 20 spectra again, but now with each containing only 20 Lorentz-functions for reasons of lesser time consumption. For the purpose of better reflecting real-world conditions, the shape-parameter λ_i and the scale-parameter A_i of each Lorentz-function Y_i have been varied uniformly in the ranges $\lambda_i \in [0.002, 0.005]$ and $A_i \in [50, 100]$, and the positioning of each peak took place as

$$\omega_i = \omega_{i-1} + v \;\; \max(\lambda_i, \lambda_{i-1})$$

with $\omega_0 = 0.0$ and $v \sim [1.5, 2]$ as a uniformly distributed random number. 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. Distortions of the spectrum are introduced relative

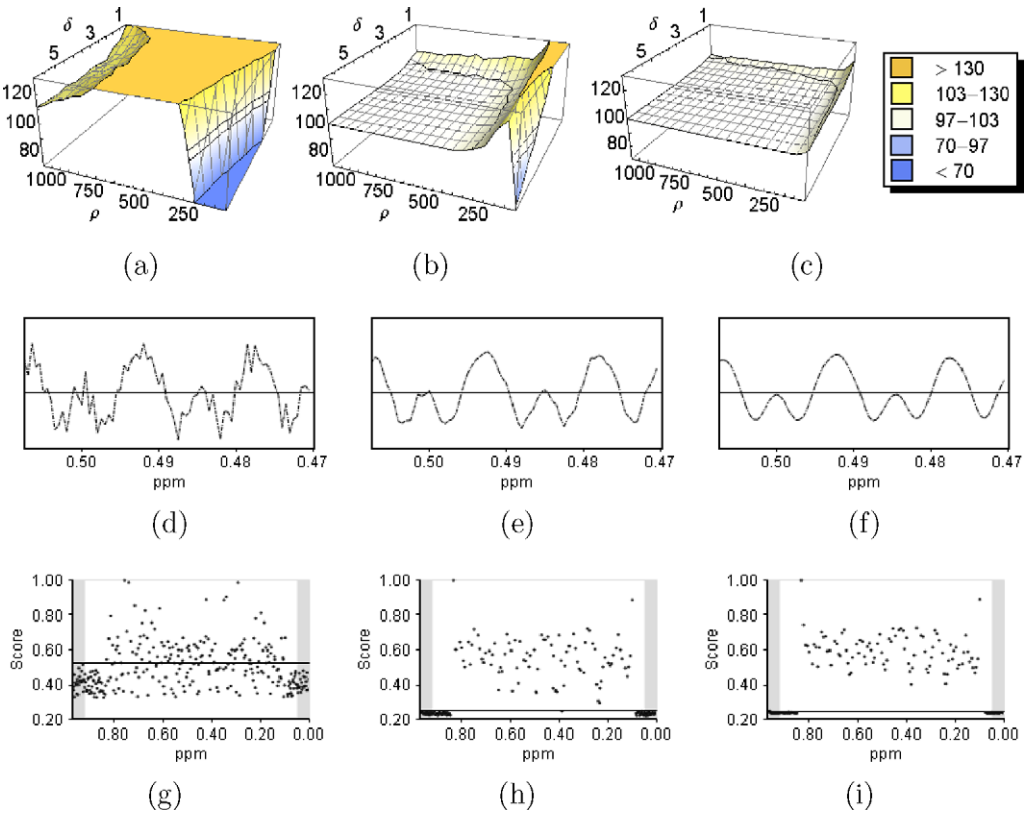

Fig. 4. Simulation results for the peak selection. Top: Number of selected peaks on average out of 20 spectra for varying distortion levels ρ and picking thresholds 5 of (a) raw data, (b) after 2,2-filtering and (c) after 3,3-filtering. Center: Example second derivative in a subregion of (a) raw data, (b) after 2,2-filtering and (c) after 3,3-filtering. Bottom: Corresponding triplet scores for ρ = 200. The grey areas denote the chosen signal-free region, and the solid lines indicate the selection score for δ = 3.0 on (g) raw data, (h) after 2,2-filtering and (i) after 3,3-filtering (see the text for more details).

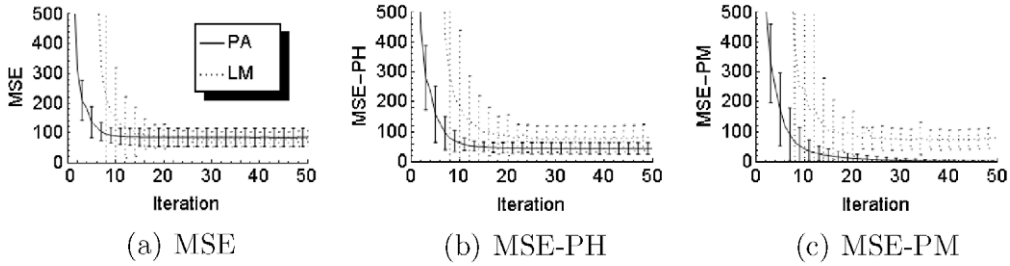

Fig. 5. Mean squared error of the proposed method *PA* (solid line) and for LEVENBERG–MARQUARDT (dotted line) for $\rho = 500$. The length of the error bars equals two times the standard deviation on average out of 20 spectra.

to the maximal peak height by a uniformly distributed random number v out of the range $v \in [0,r]$, with r given as

$$r = \frac{\max_{j\in J}\left(\frac{A_j}{\lambda_j}\right)}{\rho} \tag{14}$$

For evaluation purposes, three different noise ratios ρ are considered: $\rho = 1000$, $\rho = 500$ and $\rho = 200$, and a 5,3-filter (five times execution of a three-point mean filter) is applied to smoothen the spectra. Subsequently, the peaks are found by the *Curvature-Based Peak Selection* (Algorithm 1) 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 $\widehat{\omega}_j$, λ_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

$$\frac{1}{n}\sum_{i=1}^{n}\left(S(w_i) - \widehat{Y}(w_i)\right)^2 \tag{15}$$

(2) *Mean Squared Error at the Peak Hills* (MSE-PH) accounting for the mean squared error within the peak intervals $[w_{l_j}, w_{r_j}]$ Eqs. (6) and (7), given as

$$\frac{1}{|J|}\sum_{j=1}^{|J|}\left(\frac{1}{r_j - l_j}\sum_{i=l_j}^{r_j}\left(S(w_i) - \widehat{Y}(w_i)\right)^2\right) \tag{16}$$

(3) *Mean Squared Error at the Peak Maxima* (MSE-PM) accounting for the squared error at each discrete peak maximum position w_{mj} Eq. (5), given as

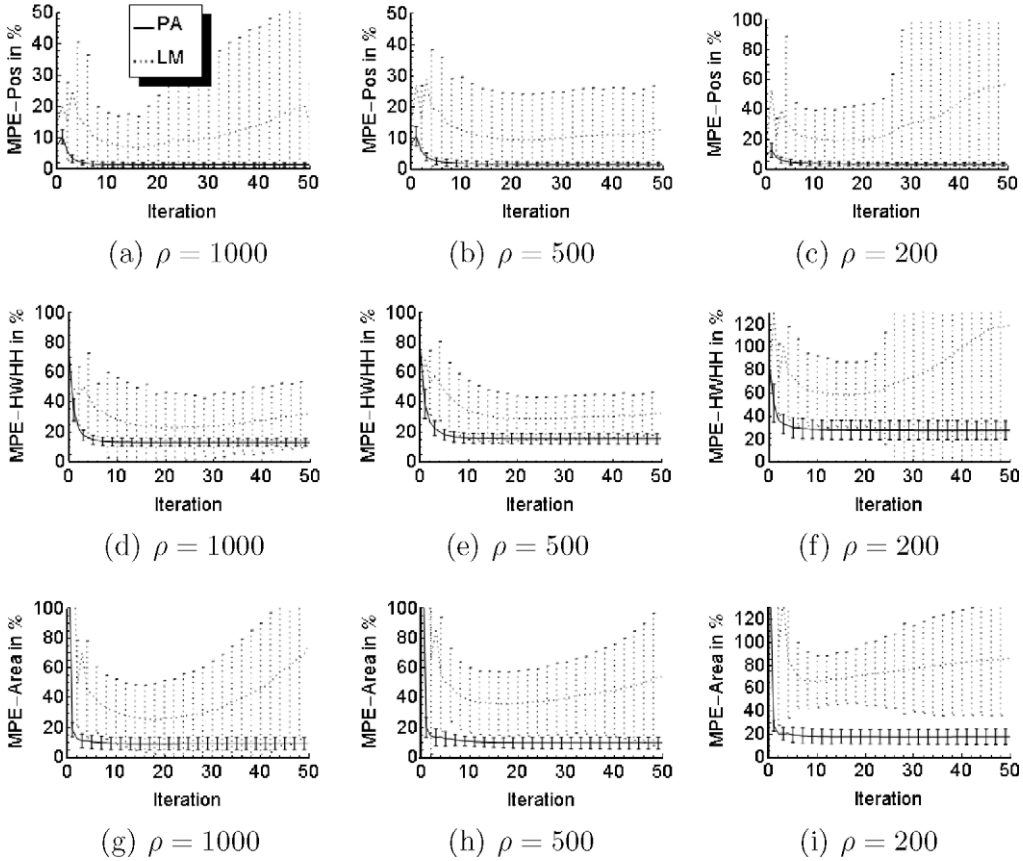

Fig. 6. Mean percentage error of the parameters for the proposed method *PA* (solid line) and for LEVENBERG–MARQUARDT (dotted line) along 50 iterations, top: position (MPE-Pos), center: HWHH (MPE-HWHH), bottom: area (MPE-Area). The length of the error bars equals two times the standard deviation on average out of 20 spectra.

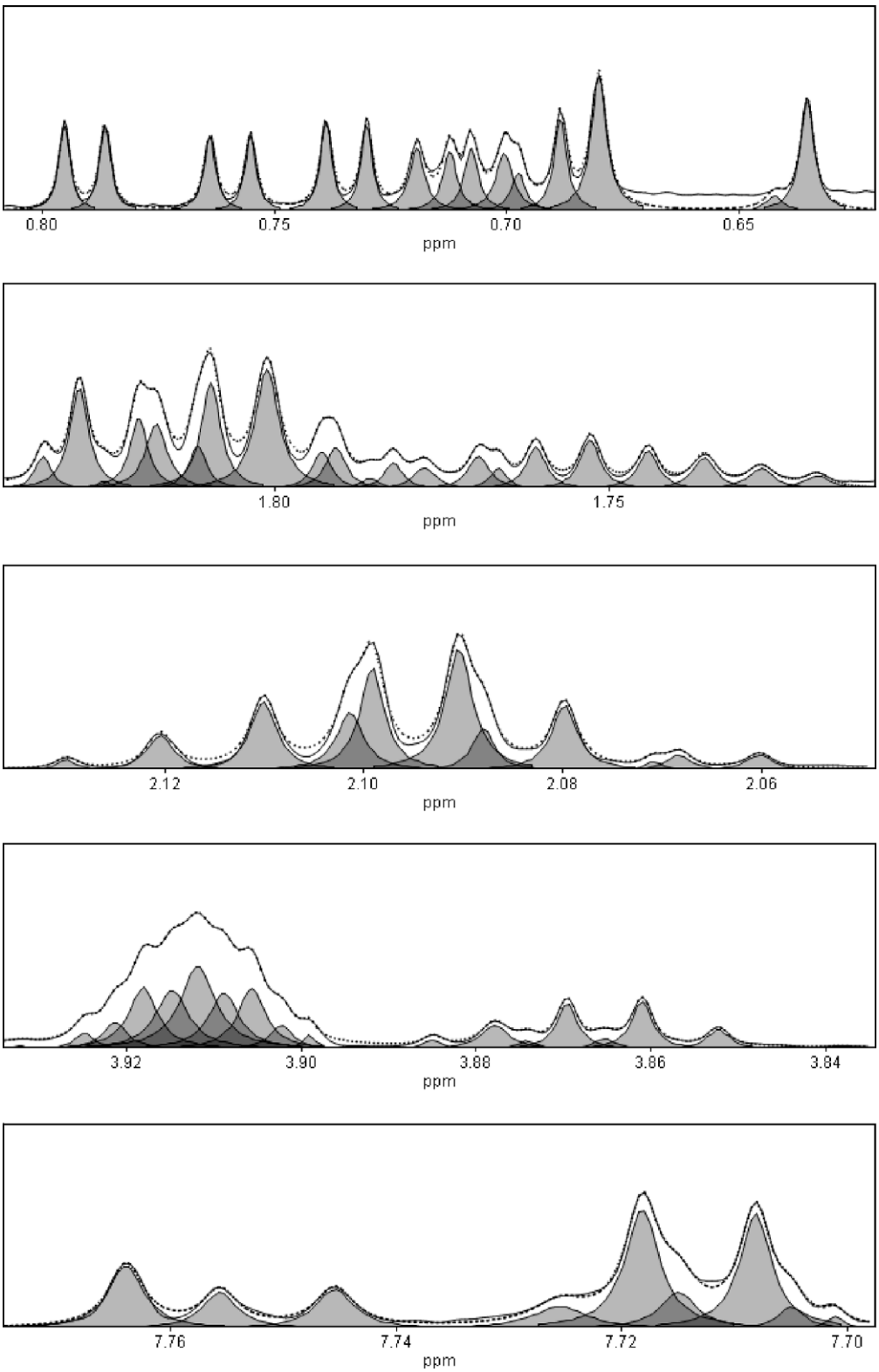

Fig. 7. Portions of an example real-world spectrum (solid line), the fitted Lorentz-functions by Algorithm 2 (grey areas) and the corresponding sum (dotted line) after 10 iteration steps.

$$\frac{1}{|J|}\sum_{j=1}^{|J|}\left(S(w_{m_j})-\widehat{Y}(w_{m_j})\right)^2 \tag{17}$$

(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

$$\frac{100}{|J|}\sum_{j=1}^{|J|}\left|\frac{(\widehat{\omega}_j-\omega_j)}{\lambda_j}\right| \tag{18}$$

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

$$\frac{100}{|J|}\sum_{j=1}^{|J|}\left|1-\frac{\widehat{\lambda}_j}{\lambda_j}\right| \tag{19}$$

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

$$\frac{100}{|J|}\sum_{j=1}^{|J|}\left|1-\frac{\widehat{A}_j}{A_j}\right| \tag{20}$$

For *LM*, the parameters have been initialized with the initial parameters $\widehat{\omega}_i^{(0)}$, $\widehat{\lambda}_i^{(0)}$ and $\widehat{A}_i^{(0)}$, found in lines 2–4 of Algorithm 2. Since the outcome of *LM* is highly dependent on the parameter initialization, five additional runs with decreased parameters $\widehat{\lambda}_i$ and $\widehat{A}_i$ as

$$\widehat{\omega}_i=\widehat{\omega}_i^{(0)}+u_\omega, \tag{21}$$

$$\widehat{\lambda}_i=\widehat{\lambda}_i^{(0)}*u_\lambda, \tag{22}$$

$$\widehat{A}_i=\widehat{A}_i^{(0)}*u_A, \tag{23}$$

with uniformly distributed random variables $u_\omega\sim[-0.001,0.001]$, $u_\lambda\sim[0.5,1.0]$ and $u_A\sim[0.5,1.0]$ are considered as well. The parameters are decreased to address the fact that the spectrum is always exceeded by the initial guess $\widehat{Y}_j^{(0)}(x_j)=S(x_j)$. For each spectrum, the fit which minimizes *MSE* (Eq. (15)) out of the initial and the additional five runs with decreased parameters as described are further considered for evaluation.

Fig. 5 shows the mean squared error performance on average out of 20 spectra for a noise ratio of $\rho=500$ (see Eq. (14)) and 50 iterations.[2] 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 Fig. 5(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 datapoints. In addition, the results of *PA* are more robust than *LM*, as can be observed by comparing the length of the error bars of the two approaches.

Fig. 6 shows the mean percentage error in the Lorentzian parameters on average out of 20 spectra of *PA* and *LM* for different noise ratios $\rho=1000$, 500 and 200 (see Eq. (14)). *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 Fig. 6(a–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 1) is indeed capable of not only selecting the number of peaks correctly, but also to identify the set of peaks with high accuracy. With an average standard deviation of ca. 1–3% (10–50%) in the position, ca. 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 behavior than *LM*.

Fig. 7 shows some regions of a real-world metabolite NMR spectrum of a human colon cancer cell line after applying the proposed automated reconstruction approach. The spectrum is obtained on a VARIAN INOVA 800 spectrometer operating at 799.77 MHz. The data was acquired using a 13 kHz spectral width, 22,114 datapoints, and 1.7 s of acquisition time. Zero filling was performed resulting in 32,768 (2^{15}) data points, and the FFT algorithm was applied without any line broadening. Baseline and phase correction were performed by the software ACD/SpecManager 6.0. After 3,3-filtering and selecting the noise regions to the ranges [12.8, 10.0] and $[-1,-3.4]$ (in ppm), 531 peaks were selected by the proposed Algorithm 1 with picking threshold $\delta=6.0$. The execution of Algorithm 2 was finished after ca. 9 s on a dual-core 1.66 GHz laptop with 1 GByte RAM and OS Windows XP. The program is freely available upon request.

5. Conclusions and outlook

In this work, 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 Lorentz-functions, a runtime-efficient selection procedure (Algorithm 1) is proposed, being able to simultaneously detect hidden and unhidden peaks at once. Simulations empirically demonstrate, that the proposed approach in conjunction with mean filtering is able to find the set of signaling Lorentz-functions properly for a broad range of varying noise amplitudes and picking thresholds. A subsequent parameter approximation scheme (Algorithm 2) is proposed, exploiting the analytical solution of a single Lorentz-function and adjusting the parameters proportionally in each step of the iteration. Empirical studies show, that the proposed approach highly outperforms the LEVENBERG–MARQUARDT algorithm in terms of minimal error and robustness in general. Especially the results for the position and area parameters are highly promising.

Several potentially interesting questions are still waiting to be answered, i.e. concerning the minimal spectral distance of a pair of Lorentz-functions to occur as a distinct second derivative minimum, or concerning convergence properties of the proportional fitting procedure. To directly enable a reliable multivariate analysis for, 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 this context, it can be concluded, that the proposed approach is in general highly suitable for solving the task of automated NMR feature extraction in terms of model selection and parameter approximation. As a final remark, 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.

Appendix A

For a Lorentz-function $Y(\omega)$ given as

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

with maximum position ω, HWHH λ and scaling factor A, the first three derivatives of Y are given as

[2] Due to lesser space occupation, only the results for $\rho=500$ are shown. The results for $\rho=1000$ and $\rho=200$ do not differ significantly.

$$Y'(\omega) = -\frac{2A\lambda(\omega-\omega_0)}{\left(\lambda^2+(\omega-\omega_0)^2\right)^2} \tag{A.1}$$

$$Y''(\omega) = \frac{8A\lambda(\omega-\omega_0)^2}{\left(\lambda^2+(\omega-\omega_0)^2\right)^3} - \frac{2A\lambda}{\left(\lambda^2+(\omega-\omega_0)^2\right)^2} \tag{A.2}$$

$$Y'''(\omega) = \frac{-48A\lambda(\omega-\omega_0)^3}{\left(\lambda^2+(\omega-\omega_0)^2\right)^4} + \frac{24A\lambda(\omega-\omega_0)}{\left(\lambda^2+(\omega-\omega_0)^2\right)^3} \tag{A.3}$$

The corresponding roots of the derivatives are given as

$$Y'(\omega)=0 \Longleftrightarrow \omega=\omega_0 \tag{A.4}$$

$$Y''(\omega)=0 \Longleftrightarrow \omega=\omega_0\pm\frac{1}{\sqrt{3}}\lambda \tag{A.5}$$

$$Y'''(\omega)=0 \Longleftrightarrow \omega=\omega_0 \vee \omega=\omega_0\pm\lambda \tag{A.6}$$

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

$$Y^{(-1)}(\omega) = A\ \arctan\left(\frac{\omega-\omega_0}{\lambda}\right), \tag{A.7}$$

the area of Y equals $A\pi$ due to

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

Appendix B

Given a discrete spectrum S containing n datapoints as

$$S=\{(\omega_1,S(\omega_1)),\ldots,(\omega_n,S(\omega_n))\},$$

the first discrete derivative S'_{discr} of S is given as

$$\begin{aligned}S'_{discr} &= \left\{\left(\omega'_1,S'(\omega_1)\right),\ldots,\left(\omega'_n,S'(\omega_n)\right)\right\}\\ &= \left\{\left(\frac{\omega_1+\omega_2}{2},S(\omega_2)-S(\omega_1)\right),\ldots,\left(\frac{\omega_{n-1}+\omega_n}{2},S(\omega_n)-S(\omega_{n-1})\right)\right\}\end{aligned}$$

With presuming equal distance between any consecutive pair of frequencies ω_i, ω_{i+1}, $i\in N$ of S, the second discrete derivative S''_{discr} is given as

$$\begin{aligned}&S''_{discr} = \left\{\left(\omega''_1,S''(\omega_1)\right),\ldots,\left(\omega''_n,S''(\omega_n)\right)\right\}\\ &\text{with}\quad \omega''_i=\omega_{i+1}\\ &\text{and}\quad S''(\omega_i)=S(\omega_{i+2})+S(\omega_i)-2S(\omega_{i+1})\end{aligned}$$

Appendix C

Lemma 1. *Given two equally scaled and shaped Lorentzian functions $Y_1(x)$ and $Y_2(x)$ as*

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

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

and given their sum $Y(x) = Y_1(x) + Y_2(x)$, the total number of local optima in Y equals 1 for

$$|\omega_1-\omega_2| \leqslant \frac{2}{\sqrt{3}}\lambda \tag{C.1}$$

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:

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

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

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

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

$$\omega_1 = \omega_2+\frac{2}{\sqrt{3}}\lambda \vee \omega_1 = \omega_2-\frac{2}{\sqrt{3}}\lambda \tag{C.3}$$

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

Lemma 2. *Given two equally scaled and shaped Lorentz-functions Y_1, Y_2, and given their sum Y as in Lemma 1, and let w.l.o.g. be $\omega_1 < \omega_2$, then it holds that*

$$\omega_2-\omega_1 < \frac{2\lambda}{\sqrt{3}} \Rightarrow Y'' \text{ has maximal two roots} \tag{C.4}$$

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. $\omega=\omega_0\pm\frac{1}{\sqrt{3}}\lambda$ (see Eq. (A.5) in Appendix A), and is negative only in the interval $[\omega_0-\frac{1}{\sqrt{3}}\lambda,\omega_0+\frac{1}{\sqrt{3}}\lambda]$, due to

$$\begin{aligned}&Y''_0(\omega)<0\\ &\frac{8A\lambda(\omega-\omega_0)^2}{\left(\lambda^2+(\omega-\omega_0)^2\right)^3}-\frac{2A\lambda}{\left(\lambda^2+(\omega-\omega_0)^2\right)^2}<0\\ &\Longleftrightarrow\ \frac{2A\lambda}{\left(\lambda^2+(\omega-\omega_0)^2\right)^2}\left(\frac{4(\omega-\omega_0)^2}{\lambda^2+(\omega-\omega_0)^2}-1\right)<0\\ &\Longleftrightarrow\ \frac{4(\omega-\omega_0)^2}{\lambda^2+(\omega-\omega_0)^2}<1\\ &\Longleftrightarrow\ 3(\omega-\omega_0)^2<\lambda^2\\ &\Longleftrightarrow\ \omega>\omega_0-\frac{1}{\sqrt{3}}\lambda\wedge\omega<\omega_0+\frac{1}{\sqrt{3}}\lambda\end{aligned}$$

In consequence, Y_0'' is positive for

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

To proof, that the summation of two equally shaped and scaled Lorentz-functions Y_1, Y_2 with $d(Y_1, Y_2) \leqslant \frac{2}{\sqrt{3}}\lambda$ 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

$$y_1 = Y''\left(\omega_0 - \frac{\lambda}{\sqrt{3}} + c\right) = \frac{8A\lambda\left(c - \frac{\lambda}{\sqrt{3}}\right)^2}{\left(\lambda^2 + \left(c - \frac{\lambda}{\sqrt{3}}\right)^2\right)^3} - \frac{2A\lambda}{\left(\lambda^2 + \left(c - \frac{\lambda}{\sqrt{3}}\right)^2\right)^2}$$

$$= \frac{54A\lambda c \overbrace{\left(3c - 2\sqrt{3}\lambda\right)}^{\alpha}}{\left(4\lambda^2 + c\underbrace{\left(3c - 2\sqrt{3}\lambda\right)}_{\alpha}\right)^3}$$

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

$$y_2 = Y''\left(\omega_0 - \frac{\lambda}{\sqrt{3}} + d\right) = \frac{8A\lambda(d - \frac{\lambda}{\sqrt{3}})^2}{\left(\lambda^2 + \left(d - \frac{\lambda}{\sqrt{3}}\right)^2\right)^3} - \frac{2A\lambda}{\left(\lambda^2 + \left(d - \frac{\lambda}{\sqrt{3}}\right)^2\right)^2}$$

$$= \frac{54A\lambda d \overbrace{(3d - 2\sqrt{3}\lambda)}^{\beta}}{\left(4\lambda^2 + d\underbrace{\left(3d - 2\sqrt{3}\lambda\right)}_{\beta}\right)^3}$$

with $d \in \mathbb{R}^{>0}$. 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:

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

for $\lambda, A > 0$ the proof succeeds as

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

□

Appendix D

With observing, that a peak triplet as defined by Eqs. (6) and (7) can have an overlap with neighboring triplets in their boundary positions l_i, r_i only, the worst-case runtime complexity of Algorithm 1 lies in $O(n + |L|)$, since finding the position triplets in line 2 takes time $O(n + |L|)$ at most, since calculating the discrete areas in line 3 needs time $O(n + |L|)$ at most, since calculating the mean and standard deviation of the scores takes time $O(|L|)$, and since the *for*-loop in lines 5–9 takes time $O(|L|)$ as well. Algorithm 1 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 $|L| \leqslant \frac{n}{2}$, Algorithm 1 has a runtime of $O(n + \frac{n}{2}) = O(n)$, and can therefore be considered as *runtime-efficient*.

Appendix E

The equation system

$$\begin{aligned}
y_1 &= A\frac{\lambda}{\lambda^2 + (\omega_1 - \omega)^2} \\
\wedge\ y_2 &= A\frac{\lambda}{\lambda^2 + (\omega_2 - \omega)^2} \\
\wedge\ y_2 &= A\frac{\lambda}{\lambda^2 + (\omega_3 - \omega)^2}
\end{aligned}$$

can be solved for the parameter ω, λ and A as

$$\omega = \frac{\omega_1^2 y_{1,2,3} + \omega_2^2 y_{1,2} y_3 + \omega_2^2 y_2(-y_{1,3})}{2\omega_{1,2} y_1 y_2 - 2(\omega_{1,3} y_1 + (-\omega_{2,3}) y_2) y_3} \tag{E.1}$$

$$\lambda = \frac{1}{\sqrt{y_{2,3}}}\sqrt{\omega_3^2 y_3 + \frac{\alpha}{4(\omega_1 y_1 y_{2,3} + \omega_3 y_{1,2} y_3 + \omega_2 y_2(-y_{1,3}))^2}} \tag{E.2}$$

$$A = \frac{-4\omega_{1,2}\omega_{1,3}\omega_{2,3} y_1 y_2 y_3 (\omega_1 y_1 y_{2,3} + \omega_3 y_{1,2} y_3 + \omega_2 y_2(-y_{1,3}))\lambda}{\left(\omega_{1,2}^4 y_1^2 y_2^2 - 2\omega_{1,2}^2 y_1 y_2\left(\omega_{1,3}^2 y_1 + \omega_{2,3}^2 y_2\right) y_3 + \left(\omega_{1,3}^2 y_1 - \omega_{2,3}^2 y_2\right)^2 y_3^2\right)} \tag{E.3}$$

with

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

Eqs. (E.1), (E.2) and (E.3) are well defined for

$$\omega_1 < \omega_2 < \omega_3 \wedge y_1 < y_2 > y_3 \wedge y_1, y_2, y_3 > 0$$

and

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

$$\wedge y_3 < 2\sqrt{\frac{(\omega_1-\omega_2)^4(\omega_1-\omega_3)^2(\omega_2-\omega_3)^2y_1^3y_2^3}{\left((\omega_1-\omega_3)^2y_1-(\omega_2-\omega_3)^2y_2\right)^4}} - \frac{(\omega_1-\omega_2)^2y_1y_2((\omega_1-\omega_3)^2y_1+(\omega_2-\omega_3)^2y_2)}{\left((\omega_1-\omega_3)^2y_1-(\omega_2-\omega_3)^2y_2\right)^2}$$

Appendix F

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 2–10 of Algorithm 2, the worst-case runtime in terms of counting the number of essential comparisons is given as $O(K \times |J|^2)$, since calculating the sum of all Lorentz-functions in line 5 takes time $O(|J|)$ for each peak triplet on its own, and since the time needed to calculate the new height values and the new parameters in lines 6 and 7 lies in $O(1)$.

References

[1] R.R. Ernst, G. Bodenhausen, A. Wokaun, Principles of Nuclear Magnetic Resonance in One and Two Dimensions, Clarendon Press, Oxford, 1987.
[2] L. Vanhamme, T. Sundin, P. Van Hecke, S. Van Huffel, MR spectroscopic quantitation: a review of time-domain methods, NMR Biomed. (2000) 14–233.
[3] S. Mierisov, M. Ala-Korpela, MR spectroscopy quantitation: a review of frequency domain methods, NMR Biomed. 14 (4) (2001) 247–259.
[4] D. Spielman, P. Webb, A. Macovski, A statistical framework for *in vivo* spectroscopic imaging, J. Magn. Reson. 79 (1988) 66–77.
[5] D.I. Hoult, P.C. Lauterbur, The sensitivity of the Zeugmatographic experiment involving human samples, J. Magn. Reson. 34 (1979) 425.
[6] C. Giancaspro, M.B. Comisarow, Exact interpolation of fourier transform spectra, Appl. Spectrosc. 37 (1983) 153–166.
[7] F.R. Verdun, C. Giancaspro, A.G. Marshall, Effects of noise, time-domain damping, zero-filling and the FFT algorithm on the "exact" interpolation of fast fourier transform spectra, Appl. Spectrosc. 42 (5) (1988) 715–721.
[8] P.F. Bernath, Fourier transform techniques, Encyclopedia of Analytical Science, vol. 3, Elsevier, Oxford, 2005. pp. 498–504.
[9] E. Bartholdi, R.R. Ernst, Fourier spectroscopy and the causality principle, J. Magn. Reson. 11 (1) (1973) 9–19.
[10] A. Ebel, W. Dreher, D. Leibfritz, Effects of zero-filling and apodization on spectral integrals in discrete Fourier-transform spectroscopy of noisy data, J. Magn. Reson. 182 (2006) 330–338.
[11] E. Massaro, V. Viti, L. Guidoni, P. Barone, High-resolution numerical filtering of NMR spectra, Phys. Med. Biol. 34 (7) (1989) 931–938.
[12] A. Asfour, K. Raoof, J.-M. Fournier, Nonlinear identification of NMR spin systems by adaptive filtering, J. Magn. Reson. 145 (2000) 37–51.
[13] J.J. van Vaals, P.H.J. van Gerwen, Novel methods for automatic phase correction of NMR spectra, J. Magn. Reson. 86 (1990) 127–147.
[14] Y. Goto, Highly accurate frequency interpolation of apodized FFT magnitude-mode spectra, Appl. Spectrosc. 52 (1998) 134–138.
[15] G.A. Morris, H. Barjat, T.J. Horne, Reference deconvolution methods, J. Prog. Magn. Res. Spec. 31 (1997) 197–257.
[16] K.R. Metz, M.M. Lam, A.G. Webb, Reference deconvolution: a simple and effective method for resolution enhancement in nuclear magnetic resonance spectroscopy, Concepts Magn. Reson. 12 (1) (2000) 21–42.
[17] Y. Li, M.E. Lacey, J.V. Sweedler, A.G. Webb, Spectral restoration from low signal-to-noise, distorted NMR signals: application to hyphenated capillary electrophoresis-NMR, J. Magn. Reson. 162 (2003) 133–140.
[18] B.C.M. Potts, A.J. Deese, G.J. Steevens, M.D. Reily, D.G. Robertson, J. Theiss, NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse, J. Pharm. Biomed. Anal. 20 (2001) 463–476.
[19] Y. Wang, M.E. Bollard, H. Keun, H. Antti, O. Beckonert, T.M. Ebbels, J.C. Lindon, E. Holmes, H. Tang, J.K. Nicholson, Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning ^{1}H nuclear magnetic resonance spectroscopy of liver tissues, Anal. Biochem. 323 (2003) 26–32.
[20] M.R. Viant, Improved methods for the acquisition and interpretation of NMR metabolomic data, Biochem. Biophys. Res. Commun. 310 (2003) 943–948.
[21] R. Stoyanova, A.W. Nicholls, J.K. Nicholson, J.C. Lindon, T.R. Brown, Automatic alignment of individual peaks in large high-resolution spectral data sets, J. Magn. Reson. 170 (2) (2004) 329–335.
[22] J.J. Jansen, H.C.J. Hoefsloot, J. van der Greef, M.E. Timmerman, A.K. Smilde, Multilevel component analysis of time-resolved metabolic fingerprinting data, Anal. Chim. Acta 530 (2005) 173–183.
[23] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics, Anal. Chem. 78 (2006) 4281–4290.
[24] D. Chang, A. Weljie, J. Newton, Leveraging latent information in NMR spectra for robust predictive models, Pac. Symp. Biocomput. 12 (2007) 115–126.
[25] M. Dijkstra, H. Roelofsen, R.J. Vonk, R.C. Jansen, Peak quantification in surface-enhanced laser desorption/ionization by using mixture models, Proteomics 6 (2006) 5106–5116.
[26] A.M. Weljie, J. Newton, P. Mercier, E. Carlson, C.M. Slupsky, Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data, Anal. Chem. 78 (13) (2006) 4430–4442.
[27] J.J. Neil, G.L. Bretthorst, On the use of bayesian probability theory for analysis of exponential decay date: an example taken from intravoxel incoherent motion experiments, Magn. Reson. Med. 29 (5) (1993) 642–647.
[28] M.I. Miller, A.S. Greene, Maximum-likelihood estimation for nuclear magnetic resonance spectroscopy, J. Magn. Reson. 83 (3) (1989) 525–548.
[29] L. Vanhamme, T. Sunidn, P.V. Hecke, S.V. Huffel, MR spectroscopy quantitation: a review of time-domain methods, NMR Biomed. 14 (2001) 233–246.
[30] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, SIAM J. Appl. Math. 11 (2) (1963) 431–441.
[31] J. Järvi, S. Nyman, M. Komu, J.J. Forsström, A PC programm for automatic analysis of NMR spectrum series, Comput. Methods Prog. Biomed. 3 (52) (1997) 213–222.
[32] J.-L. Pons, T.E. Malliavin, M.A. Delsuc, Gifa V.4: a complete package for NMR data set processing, J. Biomol. NMR 8 (1996) 445–452.
[33] G.J. Metzger, M. Patel, X. Hu, Application of genetic algorithms to spectral quantification, J. Magn. Reson. B 110 (1996) 316320.
[34] M. Karakaplan, Fitting Lorentzian peaks with evolutionary genetic algorithm based on stochastic search procedure, Anal. Chim. Acta 587 (2) (2007) 235–239.
[35] G.L. Bretthorst, W.C. Hutton, J.R. Garbow, J.J.H. Ackerman, Exponential model selection (in NMR) using bayesian probability theory, Concepts Magn. Reson. 27A (2) (2005) 64–72.
[36] R. Koradi, M. Billeter, M. Engeli, P. Güntert, K. Wüthrich, Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY, J. Magn. Reson. 135 (1998) 288–297.
[37] H.N.B. Moseley, N. Riaz, J.M. Aramini, T. Szyperski, G.T. Montelione, A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra, J. Magn. Reson. 170 (2004) 263–277.
[38] H.W. Koh, S. Maddula, J. Lambert, R. Hergenröder, L. Hildebrand, Feature selection by Lorentzian peak reconstruction for ^{1}NMR post-processing, in: 21st IEEE Symposium on Computer-Based Medical Systems, 2008, pp. 608–613.
[39] H. Fribolin, Basic One and Two Dimensional NMR Spectroscopy, third ed., Wiley-VCH, 1999, ISBN 3-527-29514-3.
[40] Eric W. Weisstein, Levenberg–Marquardt Method, From MathWorld – A Wolfram Web Resource. Available from: <http://mathworld.wolfram.com/Levenberg-MarquardtMethod.html>.