Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain

Stanislav Sokolenko, Ghina Hajjar, Jonathan Farjon, Serge Akoka, Patrick Giraudeau, Tangi Jezequel

Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain

Stanislav Sokolenko[a,*], Tangi Jézéquel[b], Ghina Hajjar[b,c], Jonathan Farjon[b], Serge Akoka[b], Patrick Giraudeau[b,d]

[a] *Department of Process Engineering and Applied Science, Dalhousie University, 1360 Barrington St., PO Box 15000, Halifax NS B3H 4R2, Canada*
[b] *CEISAM, UMR CNRS 6230 Bât. 22 Faculté des Sciences et Techniques 2 rue de la Houssinière 44322 Nantes Cedex 03, France*
[c] *Laboratory of Metrology and Isotopic Fractionation, Research Unit: Technologies et Valorisation Agroalimentaire (TVA), Faculty of Science, Saint-Joseph University of Beirut, PO Box 17-5208 Mar Mikhael, Beirut 1104 2020, Lebanon*
[d] *Institut Universitaire de France, 1 rue Descartes, 75005 Paris Cedex 05, France*

Abstract

Quantitative NMR is intrinsically dependent on precise, accurate, and robust peak area calculation. In this work, we demonstrate how the use of complex-valued peak descriptions can improve peak fitting in the frequency domain — incorporating phase and baseline correction as well as apodization while working with commonly used Fourier-transformed data. The method has been implemented in an open source R package called rnmrfit that is available for download on GitHub (`https://github.com/ssokolen/rnmrfit`). Application to real data suggests that this approach can also result in dramatically higher precision than can be achieved with existing software. Simulation data indicates that coefficients of variation below 0.1% can be readily achieved at signal to noise (SNR) ratios of approximately 100. The use of complex-valued data in the frequency domain is demonstrated as a relatively simple and effective means of improving peak fitting for quantitative NMR analysis.

Keywords: NMR, lineshape, fitting, quantitative, frequency, imaginary

*Corresponding author
Email address: `Stanislav.Sokolenko@dal.ca` (Stanislav Sokolenko)

1. Introduction

The calculation of NMR resonance peak areas is a key aspect of NMR analysis and forms the basis of so-called "quantitative" NMR (typically referred to as qNMR). This quantification relies on the basic principle that the area of a resonance is proportional to the number of nuclei that corresponds to that resonance [1], allowing the calculation of relative compound concentrations from ratios of peak areas (and the calculation of absolute concentrations with the addition of an internal standard). qNMR has been used across a wide array of applications that include metabolomics [2], natural product analysis [3], drug discovery and development [4], as well as isotopic analysis (the most demanding application as it requires coefficients of variation less than 0.1%) [5]. And whereas there are considerable differences across these different fields, one thing they all have in common is the need for precise and accurate peak area calculation.

The basic method of area quantification is the integration of the NMR lineshape between two chemical shifts. Malz and Jancke [6] have shown that the simple integration of isolated peaks can achieve 95% confidence intervals as low as 1.5% of the compound concentrations. However, particularly in the analysis of natural products and biofluids typical of metabolomics, isolated peaks may be few and far between. Furthermore, the same researchers [6] have also commented that the integration of even well isolated peaks can result in relative uncertainties of 11% with even "slightly wrong" phase and baseline corrections [6]. The general alternative to direct integration of the spectrum is some form of lineshape fitting with the overall goal of separating the contribution of different chemical species from each other and confounding components such as baseline and phase error. Roughly speaking, this process can be divided into two steps — the identification of the chemical species (or residues) that make up the observed resonance and their subsequent quantification. In many cases, the compounds or residues to be quantified are known in advance — making the final step of quantification the more general of the two. And in many applications of qNMR, the basic challenge is how to achieve the most precise quantification

of resonance peaks [5, 7].

From a mathematical perspective, one of the basic divisions of existing lineshape fitting approaches is whether the NMR lineshape is modelled in the time or frequency domain. In the time domain, an NMR resonance is typically modelled as a decaying sinusoidal function (more specifically, a complex-valued exponential) with constraints added to account for multiplet relations. The use of Bayesian inference to estimate the number of resonances as well as resonance frequency, decay rate, and intensity has been described as early as 1990 [8, 9, 10]. Some of the more recent advances include the improvement of the Markov Chain Monte Carlo technique used to draw samples [11, 12], the implementation of downsampling for isolating specific regions of interest [13], and the introduction of more generalized lineshapes to account for distortions [14]. In the frequency domain (following the Fourier transform of the FID), lineshapes are typically fit as Lorentz peaks using real-valued data following phase correction. Most approaches rely on a database of compounds, whether generated in-house or available online [15, 16, 17, 18], and many of these are also based on Bayesian inference [16, 17, 18]. The use of libraries can be very effective for standardized samples and pulse sequences (such as those used in metabolomics), but their application is limited to specific combinations of compounds, biological matrices, and pulse sequences on which the database is built. However, individual peaks (with constraints to define multiplet relations) have also been fit using classical non-linear optimization techniques based on the least-squares fit [19, 20].

Some general trade-offs can be identified when considering the various approaches. Lineshape fitting in the time domain has the advantage of being more general, with the use of complex-valued data allowing for built-in phase correction. However, even a relatively crowded NMR spectrum contains large areas of relatively signal-free regions in the frequency domain that only contribute noise to an overall fit in the time domain. Although it is possible to downsample [13], this adds extra processing steps and potential for artefacts, whereas the Fourier transform is already a well accepted means of isolating frequencies of interest

and disregarding more problematic ones. Fitting in the frequency domain can be used to target specific peaks of interest while greatly speeding up the fit process (by simply considering fewer points). There is a similar trade-off between Bayesian and classical non-linear optimization techniques. Bayesian approaches are more general as they seek out a global solution whereas optimization based on gradient descent or related techniques require good initial guesses to avoid getting trapped in local solutions that are not globally optimal. On the other hand, Bayesian fitting methods are also considerably slower for high precision calculations.

The method proposed in this work bridges some of these trade-offs by considering a more generalized form of fitting in the frequency domain. One way that is accomplished is by describing NMR peaks as complex-valued functions. There are a number of practical and theoretical advantages to this approach. First, phase correction can be implemented as part of the fitting process in the frequency domain, thus avoiding user-dependent phasing that affects the accuracy and precision of peak area determination. Second, baseline distortions can be better separated from the peak shapes as one set of peak parameters is used to describe two different shapes (as they appear in the real and imaginary components). Third, the broader shape of typical Lorentz/Voigt peaks in the imaginary domain allows for more robust gradient-descent style optimization — initial guesses at peak position can be further away from their real positions without resulting in a spurious fit. It should be noted that the use of complex-valued data in the frequency domain has been suggested in the past (e.g. [21]), but to the best of our knowledge, the approach has not been described or explored in-depth. On top of using both real and imaginary data, the proposed method also adds a convolution term that can be used to incorporate changes in peaks shape that result from truncation or different forms of apodization. The proposed approach is implemented in an open source R package called rnmrfit that is available for download on GitHub (`https://github.com/ssokolen/rnmrfit`). Here, we present the theory underlying this approach, as well as an evaluation of its performance on both simulated and experimental data.

2. Theory

2.1. Complex-valued Lorentz function

Using the general notation of Keeler [22], a simple description of a magnetic resonance singlet in the time domain is that of a decaying complex exponential expressed as:

$$S(t) = S_0 \exp(i\Omega t) \exp(-Rt) \tag{1}$$

where the signal (S) is a function of time (t), relative intensity (S_o), resonance frequency (Ω), and relaxation rate constant (R, which can also be expressed as the reciprocal of the relaxation time $1/T_2$). The Fourier transform of Equation 1 is the complex-valued Lorentz function:

$$FT[S(t)] = S_0(A(\omega) + iD(\omega)) \tag{2}$$

$$= S_0 \left(\frac{R^2 + iR(\omega - \Omega)}{R^2 + (\omega - \Omega)^2} \right) \tag{3}$$

where ω is frequency and A and D are the absorption and dispersion lineshapes. Written explicitly:

$$A(\omega) = \frac{R^2}{R^2 + (\omega - \Omega)^2} \tag{4}$$

$$D(\omega) = \frac{R(\omega - \Omega)}{R^2 + (\omega - \Omega)^2} \tag{5}$$

S_0 is generally taken to have arbitrary units while ω, Ω, and R are all in the same units of frequency, so a variable transformation can be used to make the peak function dimensionless:

$$f(z) = S_0 \left(\frac{1 + iz}{1 + z^2} \right) \tag{6}$$

$$z = \frac{\omega - \Omega}{R} \tag{7}$$

As z is dimensionless, the units of Ω and R will be ignored for the rest of this section (with Hz and ppm used interchangeably).

2.2. Parameter estimation

Observed Fourier transformed data can be considered as a collection of (x, y) pairs, where y has both a real $\Re(y)$ and imaginary $\Im(y)$ component. Minimizing the sum of squared deviation of both real and imaginary data offers a convenient optimization condition. Considering the fit of a single Lorentz peak to j points, the deviation of the fit from the observed data is expressed as:

$$z = \frac{x - \Omega}{R} \tag{8}$$

$$f(z) = S_0 \left(\frac{1 + iz}{1 + z^2} \right) \tag{9}$$

$$\varepsilon = y - f(z) \tag{10}$$

$$SS = \sum \Re(\varepsilon)^2 + \Im(\varepsilon)^2 \tag{11}$$

where x and y are vectors of data points and SS is the total sum of squares across all points. The calculation of analytical derivatives allows for the convenient use of gradient-descent optimization algorithms:

$$\frac{\partial \varepsilon}{\partial S_0} = - \left(\frac{1 + iz}{1 + z^2} \right) \tag{12}$$

$$\frac{\partial \varepsilon}{\partial z} = -S_0 \left(\frac{-2z + i(1 - z^2)}{(1 + z^2)^2} \right) \tag{13}$$

$$\frac{\partial \varepsilon}{\partial \Omega} = \frac{\partial \varepsilon}{\partial z} \frac{\partial z}{\partial \Omega} \tag{14}$$

$$\frac{\partial \varepsilon}{\partial R} = \frac{\partial \varepsilon}{\partial z} \frac{\partial z}{\partial R} \tag{15}$$

Thus, the derivative of SS with respect to any parameter θ can be calculated as:

$$\frac{\partial SS}{\partial \theta} = \sum 2\Re(\varepsilon)\Re \left(\frac{\partial \varepsilon}{\partial \theta} \right) + 2\Im(\varepsilon)\Im \left(\frac{\partial \varepsilon}{\partial \theta} \right) \tag{16}$$

2.3. Gauss and Voigt functions

The same principles can be expanded to consider both Gauss and Voigt functions. The Voigt function can be seen as an extension of the Lorentz, where the signal decay rate is not constant, but takes on normally distributed values — essentially leading to convolution of the Lorentz and Gauss peaks. Although

it is possible to approximate the Voigt function using a sum of Lorentz and Gaussian terms, performing the optimization with complex values allows for a more concise representation. The optimization relies on the same general equations as above (10, 11, 16), with the specific parameters outlined in Table 1. It should be noted that the direct evaluation of $\exp(-z^2)\mathrm{erfc}(-iz)$ is numerically unstable, so it is necessary to use a numerically optimized implementation of the Voigt function for the calculation (often referred to as the Faddeeva function) [23].

Table 1: Summary of Lorentz, Gauss, and Voigt functions, where S_0 is relative intensity of the peak (its relative height), Ω is the resonance frequency (relative peak position), and R is the relaxation rate constant (relative peak width). The R_G term in the Voigt function relates to the distribution of possible R values (it can be seen as the Gaussian component of the overall peak).

	Lorentz	Gauss	Voigt
$f(z)$	$S_0\left(\frac{1+iz}{1+z^2}\right)$	$S_0\exp(-z^2)\mathrm{erfc}(-iz)$	$S_0\exp(-z^2)\mathrm{erfc}(-iz)$
z	$\frac{x-\Omega}{R}$	$\frac{x-\Omega}{\sqrt{2}R}$	$\frac{x-\Omega+iR}{i\sqrt{2}R_G}$
$\frac{\partial f(z)}{\partial z}$	$\frac{-2z+i(1-z^2)}{(1+z^2)^2}$	$-2zf(z)+\frac{i2}{\sqrt{\pi}}$	$-2zf(z)+\frac{i2}{\sqrt{\pi}}$

2.4. Baseline and phase correction

Apart from considering Lorentz, Gauss, and Voigt peak functions, $f(z)$, it is also possible to add baseline terms $f_b(z)$. In rnmrfit, $f_b(z)$ is modelled as a complex-valued spline polynomial with a variable order and a variable number of equidistant knots. The result is that the baseline is expressed as a piece-wise polynomial that is capable of modelling most observed baseline distortions. Baseline smoothness can be indirectly controlled by both polynomial order and knot number – with 2nd or 3rd order polynomials and 1 or 2 interior knots sufficient for correcting minor distortions. In principle, the baseline for the real and imaginary domains should be the same. In practice, however, differences can arise from a number of possible sources. For example, baseline distortions caused by macromolecules will naturally have different real and imaginary components.

Furthermore, peaks are considerably wider in the imaginary domain than in the real and therefore feature a greater degree of overlap, so fitting a subsection of the spectrum may feature contributions from outside of the selected region that only appear in the imaginary data. To deal with these discrepancies, the default option in rnmrfit is to fit separate real and imaginary baselines, but extra precision can be gained by forcing a single common baseline where appropriate. Despite the use of two different baselines, peaks are estimated using a single set of parameters, so more information is nonetheless captured than if the imaginary domain is entirely ignored.

The use of complex-valued data also allows the inclusion of phase correction directly in the fit. A phase transform in the frequency dimension can be expressed as y', with the real and imaginary components calculated as:

$$\Re(y') = \Re(y)\cos(\phi) + \Im(y)\sin(\phi) \tag{17}$$

$$\Im(y') = \Im(y)\cos(\phi) - \Re(y)\sin(\phi) \tag{18}$$

with the corresponding change in ε to ε':

$$\Re(\varepsilon') = \Re(y') - \Re(f(z)) \tag{19}$$

$$\Im(\varepsilon') = \Im(y') - \Re(f(z)) \tag{20}$$

Apart from the difference in nomenclature, this change does not impact any of the previously developed equations. But it does add another set of derivatives to fit the local phase angle ϕ:

$$\frac{\partial\Re(\varepsilon')}{\partial\phi} = -\Re(y)\sin(\phi) + \Im(y)\cos(\phi) \tag{21}$$

$$\frac{\partial\Im(\varepsilon')}{\partial\phi} = -\Im(y)\sin(\phi) - \Re(y)\cos(\phi) \tag{22}$$

2.5. Apodization

Apodization and truncation can be expressed as a simple product in the time domain or a convolution in the frequency domain. Although the time domain calculation is naturally more efficient, the difference in computation time is less pronounced when considering that the frequency domain fit typically considers

only a fraction of the data points. The addition of a convolution term can be directly included in the model fit, transforming Equations 10 and 13 into:

$$\varepsilon'' = y - f(z) * g \tag{23}$$

$$\frac{\partial \varepsilon''}{\partial z} = -\frac{\partial f(z)}{\partial z} * g \tag{24}$$

where g is a convolution vector with a Fourier transform. Examples of g include the Fourier transforms of a step function to represent truncation or common apodization functions such as exponentials and sinusoidals. The result is that the change in shape of the peak resulting from truncation (the formation of "sinc wiggles") or apodization (line sharpening/broadening) can be incorporated directly in the fit. The apodization can be combined with any of the peak functions as well as baseline and phase correction. Furthermore, g can take on any function that has a Fourier transform — allowing the application of reference deconvolution or other forms of signal correction in addition to apodization.

2.6. Multiple peaks

Although the above derivation considers the fit of only one peak at a time, multiple peaks can be fit by including a sum of all peak contributions, with minimal changes to the derived equations. The definition of multiplets requires the use of constrained optimization algorithms, where the distance between multiple peaks and intensity ratios can be specified. The rnmrfit package allows for both hard (linear) and soft (inequality) constraints to allow for approximate multiplet definitions when the coupling constant may not be known exactly.

3. Materials and methods

3.1. Software implementation

The proposed algorithm has been implemented in pure R in the rnmrfit package by leveraging the nloptr [24] and RcppFaddeeva [23] packages (which provide R wrappers around the C/C++ NLopt and Faddeeva code developed by the Ab Initio group). Of the various optimization algorithms available through

NLopt, the sequential quadratic programming (SQP) algorithm for nonlinearly constrained gradient-based optimization (SLSQP) was chosen for its ability to support both inequality and equality constraints. The rnmrfit package provides the user with a relatively simple API for defining, fitting, and plotting NMR multiplets within 3-5 lines of code. An in-depth tutorial for the package is available on GitHub (`https://github.com/ssokolen/rnmrfit`).

3.2. Simulated NMR data

Sample NMR data was generated in the frequency domain using the Lorentz and Voigt lineshapes (with the Gauss component of the Voigt lineshape set to 50% of the Lorentz). Three peak combinations were chosen to compare algorithm performance with various levels of overlap: a singlet, a singlet and a doublet, and two doublets. Each dataset was generated with 256 real and 256 imaginary data points, with the width of each peak at half height corresponding to approximately 10-20 points. The ideal peak shapes were distorted using different levels of noise as well as phase and baseline errors. Gaussian noise was applied to generate signal to noise (SNR) ratios of 5, 10, 50, and 100, where SNR is calculated as the variance of the pure peak data divided by the variance of the noise. Baseline errors were generated as quadratic polynomials with the position of the maximum/minimum randomly selected to fall within the domain of the data and the magnitude set to 20% of the maximum peak intensity. Phase errors were added at three different levels: 0°, 15°, and 30°, with the sign of the phase error chosen at random.

3.3. Cholesterol NMR data

For the basic illustration of the fitting approach, a ^{13}C-NMR spectrum of cholesterol was recorded using the adiabatic INEPT pulse sequence reported in [25]. 90 mg of cholesterol was dissolved in 600 μL of $CDCl_3$. The spectrum was recorded on a 500 MHz Bruker Avance-III spectrometer equipped with a 5 mm dual ^{13}C/^{1}H cryoprobe (tuned to the ^{13}C recording frequency of 125.76 MHz). The temperature of the probe was set to 288 K. The 90° ^{1}H and ^{13}C pulse

widths were calibrated to 10 μs and 11 μs, respectively, with an acquisition time of 0.8 s, recovery time of 10.8 s, 4 dummy scans, and 16 scans. Both inversion and ^{1}H decoupling were performed with an adiabatic full-passage pulse [26]. Refocusing was achieved with a composite adiabatic pulse.

3.4. Vanillin NMR data

^{1}H-NMR spectra of vanillin were used to evaluate the accuracy and precision of the algorithm on real data (previously reported in [27]). Briefly, 77 mg of $DMSO_2$ and 250 mg of vanillin were dissolved in 510 μL of acetone-d_6 (with 11.76 mM $Cr(acac)_3$ relaxing agent). ^{1}H-NMR spectra were recorded using the recently published pulse sequences for high-precision ^{1}H quantitative NMR (DWET, MWET and PWET) [27] on a 400 MHz Bruker Avance-I spectrometer equipped with a 5 mm dual ^{13}C/^{1}H probe. The temperature of the probe was set to 303 K. The 90° ^{1}H pulse width was calibrated to 10 μs, with an acquisition time of 4s, delay time of 5s, 2 dummy scans, and 4 scans.

Seven spectra were collected for each pulse sequence using a single sample to estimate precision and accuracy. Peak areas were calculated using PERCH Software (Perch solutions Ltd, Kuopio, Finland), global spectral deconvolution (GSD) implemented in Mnova 12.0 (Mestrelab Research, S.L., Santiago de Compostela, Spain), as well as the proposed algorithm. The overall spectra were divided into 5 regions, corresponding to the methoxy, aromatic, hydroxy, and aldehyde groups of vanillin along with the $DMSO_2$ peak. All but the aromatic regions were fit with one singlet each, while the aromatic region required four singlets.

3.5. Synthetic mixture and plasma NMR data

^{1}H-NMR spectra of a synthetic mixture and a human plasma sample were used to provide examples of algorithm performance on more complex data. The synthetic metabolite mixture consisted of 2.27 mM valine, 2.12 mM lactate, 2.25 mM n-acetylaspartate, 2.31 mM methionine, 2.28 mM glutamate, 2.06 mM creatine, 2.31 mM phenylalanine, 2.35 mM taurine, 2.22 mM histidine, 2.33 mM

glycine, and 0.167 mM TMSP as reference. All compounds were purchased from Sigma-Aldrich and dissolved in a 50 mM phosphate buffer. The pH of the solution was adjusted to 7.4 using a 1 M hydrochloride solution. The spectrum was recorded on a 500 MHz Bruker Avance-III spectrometer equipped with a 5 mm dual $^{13}C/^{1}H$ cryoprobe (tuned to the ^{1}H recording frequency of 500.13 MHz). The temperature of the probe was set to 288 K. Five consecutive spectra were acquired with the noesy-1d sequence, 64 scans and 16 dummy scans, a sampling period of 1.5 s, a recovery delay of 4.5 s, and a spectral width of 16 ppm.

4. Results and discussion

4.1. Basic fit

A basic demonstration of the fit is shown in Figure 1 using ^{13}C-NMR analysis of a cholesterol sample. Although a similar high quality fit of the basic lineshape can be expected using existing approaches, Figure 1 demonstrates the versatility of the proposed algorithm in correcting phase and baseline errors as well as the incorporation of apodization directly in the fit. The incorporation of resolution-enhancing apodization is likely to be useful for fitting highly overlapping spectra or dealing with highly truncated spectra without the need for significant line broadening. The overall fit including phase and baseline correction is achieved in a fraction of a second on a mid-range laptop with an i7-2640M processor (with the apodization adding a couple of extra seconds). Although the overall calculation time scales with increasing numbers of peaks, there is potential for speed improvement by translating the main optimization code from R into C++ using the Rcpp package [28] — ensuring rapid calculation even at larger scales.

4.2. Varying noise and error

The impact of signal to noise ratio (SNR) as well as baseline and phase errors was explored using simulated data. Three different resonance combinations were considered as simple test cases — a singlet, a singlet and doublet, as well as

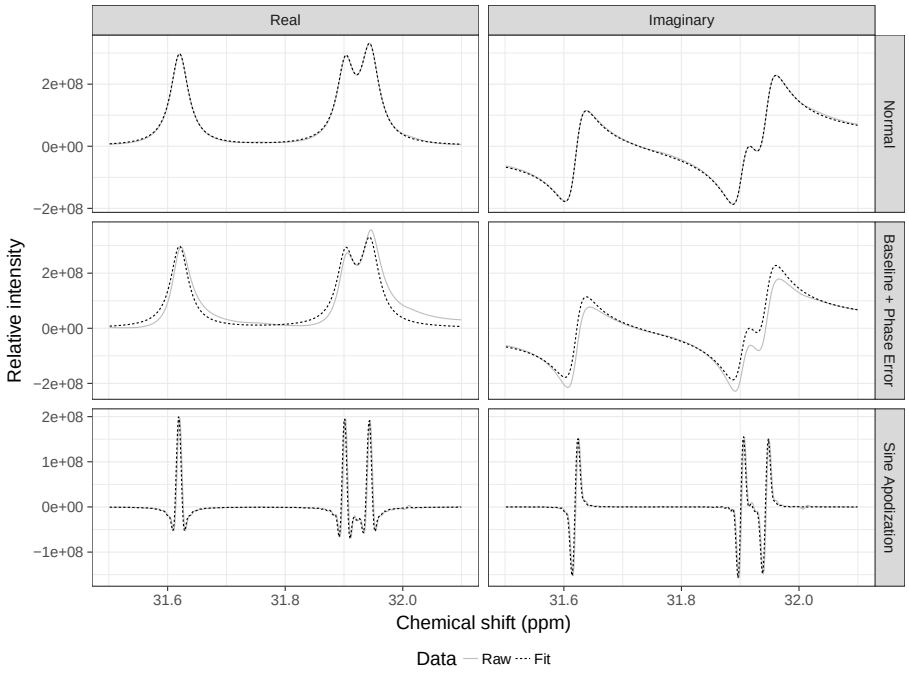

Figure 1: Demonstration of proposed algorithm using the C2, C7, and C8 resonances of cholesterol (see Figure S2 in the Supplementary Information for reference). The data fit in the first row is the original data with 2 Hz line broadening, the data fit in the second row had simulated baseline and phase errors added to it, and the data fit in the third row has had sine bell apodization applied.

two doublets. A summary of the results is presented in Figure 2. Overall neither phase errors of up to 30° nor quadratic polynomial baseline errors with a magnitude of up to 20% of the signal contributed to the overall accuracy and precision of the fit. With an SNR of 50, the median error of the calculated peak area was approximately 0.25% and the coefficient of variation (CV) was approximately 0.35%. If an SNR of 100 can be achieved, then both values fall below 0.1%. Reducing the SNR down to 5 modestly increases both accuracy and precision to approximately 2-3%.

Although Figure 2 presents the data for only the Lorentz peak shape, the results were practically equivalent for the more general Voigt shape (see Figure S1

in the Supplementary Information). It was found that the extra flexibility offered by the Voigt shape (in modulating the relative width of a peak at different peak heights) led to an increased possibility of optimization divergence. Effectively, a dramatic change in the baseline and phase terms led to a sub-optimal (global) convergence before the peak width could be correctly fit. However, this effect was eliminated by fitting Lorentz peaks first and using the results as an initial guess to fit the Voigt peaks. Overall, the best results were obtained by fitting Lorentz peaks without phase or baseline correction, refitting with the addition of a phase term, and then changing the Lorentz peak type to Voigt for a final fit. It should be noted that the rates of divergence were rarely higher than 1%, meaning that a direct Voigt fit should be sufficient for most general cases.

4.3. Varying initial parameter values

Despite the potential for a high degree of precision and accuracy, iterative optimization algorithms can be very sensitive to the initial parameter values used to begin the optimization. However, the use of both real and imaginary data has made the proposed algorithm more robust to poor initial values. As it is currently implemented, the algorithm requires the user to supply an estimate of only the multiplet chemical shift (as well as the coupling pattern) — the peak heights are estimated as the intensity value corresponding to the chemical shift of the singlets that make up the multiplet and the initial peak width is taken as 1 Hz (which was found to be sufficient for all real data tested, although there is an option to override the default value). The ability of the algorithm to converge onto a global optimum was tested on a simulated singlet with an SNR of 5, with and without phase and baseline error correction. The results are summarized in Figure 3. Despite the significant noise and phase or baseline errors, 100% convergence was achieved for any initial chemical shift that fell within approximately 25% of the maximum peak height. With no baseline or phase errors, the initial chemical shift could be as far away as 5% or 10% of maximum peak height while ensuring convergence. Such estimates could be

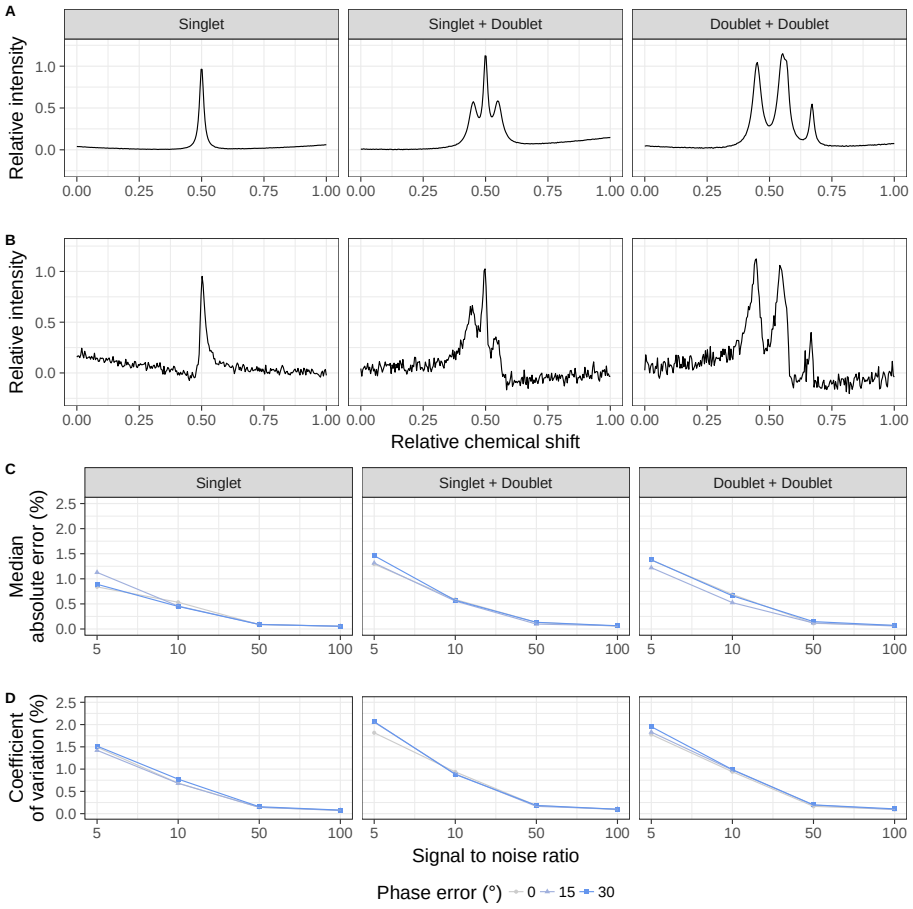

Figure 2: Summary of Lorentz peak fit accuracy and precision as a function of SNR as well as phase and baseline error. 100 peaks were randomly generated for each condition — combining an SNR of 5, 10, 50, or 100 with a phase error of 0°, 15°, or 30°. A quadratic baseline with a magnitude of up to 20% of the maximum peak height was added at each condition. **A** Example spectrum at an SNR of 100 with 0° phase error. **B** Example spectrum at an SNR of 5 with 30° phase error. **C** Median error of calculated area as a percent of the true area. **D** Coefficient of variation of the calculated area (standard deviation divided by the average area).

made by eye or with the aid of separate peak-picking software. The initial peak width was found to have minimal impact (data not shown) and was set at 3× narrower than the true width for the simulation. Some complex combinations

of multiplets may suffer from worse performance, but it is unlikely that these can be identified *a priori*. Performing an initial fit without phase or baseline correction to improve on the initial chemical shift values was sufficient to ensure convergence in all practical cases tested (data not shown).

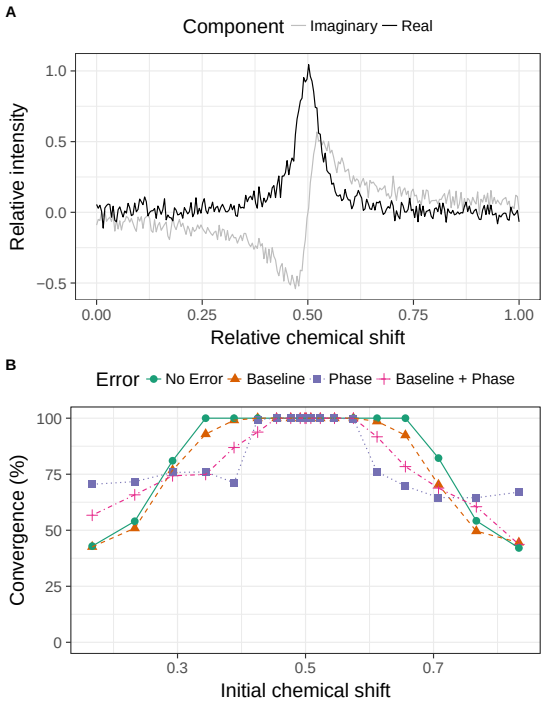

Figure 3: Impact of initial chemical shift on global convergence with and without baseline and phase error correction. **A** Example singlet with an SNR of 5 and no phase or baseline error. **B** Percent convergence based on 1000 iterations as a function of initial chemical shift. When present, the phase error was set to $\pm 30°$ and the baseline error was set to a quadratic function with a magnitude of up to 20% of the maximum peak height.

4.4. Application to high precision data

The performance of the proposed approach was assessed in the context of high precision quantitative ^{1}H-NMR using recently published data [27]. ^{1}H-NMR spectra were obtained with the DWET, MWET and PWET pulse sequences capable of yielding a high degree of precision on peak areas from successively recorded spectra, as required in the context of isotopic analysis for

which they have been developed. Briefly, these pulse sequences combine several spatial encoding elements to remove the effect of radiation damping in concentrated samples with the goal of determining the relative amount of internal standard at a higher level of accuracy that would be possible with gravimetric analysis. Seven ^{1}H spectra of vanillin for each pulse sequence were fit using the proposed algorithm and the resulting precision was compared to that achieved by PERCH [29] and Mnova software. PERCH is commonly used for NMR peak fitting and is recognized by the qNMR community for its good performance — particularly in the field of isotopic NMR where it acts as a reference, whereas Mnova offers the added benefit of automated peak-picking on top of spectral deconvolution. The results of the comparison are summarized in Figure 4 with the complete area quantification data available as a Microsoft Excel file in the Supplementary Information (the Mnova CV values are excluded from the figure as they were considerably higher than both of the other methods). Overall, the proposed approach was found to be more precise across practically all spectral regions and all three of the pulse sequences tested, in both absolute terms (Figure 4C) and when normalized by the area of the $DMSO_2$ peak (Figure 4D). The magnitude of the difference, however, varied considerably across the different regions. The absolute quantification of $DMSO_2$ and methoxy peak areas using the proposed approach was observed to have a coefficient of variation in the range of 0.02%-0.05%, approximately five times more precise than with PERCH. As shown in Figure 4B, the signal to noise ratio for these two peaks was visibly larger than for the other spectral regions and suggests that the proposed approach can take advantage of strong signal data — at smaller signal to noise ratios, the fitting algorithm is unlikely to have as much impact. However, it should be noted that the primary strength of the proposed algorithm is its robustness and generality, with the increase in precision serving as a useful side benefit.

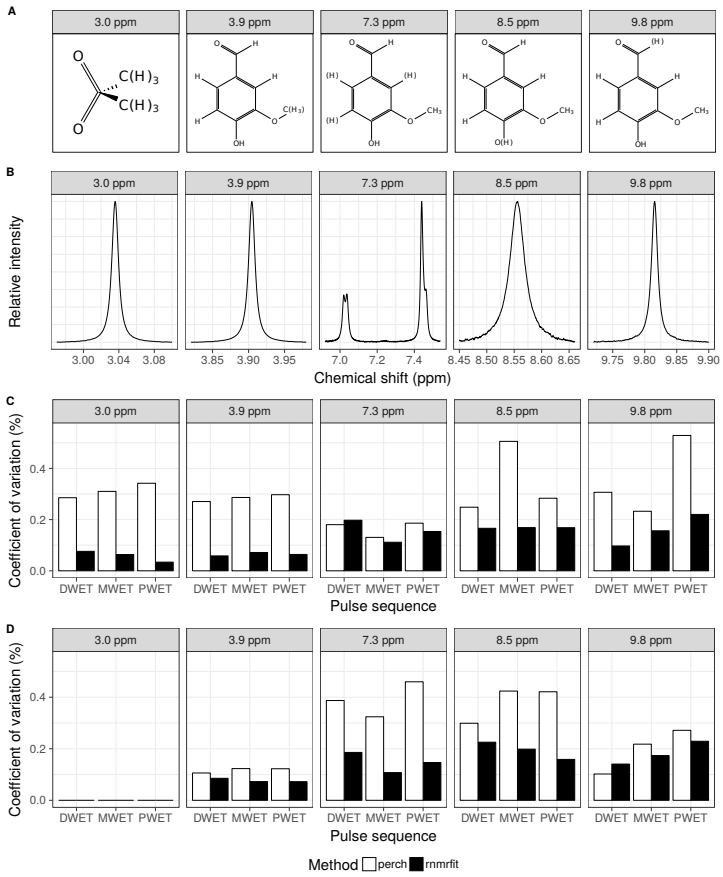

Figure 4: Comparison of peak area quantification precision between rnmrfit and PERCH Software based on the analysis of 7 ^{1}H-NMR spectra of vanillin (see Materials and methods for more detailed sample description). **A** Hydrogens corresponding to each spectral region highlighted with parentheses. **B** Example spectra corresponding to each spectral region fit with the software. **C** The coefficient of variation (in percent of mean value) from the quantification of area for each spectral region. **D** The coefficient of variation (in percent of mean value) from the quantification of area for each spectral region, normalized by the area of the $DMSO_2$ peak.

4.5. Application to more complex data

Whereas the analysis above has focused on relatively simple spectra, the proposed approach can also be used for more complex data, featuring significant spectral crowding or baseline distortions. The following examples are not in-

tended to be comprehensive but offer a glimpse of what the approach is capable of. Moving beyond basic singlets, a mixture of 17 peaks with some light overlap (stemming from glutamate, methionine, creatine, histidine, and phenylalanine in the 3.70-4.05 ppm range) could be fit within a couple of seconds and achieve a coefficient of variation of approximately 0.5% for the peak area corresponding to each compound (see Figure 5 for a visualization of the fit). The peaks were fit as four double doublets and one singlet, with the J-coupling frequency obtained from the Human Metabolome Database (HMDB, [30]) — rnmrfit can also make use of inequality constraints to allow for cases where the J-coupling frequency estimates may not be exact. An example from a human plasma sample featuring a complex baseline is presented in Figure 6. Modelling the complex baseline required the number of interior baseline knots to be increased from 2 or 3 (used in the previous examples) up to 25, but the resulting fit appears to be quite good — the baseline estimate follows the general contour of the data while leaving a sufficient gap at the base of the peaks. Despite the increased sample complexity, the median coefficient of variation for the peak area was found to be approximately 2%, with the fit taking several seconds per spectrum.

Although rnmrfit is primarily intended to be used for fitting spectra where the identity of many compounds are known ahead of time (and their corresponding NMR spectra can be defined as a series of multiplets), the fit algorithm itself can also be used in the context of global spectrum decomposition, where a complex spectrum is broken down into a series of singlets (or other simple spectral features). An example from a human plasma sample featuring complex spectral crowding is presented in Figure 7, where the proposed approach demonstrates a very good fit of the data using 86 singlets and a baseline with 20 interior knots. Despite the good fit, it is also important to note a number of key limitations. First, the initial position of the singlets was identified by eye. Automated peak selection is a challenging task that was deemed to be outside the scope of the current implementation, which focuses squarely on the fit process itself[1].

[1] Future versions of the software may include an implementation of an existing peak-picking

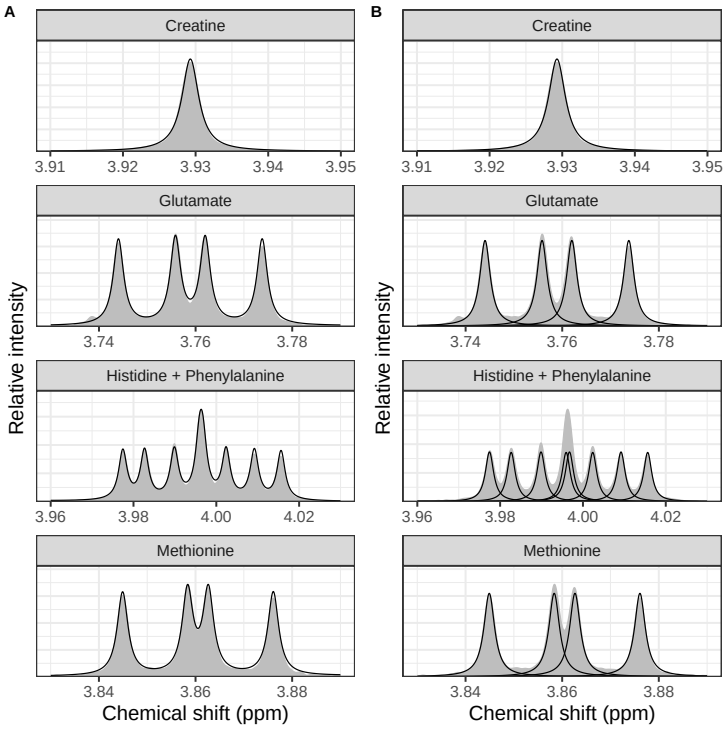

Figure 5: Demonstration of proposed algorithm on a subspectrum of a simple mixture (see Materials and methods for more details). **A** Overlap of the overall fit (black lines) over the observed data (gray fill). **B** Breakdown of the overall fit into constituent peaks.

Second, achieving a reasonable fit of complex data required liberal use of optimization constraints. Once picked, peak positions were not allowed to vary more than 1% of the subspectrum ppm range (or 0.004 ppm for the 0.4 ppm range of data being fit) and peak widths were constrained to be in the range of 0.5-2 Hz to better separate peak and baseline contributions. Furthermore, singlets were constrained from "passing" or completely overlapping each other during optimization — thereby limiting the possibility of the optimization converging on a non-useful local minimum. Phase correction was also constrained to within approximately 10° as it was expected that dedicated phase correction software

method (e.g. [31, 32, 33]) or present a novel approach

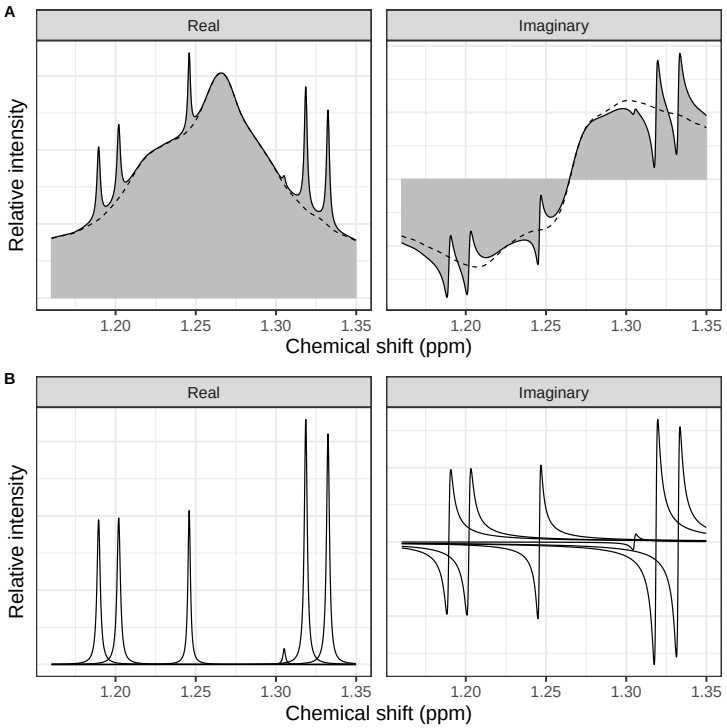

Figure 6: Demonstration of proposed algorithm on a subspectrum of a human plasma sample (see Materials and methods for more details). **A** Overlap of the overall fit (black lines) and estimated baseline (dashed line) over the observed data (gray fill). **B** Breakdown of the overall fit into constituent peaks.

should be capable of getting relatively close to the true value. Although rnmrfit assumes rational defaults for all of these constraints, the default values may not be suitable in all cases. Third, increasing the number of peaks in a single fit considerably slowed down the overall optimization time. The time it takes to achieve a reasonable overall fit increases approximately linearly with each peak and baseline knot — so 50-100 peaks can be fit within a couple of minutes. However, a small fraction of peaks require considerably more time than the rest of the peaks to achieve a good fit, slowing down computation to more than an hour to ensure high levels of precision for all peaks. One solution to this problem is computational — implementing the minimization function in Fortran sped up

computation by a factor of 10-100 but the resulting code had to be dropped as it was not easily portable across all platforms (a C++ version is planned for the near future). Another solution is to segment the overall spectrum into smaller regions prior to fitting, thereby preventing one or more problematic peaks from slowing down the fit.

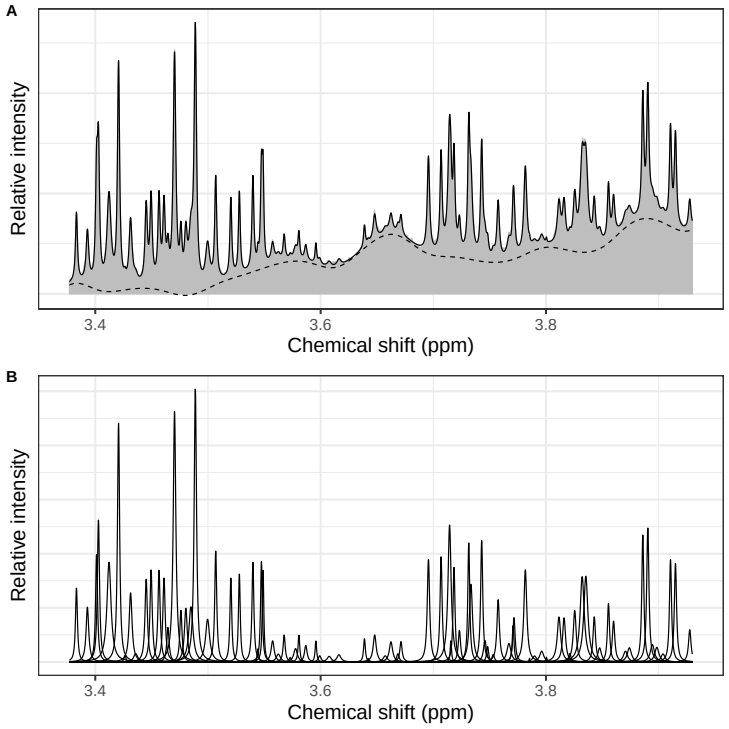

Figure 7: Demonstration of proposed algorithm on a subspectrum of a human plasma sample (see Materials and methods for more details). **A** Overlap of the overall fit (black lines) and estimated baseline (dashed line) over the observed data (gray fill). **B** Breakdown of the overall fit into constituent peaks.

5. Conclusion

The imaginary component of NMR data is not typically considered when fitting peaks in the frequency domain. However, adding this component is relatively straightforward and enables the incorporation of both phase and baseline

correction as well as apodization directly in the peak fitting process — resulting in the robust fit of small or overlapping peaks with the added computational efficiency of working in the frequency domain. Application to real data suggests that this approach can also result in dramatically higher precision than can be achieved with commonly used PERCH software (although the increase in precision is not uniform across different spectral regions or different spectra). The proposed algorithm has been implemented in the rnmrfit package for the R programming language (which has been made available on GitHub), but it is not specifically tied to this particular software. The implementation of this approach in other software is likely to be of benefit for a broad community of qNMR users.

Acknowledgment

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for the NSERC Postdoctoral Fellowship awarded to SS as well as the Lebanese CNRS and the research council of the Saint Joseph University of Beirut for their financial support of GH.

The authors would also like to thank two anonymous reviewers whose input helped improve the manuscript.

Conflict of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. K. Bharti, R. Roy, Quantitative 1H NMR spectroscopy, TrAC Trends in Analytical Chemistry 35 (2012) 5–26. doi:10.1016/j.trac.2012.02.007.

[2] D. S. Wishart, Quantitative metabolomics using NMR, TrAC Trends in Analytical Chemistry 27 (3) (2008) 228–237. doi:10.1016/j.trac.2007.12.001.

[3] G. F. Pauli, B. U. Jaki, D. C. Lankin, Quantitative 1H NMR: development and potential of a method for natural products analysis, Journal of Natural Products 68 (1) (2005) 133–49. doi:10.1021/np0497301.

[4] S. Singh, R. Roy, The application of absolute quantitative (1)H NMR spectroscopy in drug discovery and development, Expert Opinion on Drug Discovery 11 (7) (2016) 695–706. doi:10.1080/17460441.2016.1189899.

[5] T. Jézéquel, V. Joubert, P. Giraudeau, G. S. Remaud, S. Akoka, The new face of isotopic NMR at natural abundance, Magnetic Resonance in Chemistry 55 (2) (2017) 77–90. doi:10.1002/mrc.4548.

[6] F. Malz, H. Jancke, Validation of quantitative NMR, Journal of Pharmaceutical and Biomedical Analysis 38 (5) (2005) 813–23. doi:10.1016/j.jpba.2005.01.043.

[7] P. Giraudeau, Challenges and perspectives in quantitative NMR, Magnetic Resonance in Chemistry 55 (1) (2017) 61–69. doi:10.1002/mrc.4475.

[8] G. L. Bretthorst, Bayesian analysis. I. Parameter estimation using quadrature NMR models, Journal of Magnetic Resonance 88 (1990) 533–551. doi:10.1016/0022-2364(90)90287-J.

[9] G. L. Bretthorst, Bayesian analysis. II. Signal detection and model selection, Journal of Magnetic Resonance 88 (1990) 552–570. doi:10.1016/0022-2364(90)90288-K.

[10] G. L. Bretthorst, Bayesian analysis. III. Applications to NMR signal detection, model selection, and parameter estimation, Journal of Magnetic Resonance 88 (1990) 571–595. doi:10.1016/0022-2364(90)90289-L.

[11] D. V. Rubtsov, J. L. Griffin, Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy, Journal of Magnetic Resonance 188 (2) (2007) 367–79. doi:10.1016/j.jmr.2007.08.008.

[12] D. V. Rubtsov, C. Waterman, R. A. Currie, C. Waterfield, J. D. Salazar, J. Wright, J. L. Griffin, Application of a Bayesian deconvolution approach for high-resolution (1)H NMR spectra to assessing the metabolic effects of acute phenobarbital exposure in liver tissue, Analytical Chemistry 82 (11) (2010) 4479–85. doi:10.1021/ac100344m.

[13] K. Krishnamurthy, CRAFT (complete reduction to amplitude frequency table) – robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR, Magnetic Resonance in Chemistry 51 (12) (2013) 821–9. doi:10.1002/mrc.4022.

[14] Y. Matviychuk, E. von Harbou, D. J. Holland, An experimental validation of a Bayesian model for quantification in NMR spectroscopy, Journal of Magnetic Resonance 285 (2017) 86–100. doi:10.1016/j.jmr.2017.10.009.

[15] P. Mercier, M. J. Lewis, D. Chang, D. Baker, D. S. Wishart, P. Mercier, M. J. Lewis, D. Chang, D. Baker, D. S. Wishart, Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra, Journal of Biomolecular NMR 49 (3-4) (2011) 307–23. doi:10.1007/s10858-011-9480-x.

[16] C. Zheng, S. Zhang, S. Ragg, D. Raftery, O. Vitek, Identification and quantification of metabolites in (1)H NMR spectra by Bayesian model selection, Bioinformatics 27 (12) (2011) 1637–44. doi:10.1093/bioinformatics/btr118.

[17] J. Hao, W. Astle, M. De Iorio, T. M. D. Ebbels, BATMAN – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model, Bioinformatics 28 (15) (2012) 2088–90. doi:10.1093/bioinformatics/bts308.

[18] S. Ravanbakhsh, P. Liu, T. C. Bjorndahl, T. C. Bjordahl, R. Mandal, J. R. Grant, M. Wilson, R. Eisner, I. Sinelnikov, X. Hu, C. Luchinat, R. Greiner, D. S. Wishart, Accurate, fully-automated NMR spectral profiling for metabolomics, PloS One 10 (5) (2015) e0124219. doi:10.1371/journal.pone.0124219.

[19] V. V. Mihaleva, S.-P. Korhonen, J. van Duynhoven, M. Niemitz, J. Vervoort, D. M. Jacobs, Automated quantum mechanical total line shape fitting model for quantitative NMR-based profiling of human serum metabolites, Analytical and Bioanalytical Chemistry 406 (13) (2014) 3091–102. doi:10.1007/s00216-014-7752-5.

[20] M. A. Bernstein, S. Skora, C. Peng, A. Barba, C. Cobas, Optimization and automation of quantitative NMR data extraction, Analytical Chemistry 85 (12) (2013) 5778–5786. doi:10.1021/ac400411q.

[21] R. Dunkel, X. Wu, Identification of organic molecules from a structure database using proton and carbon NMR analysis results, Journal of Magnetic Resonance 188 (1) (2007) 97–110. doi:10.1016/j.jmr.2007.06.007.

[22] J. Keeler, Understanding NMR spectroscopy, John Wiley and Sons, Chichester, U.K., 2010.

[23] B. Auguie, D. Eddelbuettel, S. G. Johnson, RcppFaddeeva: 'Rcpp' Bindings for the 'Faddeeva' Package, R package version 0.1.0 (2015).
URL https://CRAN.R-project.org/package=RcppFaddeeva

[24] J. Ypma, H. W. Borchers, D. Eddelbuettel, nloptr: R interface to NLopt, R package version 1.0.4 (2017).
URL https://CRAN.R-project.org/package=nloptr

[25] U. Bussy, C. Thibaudeau, F. Thomas, J.-R. Desmurs, E. Jamin, G. S. Remaud, V. Silvestre, S. Akoka, Isotopic finger-printing of active pharmaceutical ingredients by 13C NMR and polarization transfer techniques as a tool to fight against counterfeiting, Talanta 85 (4) (2011) 1909–14. doi:10.1016/j.talanta.2011.07.022.

[26] E. Tenailleau, S. Akoka, Adiabatic 1H decoupling scheme for very accurate intensity measurements in 13C NMR, Journal of Magnetic Resonance 185 (1) (2007) 50–8. doi:10.1016/j.jmr.2006.11.007.

[27] T. Jézéquel, V. Silvestre, K. Dinis, P. Giraudeau, S. Akoka, Optimized slice-selective H NMR experiments combined with highly accurate quantitative C NMR using an internal reference method, Journal of Magnetic Resonance 289 (2018) 18–25. doi:10.1016/j.jmr.2018.02.002.

[28] D. Eddelbuettel, R. François, Rcpp: Seamless R and C++ integration, Journal of Statistical Software 40 (8) (2011) 1–18. doi:10.18637/jss.v040.i08.
URL http://www.jstatsoft.org/v40/i08/

[29] R. Laatikainen, M. Niemitz, W. J. Malaisse, M. Biesemans, R. Willem, A computational strategy for the deconvolution of NMR spectra with multiplet structures and constraints: analysis of overlapping 13C-2H multiplets of 13C enriched metabolites from cell suspensions incubated in deuterated media, Magnetic Resonance in Medicine 36 (3) (1996) 359–65.

[30] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Research 46 (D1) (2018) D608–D617. doi:10.1093/nar/gkx1089.

[31] A. Abbas, X.-B. Kong, Z. Liu, B.-Y. Jing, X. Gao, Automatic peak selection by a Benjamini-Hochberg-based algorithm, PloS One 8 (1) (2013) e53112. doi:10.1371/journal.pone.0053112.

[32] C. Cobas, F. Seoane, E. Vaz, M. A. Bernstein, S. Dominguez, M. Pérez, S. Sýkora, Automatic assignment of 1h-NMR spectra of small molecules, Magnetic Resonance in Chemistry 51 (10) (2013) 649–54. doi:10.1002/mrc.3995.

[33] S. Tikole, V. Jaravine, V. Rogov, V. Dötsch, P. Güntert, Peak picking NMR spectral data using non-negative matrix factorization, BMC Bioinformatics 15 (2014) 46. doi:10.1186/1471-2105-15-46.