CHALMERS

Surface Modeling for Quantification of TOCSY NMR Spectra

Master's Thesis in Engineering Mathematics and Computational Science

ANDREAS HENRIKSSON

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2015

Abstract

Nuclear magnetic resonance spectroscopy (NMR) is a measurement method, used in metabolomics, in which different chemicals in a sample give rise to peaks in a frequency spectrum. In metabolomics, one wants to measure a large number of body fluid samples in order to link metabolite concentrations to medical conditions. However, there is a trade-off between NMR measurement time, and peak separation in the resulting frequency spectra. It is therefore desirable to find mathematical methods to automatically quantify metabolites in NMR spectra with overlapping peaks. One possible way to accomplish this is to model frequency spectra parametrically using a sum of peak shape functions, and the purpose of this thesis is to study peak modeling in two-dimensional NMR spectra.

We model experimental data from blood serum using a simple theoretical model, and quantify four metabolites in the sample. Four ways of constraining optimization parameters are explored, ranging from very restricted to very free. The model shows good agreement with the data, and the estimated metabolite concentrations are fairly close to the expected values. While the most free constraining methods gave considerably smaller residuals, they were prone to overfitting. Because of this, the estimated concentrations showed higher accuracy when using more constrained parameters.

Acknowledgements

I would like to thank all the people at the Swedish NMR Centre in Gothenburg and especially my advisor Diana Bernin, for preparing samples, running the NMR experiments and providing guidance. I also want to thank my academic supervisor, Tobias Gebäck at Chalmers University of Technology.

Andreas Henriksson, Göteborg 2015-05-04

Contents

1 Introduction

Nuclear magnetic resonance spectroscopy (NMR) is an experimental method in which radio frequency pulses are applied to a sample in a magnetic field, and the response recorded. It can be used to determine molecular structure, to identify and quantify substances in a mixture, as well as a large number of other uses. In the field of metabolomics, one attempts to correlate variations in the concentrations of metabolites in the body with medical conditions. The quantitative properties of NMR are useful for this.

A common approach when doing this is to divide the NMR frequency spectrum into a number of bins, and use multivariate analysis methods such as Principal Component Analysis (PCA) to extract information from the resulting intensities [15]. However, what one would really like to do is to determine the absolute concentrations of metabolites in a large number of samples, and use this as the inputs of the multivariate analysis. In this work, we take a step in that direction with a qualitative study of one quantification method.

1.1 Background

In an NMR experiment, various substances in a sample result in a set of peaks at predictable locations in a frequency spectrum. The traditional one-dimensional NMR experiment suffers from severe peak overlap when applied to a complex sample such as blood plasma. This can be mitigated with two-dimensional experiments. As experimental equipment improves, large-scale application of 2D experiments becomes more and more viable. For this reason, automatic quantification with 2D NMR spectra is currently an active research topic in the context of metabolomics.

A large number of 2D NMR experiments are available and useful in different contexts. For quantification, there is a tradeoff between experiment time and peak separation in the spectrum. TOCSY (TOtal Correlated SpectroscopY) falls somewhere in the middle of the scale [5].

Direct numerical integration has previously been shown to work well when applied to spectra with well-separated peaks [8][10][11]. However, when peaks overlap, parts of other, unrelated peaks are inevitably included in the integral. Since this separation comes at the cost of increased experiment times, which can span tens of hours, it is desirable to find a quantification method that works in the presence of peak overlaps for large-scale metabolomic experiments. Using a parametric surface modeling approach, a set of peaks with certain positions and intensities can be found which fit the experimentally obtained spectrum.

1.2 Previous Work

Due to the high amount of peak overlap in one-dimensional NMR spectra, it seems necessary to include previously known information about metabolite peaks in the fitting procedure. A metabolite in a sample will usually result in several peaks at different positions in the spectrum, all of which have intensities proportional to its concentration. Because of this, intensities estimated from well-separated peaks can be used to resolve crowded parts of the spectrum. Curve fitting with previously known information has been used to quantify metabolites in rat brain extracts [4]. The approach is used in the BATMAN software [1][6], which is based on a Bayesian model.

In two-dimensional experiments, parametric modeling is a less explored topic. The Nmrglue Python package [7] can, among other things, perform surface fitting in several dimensions. Fast Maximum Likelihood Reconstruction (FMLR) [3] is an algorithm that finds potential peaks, and models them parametrically. On a synthetic sample, the algorithm was shown to posses greater accuracy when compared to numerical integration. McKenzie et al. use peak fitting to decide whether local maxima in a spectrum are true peaks or random noise, but not primarily for quantification [12].

1.3 Purpose

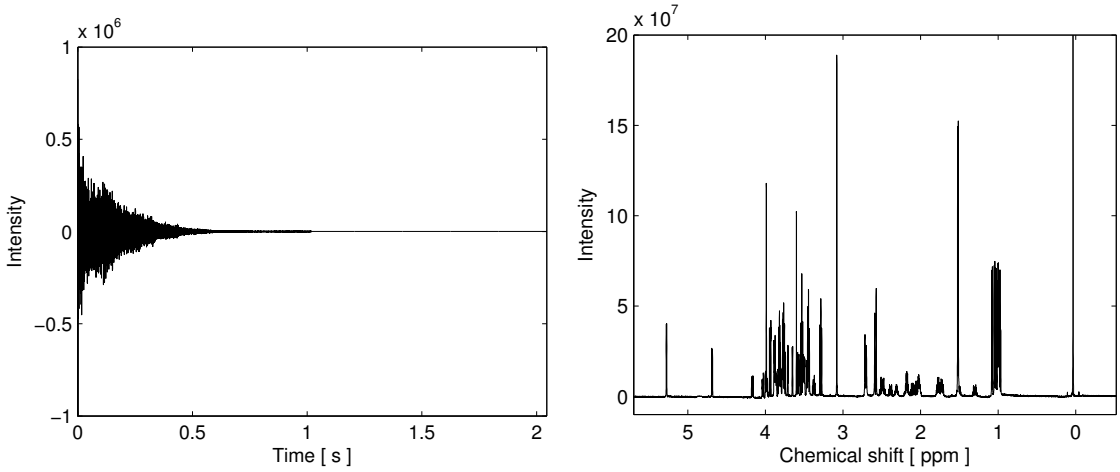
The purpose of this thesis is to qualitatively study the basic surface modeling approach. While authors have applied parametric modeling for quantification in NMR experiments, there is a lack of investigation into how optimization parameter constrains affect the calculated concentrations, and how well the underlying model fits experimental data.

2 Theory

2.1 One-Dimensional Experiment

In a one-dimensional NMR experiment, a sequence of RF pulses are applied to a sample, which results in a recorded signal called the Free Induction Decay (FID). Two coils at a 90 degree angle are used, and the FID takes the form of a sum of damped complex oscillations,

$$\tilde{S}(t) = \sum_{n=1}^{N} \hat{S}_n \exp(i\tilde{\phi}_n) \exp(i\Omega_n t) \exp(-R_n t).$$



(a) Free Induction Decay (FID) **(b)** Frequency spectrum

Figure 1: Illustration of a Free Induction Decay signal and the corresponding absorption mode frequency spectrum. The sample is a synthetic mixture with some of the metabolites found in blood.

The oscillation intensity $\hat{S}_n$ and frequency Ω_n determine the amplitude and position of a corresponding peak in the frequency spectrum. The decay rate R_n determines the width of the peak: A large R_n means that the signal decays quickly, which translates to a wide peak in the frequency spectrum. The oscillation phase $\tilde{\phi}_n$ is essentially arbitrary, but for the spectra considered in this thesis there is a phase correction constant ϕ such that $\tilde{\phi}_n \approx \phi$ for all relevant peaks.

We define $S(t)$ as the recorded FID multiplied with a phase correction factor $\exp(-i\phi)$,

$$S(t) = \exp(-i\phi)\tilde{S}(t) = \sum_{n=1}^{N} \hat{S}_n \exp(i\tilde{\phi}_n - i\phi) \exp(i\Omega_n t) \exp(-R_n t).$$

Let $\phi_n = \tilde{\phi}_n - \phi \approx 0$. After Fourier transformation, we obtain a spectrum of Lorentzian functions,

$$S(\omega) = \sum_{n=1}^{N} \hat{S}_n \exp(i\phi_n) \frac{1}{R_n + i(\omega - \Omega_n)}.$$

A Lorentzian can be divided into a real and imaginary part referred to as the absorption and dispersion lineshapes,

$$\frac{1}{R_n + i(\omega - \Omega_n)} = A_n(\omega) + iD_n(\omega)$$

where

$$A_n(\omega) = \frac{R_n}{R_n^2 + (\omega - \Omega_n)^2}\,, \quad D_n(\omega) = \frac{-(\omega - \Omega_n)}{R_n^2 + (\omega - \Omega_n)^2}.$$

Expanding the phase factor $\exp(i\phi_n)$ using Euler's formula, we get

$$S(\omega) = \sum_{n=1}^{N} \hat{S}_n(\cos\phi_n + i\sin\phi_n)(A_n(\omega) + iD_n(\omega)) \tag{1}$$

$$= \sum_{n=1}^{N} \hat{S}_n(\cos\phi_n A_n(\omega) - \sin\phi_n D_n(\omega) + i\sin\phi_n A_n(\omega) + i\cos\phi_n D_n(\omega)) \tag{2}$$

It is desirable to obtain a spectrum with absorption lineshapes for all peaks, since such peaks are more narrow and the shape is easier to interpret visually. We can see that this is the case for the real part of (2) when $\phi_n = 0$, and we therefore have

$$\mathrm{Re}\,[S(\omega)] \approx \sum_{n=1}^{N} \hat{S}_n A_n(\omega).$$

If we do not disregard the phase errors ϕ_n, we get the more general equation,

$$\mathrm{Re}\,[S(\omega)] = \sum_{n=1}^{N} \hat{S}_n L_n(\omega),$$

where

$$L_n(\omega) = \cos\phi_n\, A_n(\omega) - \sin\phi_n\, D_n(\omega)$$

is the real part of a Lorentzian function with phase error ϕ_n.

2.2 Two-Dimensional TOCSY Experiment

In two-dimensional experiments, several 1D FID's are recorded while varying a time t_1 in the pulse sequence. For TOtal Correlated SpectroscopY (TOCSY) using the States-TPPI method for phase discrimination, cosine and sine modulated signals are recorded separately,

$$S_{\cos}(t_1, t_2) = \sum_{n=1}^{N} \hat{S}_n \cos(\Omega_{n,1} t_1) \exp(-R_{n,1} t_1) \exp(i\Omega_{n,2} t_2) \exp(-R_{n,2} t_2)$$

$$S_{\sin}(t_1, t_2) = \sum_{n=1}^{N} \hat{S}_n \sin(\Omega_{n,1} t_1) \exp(-R_{n,1} t_1) \exp(i\Omega_{n,2} t_2) \exp(-R_{n,2} t_2).$$

To simplify the notation, we ignore phase errors for now. These functions are used to obtain an absorption mode lineshape in two dimensions, $A(\omega_1)A(\omega_2)$. Applying the Fourier transform in t_2, we get

$$S_{\cos}(t_1, \omega_2) = \sum_{n=1}^{N} \hat{S}_n \cos(\Omega_{n,1} t_1) \exp(-R_{n,1} t_1) \left(A_{n,2}(\omega_2) + i D_{n,2}(\omega_2)\right)$$
$$S_{\sin}(t_1, \omega_2) = \sum_{n=1}^{N} \hat{S}_n \sin(\Omega_{n,1} t_1) \exp(-R_{n,1} t_1) \left(A_{n,2}(\omega_2) + i D_{n,2}(\omega_2)\right).$$

A linear combination of the real part of both signal now yields

$$\begin{aligned} S(t_1, \omega_2) &= \mathrm{Re}\left[S_{\cos}(t_1, \omega_2)\right] + i\mathrm{Re}\left[S_{\sin}(t_1, \omega_2)\right] = \\ &= \sum_{n=1}^{N} \hat{S}_n (\cos(\Omega_{n,1} t_1) + i \sin(\Omega_{n,1} t_1)) \exp(-R_{n,1} t_1) A_{n,2}(\omega_2) = \\ &= \sum_{n=1}^{N} \hat{S}_n \exp(i\Omega_{n,1} t_1) \exp(-R_{n,1} t_1) A_{n,2}(\omega_2). \end{aligned}$$

After a final Fourier transformation in t_1, we get the desired 2D absorption mode lineshape,

$$S(\omega_1, \omega_2) = \sum_{n=1}^{N} \hat{S}_n A_{n,1}(\omega_1) A_{n,2}(\omega_2).$$

When phase errors $\phi_{n,1}, \phi_{n,2}$ are considered, the absorption mode lineshape is replaced with the mixed absorption-dispersion lineshape as in the one-dimensional case,

$$S(\omega_1, \omega_2) = \sum_{n=1}^{N} L_{n,1}(\omega_1, \phi_{n,1}) L_{n,2}(\omega_2, \phi_{n,2}) \tag{3}$$

where

$$L_{n,i}(\omega_i, \phi_{n,i}) = \cos\phi_{n,i} A_{n,i}(\omega_i) - \sin\phi_{n,i} D_{n,i}(\omega_i).$$

This is the equation that will be used to model the experimental data.

2.3 Apodization

Before Fourier transformation, it is common to multiply the FID with some function, a procedure known as apodization. The chosen function is usually decaying, which is what we consider here. This serves two purposes. Firstly, it can increase the signal-to-noise ratio by giving less weight to signal points toward the end of

the FID, which are mostly noise. Secondly, it prevents truncation artifacts in the transformed spectrum, when the acquisation time is too short to let the FID decay to zero. The drawback is that the width of the peaks increases, leading to more overlap between peaks.

For this thesis, $\exp(-R_A t)$ was chosen as the apodization function in both dimensions. This preserves the Lorentzian lineshapes in the spectrum. We can regard R_n in the previous sections as the sum of the experimental decay rate and the apodization decay rate.

2.4 Quantification

The concentration of a substance in the sample is proportional to the intensities of its resonances, denoted $\hat{S}_n$ above. In one-dimensional NMR this relationship is linear, and proportionality constants can be calculated using a single added reference compound with known concentration [9].

Things are more complicated in two-dimensional experiments, and the relationship might not even be linear. To combat this, several methods have been employed. For this thesis, a standard addition procedure was used, because it is simple and has been shown to give good results previously [5]. A model mixture, with known concentrations of the substances being studied, is prepared. The mixture is added to the sample in several different concentrations, and a spectrum is recorded for each increment. When the intensity of a peak has been estimated in all spectra, linear regression was used to find a value for the absolute concentration of a substance (illustrated in figure 2). This method works for non-linear relationships as well, by assuming linearity in the concentration region where the measurements are made.

In previous literature, a straight line has been used for the relationship between substance amount added and peak intensity. However, rather than following a straight line as expected, the peak intensities in the samples studied in this thesis sloped downward as seen in figure 3. This can be explained by the fact that the NMR samples were being diluted as the model mixture was added. To compensate, a slightly less intuitive relationship was investigated and used.

The concentration of a substance in NMR sample k is given by

$$c^{(k)} = \frac{c_p V_p + c_a V_a^{(k)}}{V_p + V_b + V_a^{(k)}}. \tag{4}$$

Here, c_p and c_a are the concentrations of the substance in the plasma and model mixture, V_p and V_b the volumes of plasma and buffer and V_a the volume of model mixture added. Our goal is to calculate c_p.

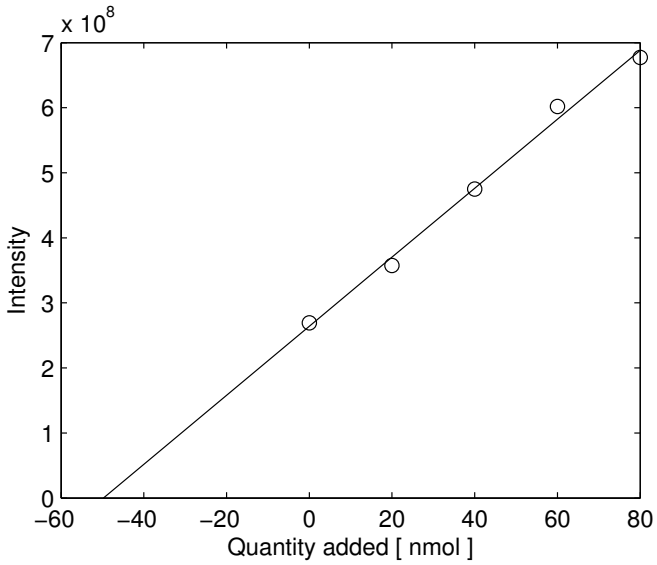
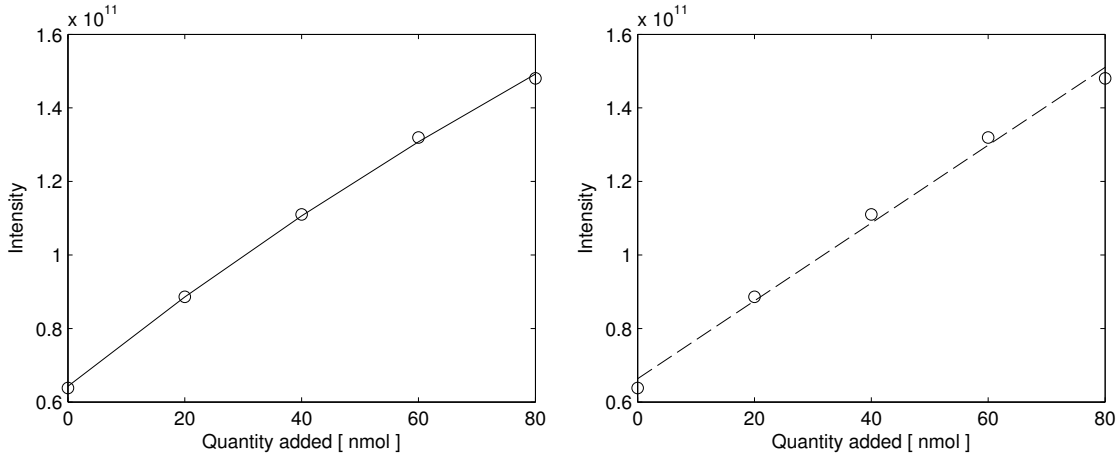


Figure 2: Example of how peak intensities in several spectra are used to calculate the absolute concentration by fitting a straight line. The intensity at 0 nmol is the peak intensity in the original sample. When the estimated line crosses the x-axis, the quantity added is the negative of the concentration (approximately 50 nmol). It can be interpreted as the molar quantity we need remove in order to reach 0 intensity.



(a) Proportional to concentration **(b)** Proportional to amount of substance

Figure 3: Comparison of two relationships between intensity and added substance quantity on data from the thesis. The peak studied is Alanine (1). **(a)** Line with decreasing slope, corresponding to equation (5). The actual quantity calculated with this line equation is not as simple to visualize as in figure 2. **(b)** Straight line, corresponding to (6).

We derive equations for two possible relationships between peak intensity and substance amount added. In the first case (figure 3a), the intensity is directly proportional to the concentration in the sample with an error term,

$$\alpha c^{(k)} = I^{(k)} + \varepsilon^{(k)}.$$

We define the total volume $V^{(k)} = V_p + V_b + V_a^{(k)}$ and combine the above equation with (4). This gives a linear equation in αc_p and α,

$$\alpha c_p \frac{V_p}{V^{(k)}} + \alpha \frac{c_a V_a^{(k)}}{V^{(k)}} = I^{(k)} + \varepsilon^{(k)}. \tag{5}$$

which can be solved by least squares.

In the second case (figure 3b), the peak intensity is instead proportional to the amount of substance $n^{(k)}$ in the NMR sample,

$$\beta n^{(k)} = \beta c^{(k)} V^{(k)} = I^{(k)} + \varepsilon^{(k)}.$$

Together with (4), we get a linear equation similar to (5) in βc_p and β,

$$\beta c_p V_p + \beta c_a V_a^{(k)} = I^{(k)} + \varepsilon^{(k)}. \tag{6}$$

Note that in an experimental procedure with no dilution, we have $V^{(k)} = const$ and the two cases above both reduce to a relationship with a straight line.

While the standard addition process is simple and accurate, the required spectrometer time is increased significantly. This is a large drawback for experiments in metabolomics. If fewer spectra are recorded for each sample, small errors in the intensity estimation can result in large concentration errors. One way to deal with this, if the lowered accuracy is acceptable, could be to record standard addition spectra for one sample, and use the resulting proportionality constants for all other samples.

2.5 Discretization

The FID is not recorded continuously, but as a sequence of equally spaced samples. These data points are transformed using the discrete Fourier transform.

An explicit formula for the resulting spectrum can be calculated, but with the frequencies and number of data points in our case, it is sufficient to consider the ideal, continuous spectrum evaluated at discrete points.

2.6 Zero-filling

In NMR, it is common practice to increase the vector of sampled FID points to twice its size by appending zeros before applying the Fourier transform. This doubles the number of mesh point and makes the peaks smoother, and increases the amount of information in the absorption part of the spectrum by incorporating information from the dispersive part [2]. Zero-filling to more than twice the amount of data points is also common. In that case, the new mesh points are only a result of an implicit interpolation process, and lack additional information. This can be helpful when interpreting a spectrum visually, but results in an unnecessarily increased computational cost when applied together with surface modeling.

2.7 Units

Above, we have derived equations in terms of $\mathrm{rad\,s^{-1}}$. This is convenient for mathematical manipulation, however from a physical perspective it makes more sense to write frequencies in Hz. The conversion formula is, by definition, $\omega = 2\pi f$, where ω is in $\mathrm{rad\,s^{-1}}$ and f in Hz.

In NMR it is common to specify frequencies with the chemical shift δ. Chemical shift is defined as

$$\delta = \frac{f - f_r}{f_s}$$

where f_r is the resonance frequency of some reference compound, and f_s is the spectrometer base frequency, typically around 300 to 900MHz. This is a very small value, so the chemical shift is generally given in parts per million, ppm. The advantage of this scale is that it is independent of spectrometer base frequency, so peak positions remain constant between different spectrometers.

The decay constant R is usually specified as the absorption mode peak width at half height, given in Hz. A straight-forward calculation yields the width of a peak, $2R\,[\mathrm{rad\,s^{-1}}]$, and hence R/π [Hz].

3 Method

A model mixture was prepared with four metabolites and caffeine, see table 2. Five NMR samples were prepared with blood serum, buffer and different amounts of model mixture, as shown in table 1. The five samples resulted in five spectra, which were used to determine absolute concentrations from the peak intensities.

The raw FID's were processed using a custom Matlab script. An exponential apodization function was applied, the FID's were zero-filled to twice the number of points, and the first FID point was multiplied by 0.5 in the direct and indirect

	Blood Serum	Buffer	Model Mixture
Spectrum 1	100 µl	100 µl	0 µl
Spectrum 2	100 µl	100 µl	10 µl
Spectrum 3	100 µl	100 µl	20 µl
Spectrum 4	100 µl	100 µl	30 µl
Spectrum 5	100 µl	100 µl	40 µl

Table 1: Volumes of blood serum, buffer and model mixture in the five measured NMR samples.

Caffeine	898 µM
Alanine	2000 µM
Glycine	1200 µM
Choline	50 µM
Lysine	800 µM

Table 2: Concentrations of the studied substances in the model mixture

dimensions. Furthermore, the process described in the matNMR code was used to handle Bruker's digital filtering [14].

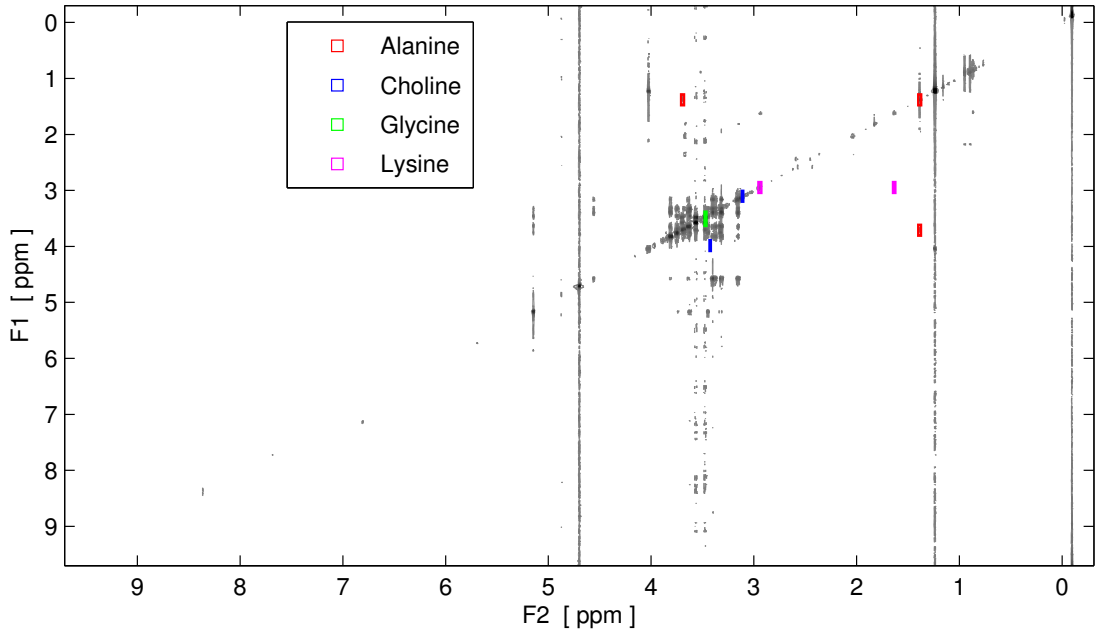
Peaks were assigned manually, using a spectrum of just the model mixture as reference.

3.1 Surface modeling

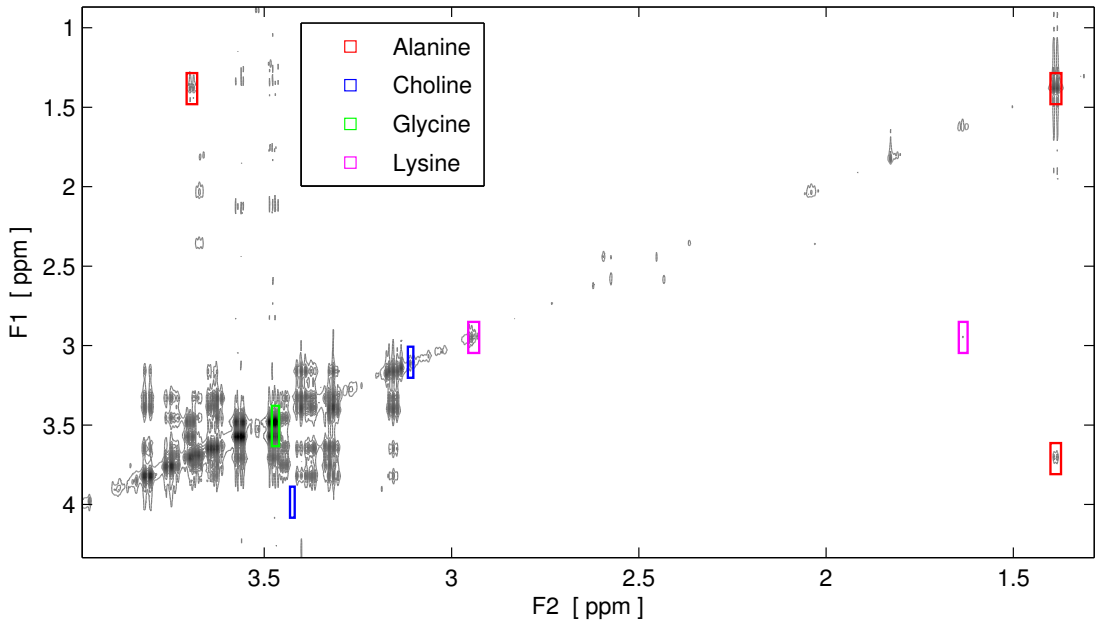
The spectra were divided into regions around the peaks of interest, and other peaks within the regions were identified. The sizes of the regions were chosen to minimize spectrum contributions of peaks outside the regions. These regions comprised only a tiny portion of the full frequency spectrum as shown in figure 4.

A sum of peak shapes of the form (3) was then fitted to each region in the measured spectra. All optimization was done with Matlab's fmincon function, minimizing the square sum of each spectral point in the residual. With this choice of objective function, the optimal intensities $\hat{S}_n$ for each peak could be calculated by linear least-squares in each iteration, given values for the other parameters.

Four different ways of constraining the optimization variables were investigated



(a) Full frequency spectrum



(b) Frequency spectrum around the studied regions

Figure 4: Contour plots of a frequency spectrum with the studied regions shown as rectangles, barely visible in (a). In general, one region can contain peaks from several different metabolites of interest, but this was not the case here. The prominent vertical patterns in the full spectrum are noise caused by large peaks (known as t_1 noise).

	Decay (R_1, R_2)	Phase (ϕ_1, ϕ_2)	Peak position (Ω_1, Ω_2)
1.	Constant	Zero	Same values in all spectra
2.	Constant	One value per spectrum	Free
3.	Constant	Free	Free
4.	Free	Free	Free

Table 3: Four ways of constraining the optimization parameters, ranging from most to least restrictive.

as seen in table 3, referred to as method 1, 2, 3 and 4.

3.2 Calculating Absolute Concentrations

A standard addition process with peak intensity proportional to substance concentration was used to determine the absolute concentrations of metabolites as described in section 2.4.

To justify equation (5), four regions without any peaks from the model mixture were selected. The total intensity of all the peaks in the regions was estimated by numerical integration in each of the 5 spectra. Numerical integration was used to make sure that the decrease in estimated intensity was not caused by the surface fitting algorithm. In this case, with no added intensity from the model mixture, equations (5) and (6) reduce to

$$\alpha c_p \frac{V_p}{V^{(k)}} = I^{(k)} + \varepsilon^{(k)} \qquad \text{and} \qquad \beta c_p V_p = I^{(k)} + \varepsilon^{(k)}$$

respectively. In these equations, V_p is constant (100 µl), c_p, α and β are unknown but constant, while $V^{(k)}$ increases from 200 µl to 240 µl

Hence, if the intensity is proportional to concentration, we should see decreasing intensity as more model mixture is added. On the other hand, if the intensity of a resonance is proportional to the amount of substance in the sample, the intensity should remain constant as the sample is diluted.

	Concentration [µM] (R^2)			
	1.	2.	3.	4.
Alanine (1)	455 (0.998)	448 (0.999)	455 (0.997)	449 (0.998)
Alanine (2)	412 (0.999)	435 (0.999)	441 (0.998)	434 (0.997)
Alanine (3)	452 (0.996)	469 (0.996)	483 (0.968)	468 (0.943)
Alanine (4)	390 (0.998)	399 (0.997)	384 (0.983)	280 (0.972)
Alanine (5)	409 (0.999)	344 (0.995)	359 (0.995)	332 (0.995)
Alanine (6)	379 (0.998)	435 (0.998)	423 (0.999)	420 (0.990)
Mean	416	422	424	397
Std.	28.6	40.7	42	67.7
Reference	427.2 ± 84.4			
Choline (1)	10.9 (0.927)	11.4 (0.960)	10.5 (0.977)	10.1 (0.640)
Choline (2)	13.6 (0.996)	12.1 (0.999)	11.2 (0.993)	12.1 (0.994)
Mean	12.2	11.8	10.9	11.1
Std.	1.34	0.332	0.33	1.01
Reference	14.5 ± 5.3			
Glycine (1)	737 (0.983)	770 (0.967)	757 (0.857)	1130 (0.621)
Reference	325.4 ± 126.8			
Lysine (1)	145 (0.997)	138 (0.998)	153 (0.988)	124 (0.994)
Lysine (2)	153 (1.000)	152 (1.000)	179 (0.981)	126 (0.968)
Lysine (3)	141 (0.999)	108 (0.976)	111 (0.981)	158 (0.890)
Lysine (4)	167 (0.998)	218 (0.957)	176 (0.967)	148 (0.953)
Mean	152	154	155	139
Std.	10.1	40.3	27	14.5
Reference	178.6 ± 58.2			

Table 4: Calculated concentrations for each peak of the studied metabolites, using the four methods described in section 3.1. R^2 values are shown inside the parentheses. Reference values are from the Serum Metabolome Database [13].

4 Results and Discussion

4.1 Calculated Concentrations

Table 4 summarizes the calculated concentrations for the identified peaks, using the four methods described in section 3.1. The values are within the expected range as found in previous studies, with the exception of glycine, which is overestimated. However, the concentrations vary considerably between peaks originating from the same substances for all methods, which indicates that the quantification is not very accurate. This is not necessarily due to errors in the peak intensity estimation, but could be a result of non-linearity in this particular NMR experiment.

4.2 Quality of Quantification

To estimate the quality of the quantification, R^2 values were calculated when fitting lines to the calculated intensities for the standard addition procedure. These are presented in table 4 inside parentheses next to the concentrations. The values were generally good, with many values above 0.99. The less restricted methods (3 and 4) show the worst linearity, likely due to overfitting. The worst R^2 value (0.621) is for the only glycine peak, using method 4. This region is the most complicated one studied, with 6 partly overlapping peaks.

4.3 Quality of Fits

An estimate for how well a modeled surface fits the spectrum is the ratio of the maximum value of the magnitude of the residual, and the peak height. Such values were calculated for all methods and optimization regions and are visualized in figure 6. In this section, we illustrate and explain some errors which caused large residuals.

All methods gave uncharacteristically large residuals around Choline (1), because of large ridges originating from peaks outside the region which were not modeled (figure 5). This explains the low R^2 value of method 4 on this peak, the many unconstrained parameters are distorted from their physical meaning in order to model the ridges.

Method 1, which was the most constrained, had the largest residuals as expected. In figure 7, the difference between model and experimental data is visible, and the residual is quite large. Figure 8 shows the same area optimized with method 2. The phase error is modeled, and peaks are allowed to move slightly in each spectrum of the standard addition process.

Another region with large residuals was around the Lysine (2), (3) and (4) peaks (figure 9). In this case, the shape of the residual suggested that the decay

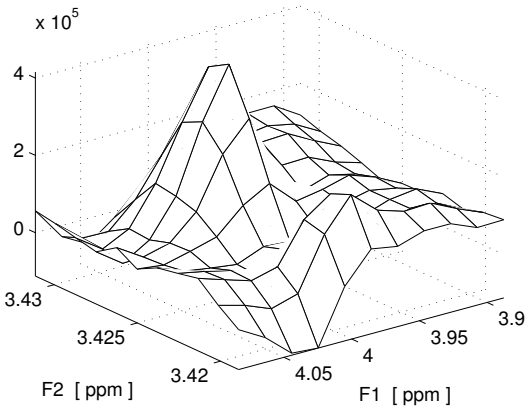


Figure 5: The region around Choline (1). The larger peak is choline, the smaller peak is an unknown substance. Ridges along the F_2 axis from large peaks outside the region are seen in the background and foreground.

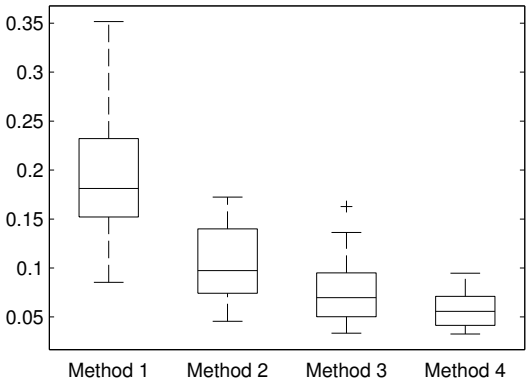
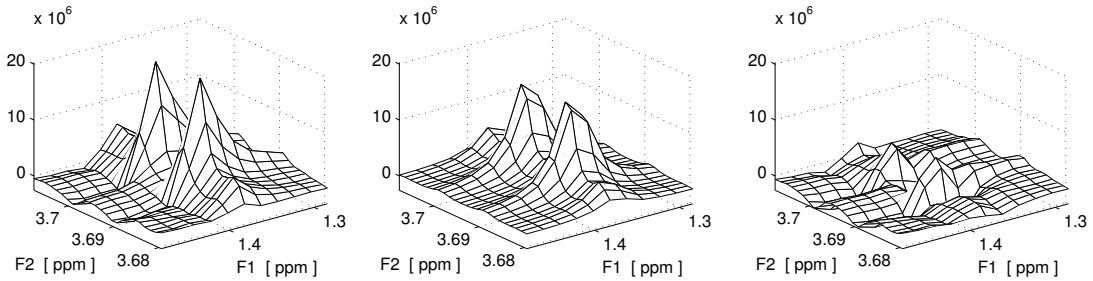
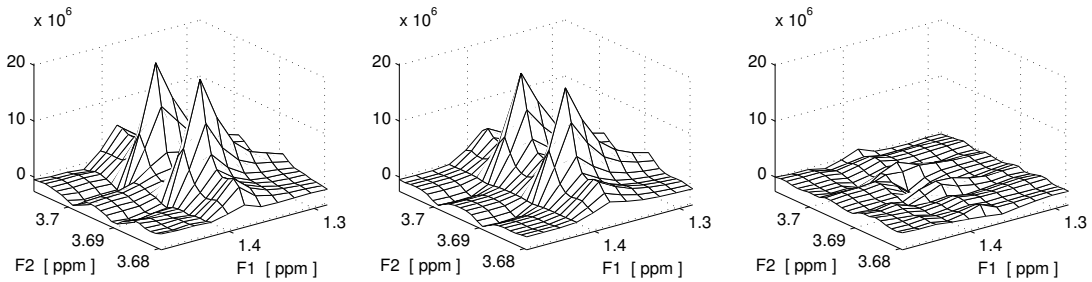


Figure 6: Boxplot of the maximum absolute value of the residual as a fraction of peak height for the four methods, excluding the Choline (1) region.



(a) Experimental data **(b)** Model **(c)** Residual

Figure 7: Surface fit using method 1 on the region around Alanine (3) and Alanine (4). The peak shapes and phase effects are not modeled correctly.



(a) Experimental data **(b)** Model **(c)** Residual

Figure 8: Surface fit using method 2 on the region around Alanine (3) and Alanine (4). When small variations in phase and peak position are allowed, the residual shrinks.

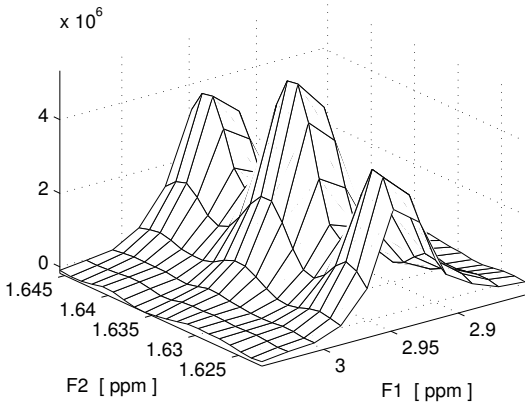


Figure 9: Experimental data of the region around Lysine (2), (3) and (4), in standard addition spectrum 5. All three peaks originate from lysine.

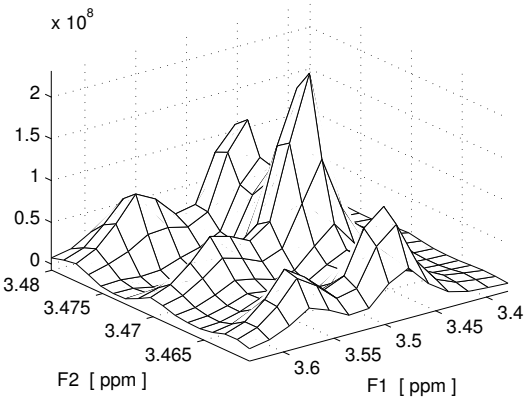
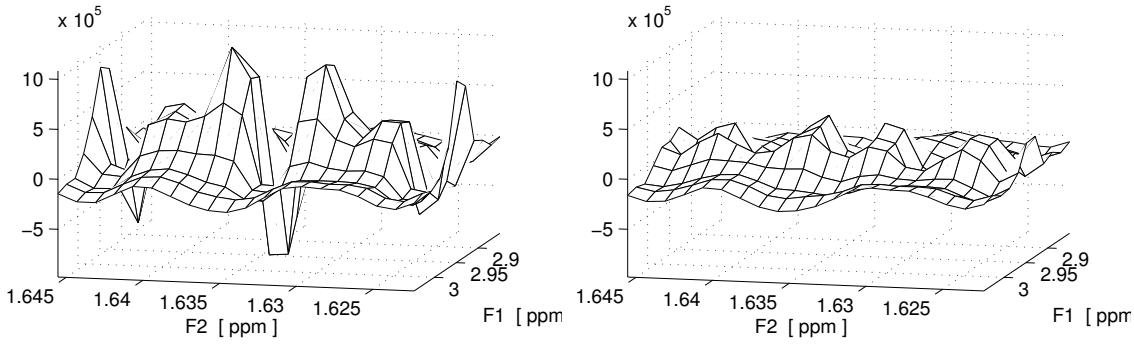


Figure 10: Region around the glycine peak. The largest peak is glycine, the five smaller peaks are unknown substances. This was the most complicated region considered in this thesis.

rate constants were incorrect for these peaks. Alternative R_1 and R_2 constants were tried, resulting in improved residuals for methods 1-3 (method 4 already allows the decay rate parameters to vary freely). Figure 11 shows the difference for method 2. In table 5 we see the different concentration estimates for original and adjusted decay rates, and the change is quite large for each individual peak. Method 1 was least sensitive to erronous decay rate constants, likely because its constrained nature prevents overfitting. This result is interesting, because it shows that the assumption of a constant decay rate for all peaks in the spectrum is not accurate enough (in this NMR experiment) if very small residuals are desired.

The smallest residuals are given by method 4 around the glycine peak. This is



(a) Original Decay Rate **(b)** Adjusted Decay Rate

Figure 11: Enlarged residuals when using method 2 on the region in figure 9, comparing different decay rate constants. The shape of the residual in (a) suggests that the three modeled peaks are too narrow in the F_2 dimension. When a larger decay rate is used in (b), the residual improves.

not unexpected, as the six peaks in this region offer many degrees of freedom in the optimization parameters. Due to the low R^2 value for this peak compared to other methods, we can conclude that the estimated model parameters lack physical significance, and are merely a result of overfitting.

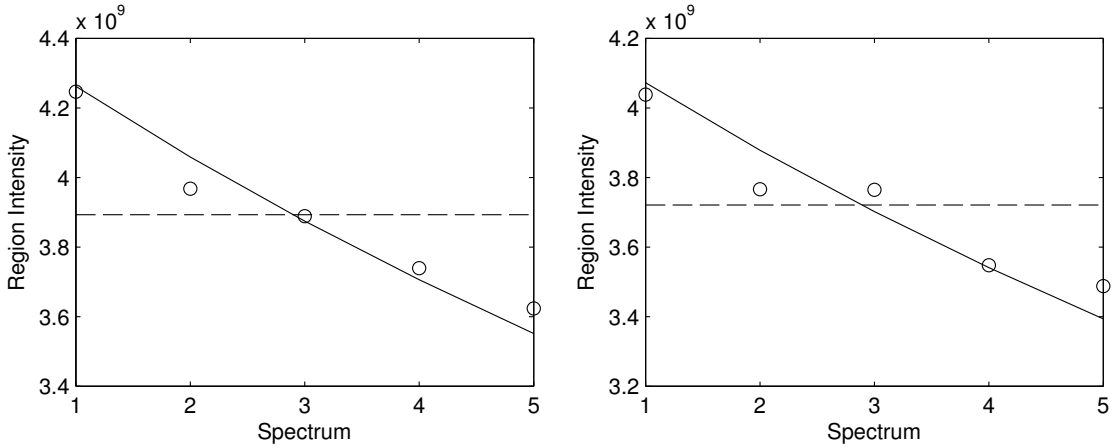
4.4 Effects of Dilution

Figure 12 shows the intensity of peaks in four regions where the standard addition mixture had no peaks. In all regions, the intensity decreases as the sample is diluted. Two models relating quantity to peak intensity are compared, intensity proportional to concentration (equation 5) and intensity proportional to amount of substance (equation 6). It is obvious that the first model fits the data better. It is possible that there is a better model, but such an investigation is out of scope of this thesis.

Not accounting for this dilution effect results in large differences in the calculated concentrations, as shown in table 6. For future work, it might be better to mix the samples in a way where this additional complication is avoided.

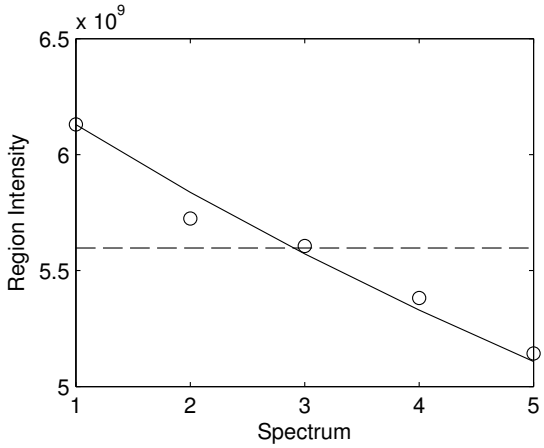
5 Conclusion

In this thesis, we have quantified four metabolites in blood serum using parametric peak modeling in two-dimensional NMR spectra. The estimated concentrations are within the expected ranges for three of the metabolites, but overestimated for one (glycine).

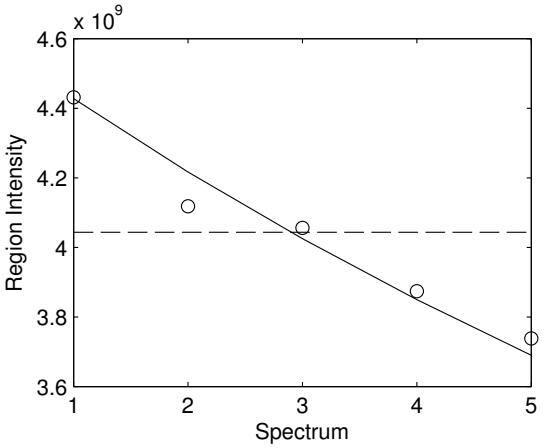


(a) Region 1

(b) Region 2



(c) Region 3



(d) Region 4

Figure 12: Total intensity in four regions as the sample is diluted. Solid line: Expected intensities according to equation (5). Dashed line: Intensities according to equation (6).

		Concentration [µM] (R^2)	
		Original Decay Rate	Adjusted Decay Rate
Method 1	Lysine (2)	153 (1.000)	155 (0.999)
	Lysine (3)	141 (0.999)	135 (0.999)
	Lysine (4)	167 (0.998)	175 (0.998)
	Mean	154	155
	Std.	13.2	20.1
Method 2	Lysine (2)	152 (1.000)	156 (1.000)
	Lysine (3)	108 (0.976)	136 (0.952)
	Lysine (4)	218 (0.957)	186 (0.983)
	Mean	160	159
	Std.	55.5	25.6
Method 3	Lysine (2)	179 (0.981)	151 (0.989)
	Lysine (3)	111 (0.981)	132 (0.974)
	Lysine (4)	176 (0.967)	181 (0.989)
	Mean	155	155
	Std.	38.2	24.7

Table 5: Comparison of estimated concentrations using different decay rates (R_1, R_2). The original decay rate was suitable for most other peaks, while the adjusted decay rate has been chosen to improve the residuals for Lysine (2), (3) and (4).

In section 2.2, we derived a peak shape model starting from two simple equations. Generally, this model allowed peak shapes which were very close to the experimental data, but artifacts in the residual were still visible even with the least constrained parameters. This could be because of inhomogeneities in the NMR instrument's magnetic field (shimming errors).

When the four methods are compared, a clear trend emerges. With less constrained optimization parameters, the residuals will be smaller, but it also means that the risk of overfitting is larger. For the spectrum regions studied in this thesis, with relatively well separated peaks of similar sizes, the constrained methods (1 and 2) are a more reliable choice. However, if one is looking to measure the intensity of a small peak overlapping a much larger one, the larger error between

	Concentration [µM] (R^2)	
	Dilution compensation	No compensation
Alanine (1)	455 (0.998)	637 (0.992)
Alanine (2)	412 (0.999)	570 (0.997)
Alanine (3)	452 (0.996)	632 (0.991)
Alanine (4)	390 (0.998)	540 (0.992)
Alanine (5)	409 (0.999)	567 (0.994)
Alanine (6)	379 (0.998)	521 (0.997)
Reference	427.2 ± 84.4	
Choline (1)	10.9 (0.927)	15.5 (0.898)
Choline (2)	13.6 (0.996)	19.3 (0.994)
Reference	14.5 ± 5.3	
Glycine (1)	737 (0.983)	1300 (0.966)
Reference	325.4 ± 126.8	
Lysine (1)	152 (0.997)	209 (0.996)
Lysine (2)	153 (1.000)	211 (0.998)
Lysine (3)	141 (0.999)	193 (0.995)
Lysine (4)	167 (0.998)	232 (0.995)
Reference	178.6 ± 58.2	

Table 6: Changes in estimated concentrations when compensating for sample dilution, using optimization method 1.

model and experimental data could be a problem.

5.1 Further work

To deal with shimming errors, previous authors have suggested that peaks be modeled with a partly Lorentzian, partly Gaussian lineshape [3][12]. This was not attempted in this work, as the data was already very close to the ideal Lorentzian shape.

Another challenge is to carry out the surface fitting in a way that allows variations in certain parameters, but avoids the overfitting problem. One solution could be to penalize parameter values in the objective functions, but finding a way to do this in a robust, automated and physically relevant fashion is not trivial.

In a larger context, the question of how to completely automate the quantification process remains. Algorithms for locating peaks in two-dimensional NMR spectra have been developed for the purpose of identifying substances, and combining such an algorithm with surface modeling is a logical next step. It is not clear if such a set up would be able to handle the crowded spectra resulting from fast NMR experiments with complicated mixtures, like one is likely to encounter in metabolomics.

References

[1] W. Astle, M. D. Iorio, S. Richardson, D. Stephens, and T. Ebbels. A bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500):1259–1271, 2012.

[2] E. Bartholdi and R. Ernst. Fourier spectroscopy and the causality principle. *Journal of Magnetic Resonance (1969)*, 11(1):9 – 19, 1973.

[3] R. A. Chylla, K. Hu, J. J. Ellinger, and J. L. Markley. Deconvolution of two-dimensional NMR spectra by Fast Maximum Likelihood Reconstruction: Application to quantitative metabolomics. *Analytical Chemistry*, 83(12):4871–4880, 2011. PMID: 21526800.

[4] R. A. de Graaf, G. M. I. Chowdhury, and K. L. Behar. Quantification of high-resolution 1H NMR spectra from rat brain extracts. *Analytical Chemistry*, 83(1):216–224, 2011. PMID: 21142125.

[5] P. Giraudeau. Quantitative 2D liquid-state NMR. *Magnetic Resonance in Chemistry*, 52(6):259–272, 2014.

[6] J. Hao, W. Astle, M. De Iorio, and 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):2088–2090, 2012.

[7] J. Helmus and C. Jaroniec. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *Journal of Biomolecular NMR*, 55(4):355–367, 2013.

[8] I. A. Lewis, S. C. Schommer, B. Hodis, K. A. Robb, M. Tonelli, W. M. Westler, M. R. Sussman, and J. L. Markley. Method for determining molar concentrations of metabolites in complex solutions from two-dimensional 1H–13C NMR spectra. *Analytical Chemistry*, 79(24):9385–9390, 2007. PMID: 17985927.

[9] F. Malz and H. Jancke. Validation of quantitative NMR. *Journal of Pharmaceutical and Biomedical Analysis*, 38(5):813 – 823, 2005. Quantitative NMR Spectroscopy principles and applications.

[10] E. Martineau, P. Giraudeau, I. Tea, and S. Akoka. Fast and precise quantitative analysis of metabolic mixtures by 2D 1H INADEQUATE NMR. *Journal of Pharmaceutical and Biomedical Analysis*, 54(1):252 – 257, 2011.

[11] E. Martineau, I. Tea, S. Akoka, and P. Giraudeau. Absolute quantification of metabolites in breast cancer cell extracts by quantitative 2D 1H INADEQUATE NMR. *NMR in Biomedicine*, 25(8):985–992, 2012.

[12] J. McKenzie, A. Charlton, J. Donarski, A. MacNicoll, and J. Wilson. Peak fitting in 2D 1H–13C HSQC NMR spectra for metabolomic studies. *Metabolomics*, 6(4):574–582, 2010.

[13] N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T. L. Pedersen, S. R. Smith, F. Bamforth, R. Greiner, B. McManus, J. W. Newman, T. Goodfriend, and D. S. Wishart. The human serum metabolome. *PLoS ONE*, 6(2):e16957, 02 2011.

[14] J. D. van Beek. matNMR: A flexible toolbox for processing, analyzing and visualizing magnetic resonance data in Matlab®. *Journal of Magnetic Resonance*, 187(1):19 – 26, 2007.

[15] B. Worley and R. Powers. Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.