Model-based signal tracking in the quantitative analysis of time series of NMR spectra

Denise Meinhardt[a,b], Henning Schröder[a,b], Jan Hellwig[a,b], Ellen Steimers[c], Anne Friebel[c], Torsten Beweries[b], Mathias Sawall[a], Erik von Harbou[d], Klaus Neymeyr[a,b]

[a] *Universität Rostock, Institut für Mathematik, 18057 Rostock, Germany*
[b] *Leibniz-Institut für Katalyse e.V., 18059 Rostock, Germany*
[c] *Technische Universität Kaiserslautern, Laboratory of Engineering Thermodynamics (LTD), 67663 Kaiserslautern, Germany*
[d] *Technische Universität Kaiserslautern, Laboratory of Reaction and Fluid Process Engineering, 67663 Kaiserslautern, Germany*

Abstract

Hard modeling of NMR spectra by Gauss-Lorentz peak models is an effective way for dimensionality reduction. In this manner high-dimensional measured data are reduced to low-dimensional information as peak centers, amplitudes or peak widths. For time series of spectra these parameters can be assumed to be smooth functions in time. We suggest to model these time-dependent parameter functions by cubic spline functions, which makes a stable quantitative analysis of NMR series possible even for crossing, highly overlapping peaks. Applications are presented for the batch distillation of methanol and diethylamine, and the reaction of acetic anhydride with 2-propanol.

Key words: NMR, Model-based decomposition, Dimensionality reduction, Time series analysis.

1. Introduction

The spectroscopic monitoring of chemical reaction systems with subsequent analysis of the spectra is a central tool for the elucidation of reaction mechanisms, the identification of intermediate species or the detection of by-products. Ideally, the spectroscopic measurement is performed in-situ on the reaction system and is combined with a computer-aided recording of the spectra. Such an application is especially of interest in the context of online-monitoring of chemical processes [17]. However, spectroscopic measurements at a high resolution in time and frequency lead to large data sets. Efficient algorithms are needed for the processing of these data sets. For UV/Vis, FT-IR or Raman spectroscopy, so-called Multivariate Curve Resolution (MCR) methods can be applied very successfully [7, 13]. With MCR methods a data matrix $D \in \mathbb{R}^{m\times n}$ of a chemical s-component system can be factorized into a matrix factor $C \in \mathbb{R}^{m\times s}$ containing in its columns the concentration profiles of the pure components and a second factor $S \in \mathbb{R}^{s\times n}$ with the pure component spectra in its rows. Therein n is the number of data points and m the number of measured spectra or time steps. Here we focus on time series of Nuclear Magnetic Resonance (NMR) spectroscopic measurements. For these an application of MCR methods is not possible in most cases, due to changing peak shapes and changing positions. Nevertheless, the idea of a simultaneous extraction of peak information along the time axis remains. Typically, the fitted peak parameters of one time layer are used as starting values for an optimization of the following layer, e.g., used in the popular Indirect Hard Modeling (IHM) method by Kriesten, Alsmeyer, Marquardt et al. [1, 11, 12].

Based on this, we propose a stronger temporal coupling of the parameters of the applied Gauss-Lorentz peak model by connecting them with spline functions along the time axis. Parameter values then only need to be optimized at a few selected time points. To determine the residuals of the model fit, the intermediate values are interpolated. The reduced number of unknowns makes the simultaneous analysis of a complete time series of NMR spectra possible and thus the challenging cases of overlapping and intersecting peaks can often be solved as well. Such moving and intersecting peaks can occur in many fields, especially when technical systems are studied, e.g., for changing pH values or temperatures. With this automated and relatively fast approach a pure component analysis is supported. The decomposition process can also be supported by suitable additional information like pure component spectra, but this

is not mandatory for the algorithm. The intended data analysis is applicable to data from high-field spectrometers and low-field benchtop devices.

1.1. Main idea of the spline-based decomposition approach

If a time series of NMR spectra is considered, then the most straightforward modeling approach is to build peak models for each spectrum. However, questions on the traceability of overlapping peaks over time result in large and numerically costly optimization problems with respect to the individual peak models. There are more effective ways than building peak models separately and independently for each time layer. The crucial point is that the parameters of the peak models can be assumed to be smooth functions along the time axis. This is illustrated on the right-hand side of Figure 1. The figure shows a three-dimensional plot of the spectra series together with red lines that mark the peak centers and run through the peak maxima. For the first of the three peaks Figure 2 shows the associated time-dependent smooth functions for the half-width, the peak center and the peak height. Our approach is to use so-called spline functions, namely piecewise interpolating polynomials, in order to represent these functions. Such splines only depend on few coefficients. This enables a low-dimensional representation of the full spectra series. In other words, the dimensions of the associated optimization problems are much smaller compared to working with non-coupled peak models generated for each spectrum separately. Moreover, the approach aims at tracing overlapping peaks and at decomposing NMR time series in order to identify the underlying pure components.

1.2. Organization of the paper

In order to make the mathematical algorithm easier to understand, we present a step-by-step approach. The peak modeling for single NMR spectra is introduced in Sec. 2. This is followed by the generalization to time series of NMR spectra in Sec. 2.3. The modeling of parameters by spline functions is described in Sec. 3, and the steps of the algorithm are presented. The new algorithm is applied to two series of NMR spectra in Sec. 4.

1.3. Notation

We use the following notation:

x	variable on the ppm-axis.
d	the real-valued single NMR signal after phase and baseline correction.
D	$\in \mathbb{R}^{n\times m}$ the real-valued time series of NMR spectra.
m	number of time layers.
n	number of data points.
N	maximal number of peaks per spectrum in a time series.
$p_{i,j}$	the parameter vector $(a_{i,j}, b_{i,j}, c_{i,j}, \lambda_{i,j})$ of the ith peak at the jth time layer.

1.4. Model data set

A simulated data set D_0 is used for accompanying explanations, see the left-hand side of Figure 1. The matrix $D_0 \in \mathbb{R}^{500\times 61}$ stores the spectra column-wise for 61 time steps with each 500 data points. The data set contains three moving and intersecting peaks.

2. Building peak models for NMR spectra series

2.1. Basic peak models

In this section, the model building for single NMR spectra is explained. The spectral line profile of a Lorentzian [10] versus the ppm-axis x reads

$$l(x; a, b, c) = \frac{ca}{a^2 + (x-b)^2}.$$

The half-width is a, the center value b and the height c. The left-hand side of Figure 3 shows a series of simulated spectra whose single spectra are all formed by $N = 3$ Lorentz curves. Typically, spectroscopic measurements are affected by noise, inhomogeneities of the static magnetic field and other sources of perturbations so that perfect

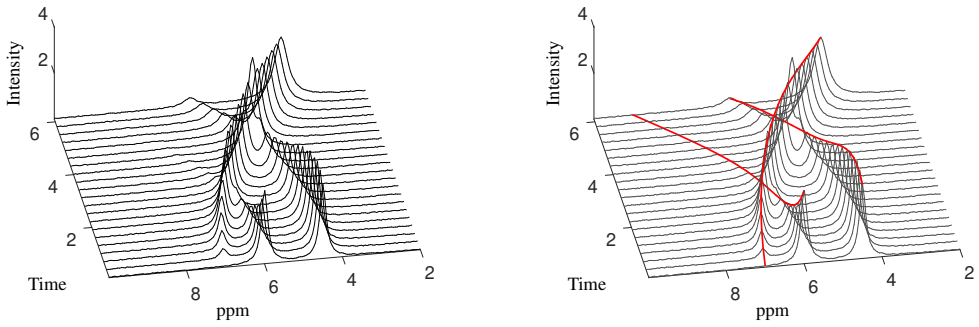


Figure 1: Left: Simulated data set $D_0 \in \mathbb{R}^{500\times 61}$ of an NMR time series with three intersecting peaks. Right: Spline-based modeling idea. The peak centers are marked by a red line.

Lorentzians cannot be assumed. Instead, an improved approximation is possible by a convex combination of a Gauss and a Lorentz curve with the form

$$h(x; a, b, c, \lambda) = \lambda g(x; a, b, c) + (1 - \lambda) l(x; a, b, c)$$

with a convex combination parameter $\lambda \in [0, 1]$. The Gaussian is given by

$$g(x; a, b, c) = c \exp\left(-\ln(2)\frac{(x-b)^2}{a^2}\right).$$

The half-width a fulfills the relation $a = \sqrt{2\ln(2)}\sigma$ where σ is the variance of the Gauss curve. Example line shapes are illustrated for $\lambda = 0$ (pure Lorentzian), $\lambda = 0.5$ and $\lambda = 1$ (pure Gaussian) on the right-hand side of Figure 3. Thus a spectrum containing N peaks can be modeled as

$$h(x, \underbrace{\{p_1, \ldots, p_N\}}_{p}) = \sum_{i=1}^{N} h(x, p_i) \tag{1}$$

with $p_i = (a_i, b_i, c_i, \lambda_i)$ being the parameter vector of the ith peak.

Future work will consist of scalar couplings and their effect on NMR spectra, resulting in a need for more general or different model functions. This will entail an even further reduction of the amount of optimization parameters.

How to determine these parameters is explained in the following subsection.

2.2. Computation of the model parameters by optimization

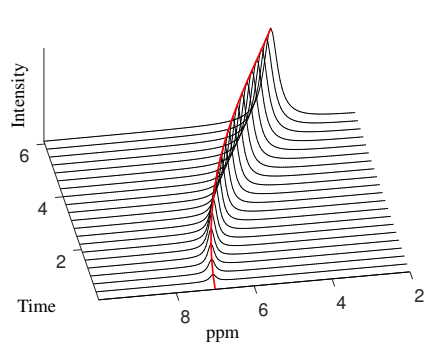
The set of optimal parameter vectors $p^* = \{p_1^*, \ldots, p_N^*\}$ of the model (1) for a given spectrum d can be obtained by numerical minimization of the lack-of-fit functional

$$r_1(p) = \|h(x, p) - d\|_2^2 = \sum_{l=1}^{n} \left(h(x_l, \{p_1, \ldots, p_N\}) - d_l\right)^2 . \tag{2}$$

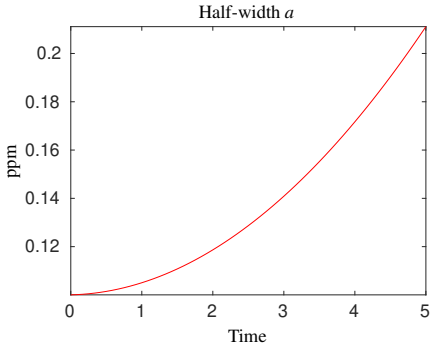
The optimization can be combined with further constraints in order to increase the stability of the algorithm. We use the constraint functions

$$r_2(p) = \sum_{i=1}^{N} (\min(0, a_i))^2 \qquad \text{(nonnegativity of } a\text{)},$$

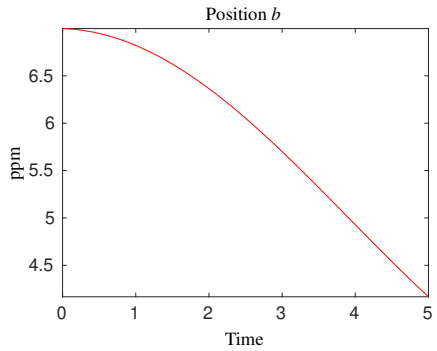
$$r_3(p) = \sum_{i=1}^{N} (\min(0, c_i))^2 \qquad \text{(nonnegativity of } c\text{)}.$$



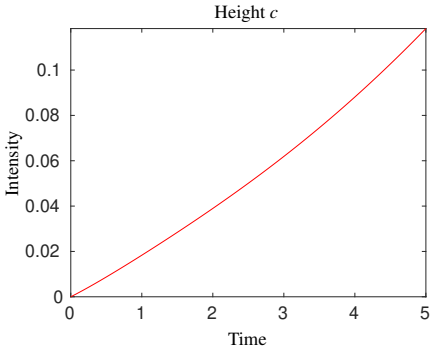
(a) A single Lorentzian peak over time.



(b) Parameter function $a(t)$ for the half-width.



(c) Parameter function $b(t)$ for the position of the peak center.



(d) Parameter function $c(t)$ for the height of the peak.

Figure 2: Modeling of a single peak (by a Lorentzian line shape) over time results in three parameter functions over time.

Hence, the final objective function $f : \mathbb{R}^{4N} \to \mathbb{R}$ is a weighted sum of the residual r_1 and the additional constraints r_2, r_3. It reads

$$f(p; d) = \sum_{k=1}^{3} \gamma_k r_k(p) \tag{3}$$

with weighting factors $\gamma_k \in \mathbb{R}, k = 1, 2, 3$. In our computations the resulting constrained minimization problem is solved by a nonlinear least-squares algorithm, namely the trust-region-reflective least squares method [3, 4] as implemented in Matlab.

2.3. Modeling of time series of NMR spectra

Let $D \in \mathbb{R}^{n \times m}$ be the NMR data matrix where n is the number of data points and m the number of time steps. Thus the NMR spectra are stored column-wise in D. We assume that each spectrum consists of N peaks and that these can be modeled by Gauss-Lorentz functions. A simple way to handle such cases is to apply the minimization of f by (3) to each column of D. It is reasonable to assume that the parameters of two consecutive columns of D differ only slightly. Thus a temporal coupling can be established, e.g., by using the optimized parameters of one time layer as starting parameters for the minimization on the next time layer.

The simulated data set $D_0 \in \mathbb{R}^{500 \times 61}$ of 61 spectra with 500 data points each is used to illustrate a key problem of this approach. The left plot of Figure 4 shows a column-wise plot of D_0 with $N = 3$ intersecting peaks. The results of the proposed consecutive minimization are shown in the center and right plot. The graphs show a misallocation of the determined centers (middle) and intensities (right) of the intersecting and overlapping peaks. Therefore, a correct

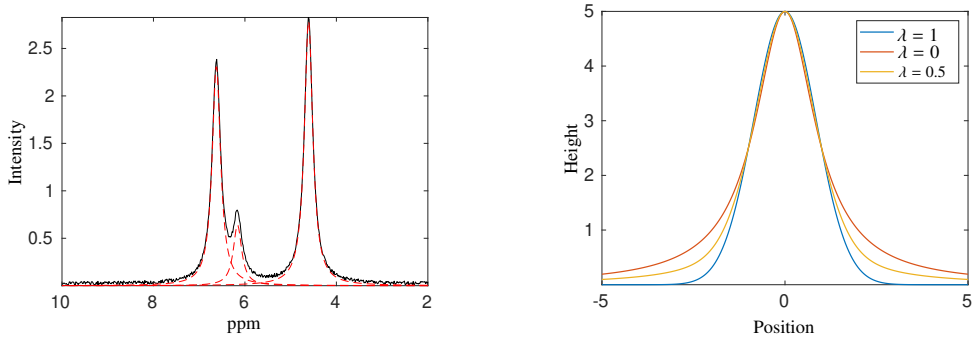


Figure 3: Left: An NMR spectrum (black) with three peaks is modeled with three Gauss-Lorentz curves (red). Right: Comparison of Gauss (blue), Lorentz (red), and a convex combination (yellow) of these curves with the parameters $a = 1, b = 0, c = 5$ and $\lambda \in \{0, 0.5, 1\}$.

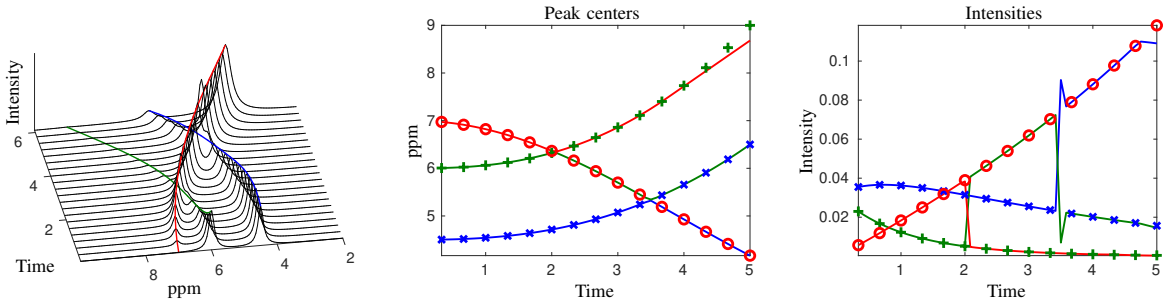


Figure 4: Step-wise fitting of a mechanistic model for each time step. Left: Simulated data set of an NMR time series. Middle: The colored points represent the simulated peak centers. Three curves interpolate the numerically reconstructed peak center values as computed for a subset of time layers. The lines interpolate the peak center values in the ordering as found by the algorithm. The resulting interpolation parameter function is continuous but not smooth. This can lead to misallocations. Right: Analogue representation for the intensities. Here the interpolating curves can even jump from one branch of the solution to another.

traceability of the peaks is not given. For instance the red curve in the centered plot is continuous, but not smooth, and the red curve in the right plot is even not continuous, but jumps from one solution branch to another branch. In order to force smooth parameter functions a spline-based approach is introduced in the next section.

3. Spline-based modeling for NMR time series

Instead of determining a large number of model parameters for each spectrum in a non-coupled way (namely the number of parameters for one spectrum times the number of spectra), we prefer to compute low-dimensional spline approximations for the parameter curves. This approach has two advantages: First, the algorithm automatically finds only smooth parameter functions, and misallocations as shown in Fig. 4 can be avoided. Second, the model benefits from a considerable dimensionality reduction since only the few coefficients of the splines serve to represent the high-dimensional NMR spectra series. Figure 2 shows typical functions $p_i(t)$ for the half-width, position and height over time for a single-peak time series. Their smoothness qualifies them for a spline approximation.

Originally, splines are known as a term for *elastic ruler* and were used for ship constructions where elastic boards were nailed to the frames of a ship. The resulting surface was as smooth as possible [5, 6]. Here splines are used to approximate the parameter functions. Splines are piecewise polynomials that pass through a number of predefined so-called knots or nodes. The left-hand side of Figure 5 illustrates a spline interpolation (blue) for 6 arbitrary nodes (blue circles). For comparison, a linear interpolation is drawn in red. We summarize that the suggested technique

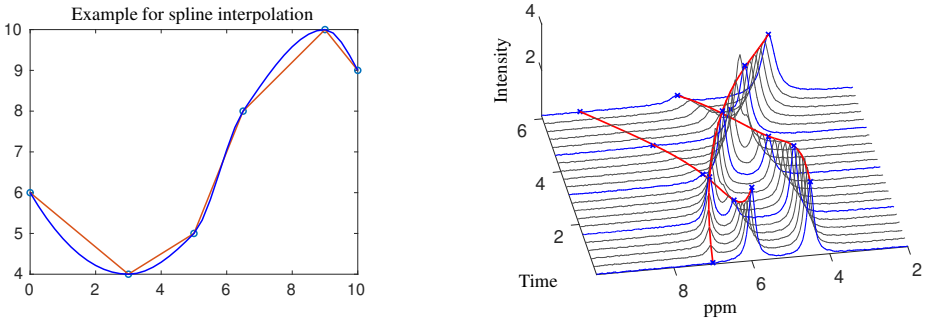


Figure 5: Left: Example for a cubic spline interpolation (blue) of the blue knots. A linear interpolation is shown in red. Right: Spline-based modeling over time. The blue spectra denote the $q = 5$ base spectra with the position and height of the intersecting peaks drawn by red lines.

requires that the changes of parameters are continuous and smooth over time. Then these functions can be expressed with few parameters, namely the coefficients defining the spline.

3.1. Spline-based modeling and optimization

In contrast to the procedure of Section 2.3 the m time layers where each spectrum has not more than N peaks are now analyzed simultaneously rather than step by step. Therefore all parameters are saved in a matrix-like structure. More precisely, it is a tensor (although a detailed introduction is not relevant for the paper)

$$P = \begin{Bmatrix} p_{1,1} & \cdots & p_{1,m} \\ \vdots & & \vdots \\ p_{N,1} & \cdots & p_{N,m} \end{Bmatrix} \quad \text{with} \quad p_{i,j} = (a_{i,j}, b_{i,j}, c_{i,j}, \lambda_{i,j}) \tag{4}$$

with $4Nm$ degrees of freedom. This data structure supports typical operations, e.g., the access of all parameters of the jth time layer by $P_{:,j}$. A generalized objective function $\tilde{f} : \mathbb{R}^{4N\times m} \to \mathbb{R}$ is defined, which is based on Eq. (3) and reads

$$\widetilde{f}(P; D) = \sum_{j=1}^{m} f(P_{:,j}; D_{:,j}) \,. \tag{5}$$

The sum expresses the column-wise application of f to P and D with a summation of the residuals. In the present form a minimization of (5) is not effective since a simultaneous optimization of the $4Nm$ parameters is too costly. Further, the time layers are still not coupled. Both problems can be solved by using splines to model the parameters P in the time direction. To this end we select q values on the time grid of measurements and collect the indices in a set I. Further, P_I is the corresponding column-wise restriction of P that contains only the parameters for the selected time layers. To give an example, we consider the data matrix $D_0 \in \mathbb{R}^{500\times 61}$ in Subsection 2.3 for which the full time grid is $\{t_1, t_2, \ldots, t_{61}\}$. A reasonable, since equidistant example subset is $\{t_1, t_{16}, t_{31}, t_{46}, t_{61}\}$ with $I = \{1, 16, 31, 46, 61\}$, see Section 3.2.3 on the determination of the node set I. An optimization of the parameters only with respect to the time layers defined by I reduces the number of degrees of freedom from $(4N) \cdot 61$ to $(4N) \cdot 5$.

In order to exploit all available information, the objective function should still be evaluated for all time layers. For this purpose P is approximated based on P_I. For the sake of simplicity just one parameter of the first modeled peak is considered, e.g., the half-width $a_{1,j},\ j = 1, \ldots, m$. Let $s_1^a(t)$ be the corresponding continuous spline function. It is defined piecewise in the ranges determined by I with smooth transitions. We use cubic Hermite spline functions, which result in less overshoot behaviour than the commonly used natural cubic splines. For the optimal parameters $a^*_{1,j}$ the spline fulfills the equations $s_1^a(t_j) = a^*_{1,j}$ for $j \in I$ and the approximations $s_1^a(t_j) \approx a^*_{1,j}$ for $j \notin I$. The spline is fully determined by the selection of time layers I and the parameters $a_{1,j},\ j \in I$, in the transition points. Hence, no further unknowns are added to the optimization process.

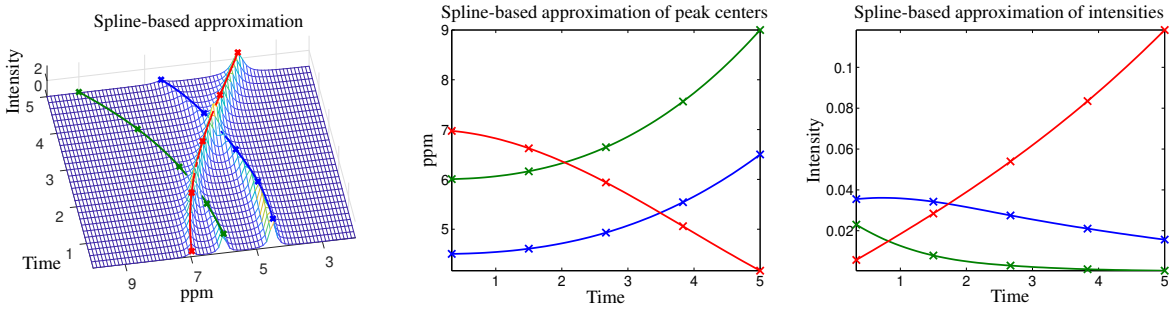


Figure 6: Spline-based approximations with $q = 5$ base points for the data set $D_0 \in \mathbb{R}^{500\times 61}$. The left graph shows an overview of the $N = 3$ intersecting peaks. The condensed splines are depicted in red, green and blue. The peak centers $s_b(t)$ (middle) and intensities $s_c(t)$ (right) are shown for the three peaks. The stars indicate the base points as given by the set I.

Next the procedure is extended to all parameters. For the ith modeled peak the vector-valued function $s_i(t) = (s_i^a(t), s_i^b(t), s_i^c(t), s_i^\lambda(t))$ contains a spline for each of the parameters half-width, center, height and convexity. Then an approximation of the parameters P in (4) is determined by

$$P_s = \begin{Bmatrix} s_1(t_1) & \cdots & s_1(t_m) \\ \vdots & & \vdots \end{Bmatrix}. \tag{6}$$

Based on (5) an objective function $F(P_I; D) := \widetilde{f}(P_s; D)$ is defined, that solely depends on the parameters of the selected time layers but still evaluates the model for the complete time grid.

In summary, the spline-based approximation is a global approach for the decomposition problem of NMR time series. There are N peaks with 4 parameters (half-width, center, height, convexity) that are evaluated on q node points in the time direction. Therefore, the dimensions of the nonlinear optimization problem can be decreased to $4Nq$ where $q \ll m$. Additionally, when dealing with intersecting peaks, spline-based modeling of the parameter functions enforces smoothness and avoids unwanted jumps or bends. Whereas Figure 4 (middle and right) shows that a time layer oriented step-by-step approach can lead to jumps in the parameter functions for the position and intensity, the middle and left graphs in Figure 6 illustrate that all functions are smooth with the spline-based approximation. Misallocations can be avoided since the optimization forces the solution towards smooth approximating functions whereas continuous, but non-smooth solutions are hardly representable by splines.

3.2. Implementation of the spline-based NMR time series modeling

The algorithm starts with the real part of the phase and baseline corrected NMR data $D \in \mathbb{R}^{n\times m}$ where n is the number of data points and m the number of time steps. An automated and thus homogeneous correction of the phase and baseline for spectral data sets is particularly important as a starting point. To this end, the data is preprocessed by the SINC software [16], which provides an automated and simultaneous correction. The following algorithm is applied:

Algorithm 1 Spline-based NMR modeling algorithm

Require: phase and baseline corrected NMR data $D \in \mathbb{R}^{n \times m}$

1. An automatic peak detection determines the maximal number of peaks N per spectrum in D, see Sec. 3.2.1.
2. One spectrum at time step j^* with N peaks is chosen as start spectrum and an initial minimization using the objective function (3) is carried out. This results in the optimal parameters $P^*_{:,j^*}$.
3. Initially, a constant parameter function is assumed. For this, the optimal parameters $P^*_{:,j^*}$ at time step j^* are used as initial values for all time steps, see Sec. 3.2.2. The optimization for the spline functions $s_1(t_I), \ldots, s_N(t_I)$ at a fixed node set I, see Sec. 3.2.3, is then conducted by minimizing the objective function $F(P_I; D) := \widetilde{f}(P_s; D)$.

3.2.1. Automatic peak detection

The negative minima of the second derivative of a spectrum approximate the peak centers as reviewed by Antonov and Neltcheva [2] for UV-Vis spectroscopy. Here, the second derivative of the spectra is calculated by a Savitzky-Golay filter [15, 14]. Since the Savitzky-Golay smoothing manages to remove a certain level of noise while also preserving small peaks, the peak detection is quite robust to noisy input data. If the peak detection algorithm nevertheless fails to assign the peaks correctly, then mistakenly found peaks are less problematic than missed ones, since their heights can be reduced to 0 during the optimization phase. To automatically detect the peak centers in the presence of noise and to prevent the algorithm from overlooking existing peaks, three parameters can be adjusted [9]: threshold for the negative value in the second derivative, threshold for the minimum peak height and a window width.

3.2.2. Initialization

The initialization of the spline functions can be conducted in various ways. One simple first try may be a constant function over time. With the second step of the algorithm (see 3.2) a first initial optimization gives the optimal parameter vector $P^*_{:,j^*}$ at time step j^*. These are assumed as constant values over time for the initial spline functions. Other initialization is possible. Successive building of the splines may be a way to stabilize the algorithm and is of interest for future work.

3.2.3. Determination of the node set I

The choice of the node set I has a strong effect on the optimization result. The nodes do not need to be equidistant. An area with rapid changes in location or intensity of peaks should have more nodes than areas with almost linear behaviour. The node set is manually stated and fixed for the optimization. Generally, the more nodes are selected, the more stable the optimization is, but the less use is made of the advantages of the splines.

3.2.4. Data with phase and baseline errors

To illustrate the robustness of the algorithm, we first add quadratic-polynomial-shaped baselines with different amplitudes to the simulated data matrix D_0 from Section 2.3. In a second series of tests, the data D_0 is phased by three different angles. The exact parameters of all time layers are known for the given model data set. For this reason, the average distances between the correct and modeled values of all parameters and all time layers can be calculated and allow us to determine how well the algorithm behaves. In all subsequent runs, the node set is set to $I = [t_1, t_{13}, t_{25}, t_{37}, t_{49}, t_{61}]$ and $j^* = 13$. The results are displayed in Table 1.

For a phasing with the phase angles $\pi/9$ and $\pi/6$ the algorithm has difficulties detecting the second peak (drawn in green), which explains the relatively large errors of the peak center variables $b_{i,j}$. Table 2 lists these errors without the parameters of the second peak.

Figure 7 shows for a certain time layer the three detected peaks by broken red lines. In all test runs exactly three peaks are found. Before considering the results presented in the two tables, we note that the average errors of $a_{i,j}, b_{i,j}$ and $c_{i,j}$ lie within the intervals $[0, 0.25], [0, 8]$ and $[0, 1]$, respectively. Since adding a baseline profile does not significantly change the peak center values, the average error for the $b_{i,j}$ remains relatively small. The heights $c_{i,j}$

Description	global rel. error	∅ error of $b_{i,j}$	∅ error of $a_{i,j}$	∅ error of $c_{i,j}$
D_0 non-modified	0.040	0.004	0.003	0.006
baseline $\leq 1\% \cdot \max(D_0)$	0.063	0.023	0.013	0.009
baseline $\leq 5\% \cdot \max(D_0)$	0.179	0.036	0.030	0.020
baseline $\leq 10\% \cdot \max(D_0)$	0.300	0.019	0.047	0.031
phasing by $\pi/18$	0.110	0.046	0.034	0.013
phasing by $\pi/9$	0.182	0.505	0.039	0.055
phasing by $\pi/6$	0.273	0.524	0.047	0.084

Table 1: Relative and averaged parameter errors after baseline addition and phasing for the model data set D_0 from Section 2.3.

Description	global rel. error	∅ error of $b_{i,j}$	∅ error of $a_{i,j}$	∅ error of $c_{i,j}$
phasing by $\pi/9$	0.182	0.025	0.030	0.021
phasing by $\pi/6$	0.273	0.037	0.038	0.038

Table 2: Relative and averaged parameter errors after phasing without taking into account the parameters of the second peak.

increase with the baseline offset, but not with the same magnitude, and also the half-widths $a_{i,j}$ are increasing. These are expected effects.

The phasing of D_0 makes the peaks narrower and they lose their characteristic shape, which makes it more difficult to approximate them by a single Gauss-Lorentz curve. While the added baseline preserves the general shape of the keys, the phasing leads to degenerate shapes and areas with negative values. The algorithm loses the track of the green peak as seen on the right side of Figure 7 for a time layer after the peak crossing. Additionally, the uncharacteristic peak shape forces the algorithm to use two Gauss-Lorentz profiles for the approximation. To summarize, if the center values are found more or less correctly, then the height values are more resilient than the half-widths. Peaks tend to be broader than expected if either the baseline or phase correction fails.

4. Numerical results for experimental NMR spectra

Next the suggested spline-based modeling approach is tested for two experimental NMR data sets that contain overlapping peaks. First the data sets are introduced. Then the three steps of the algorithm and the results are presented.

Data set 1. Batch distillation of methanol (CH_3OH) and diethylamine ($C_4H_{11}N$), see Fig. 8. Details on the experiment can be found in Friebel et al. [8]. Thus, only a brief description is given here. For the isobaric batch distillation

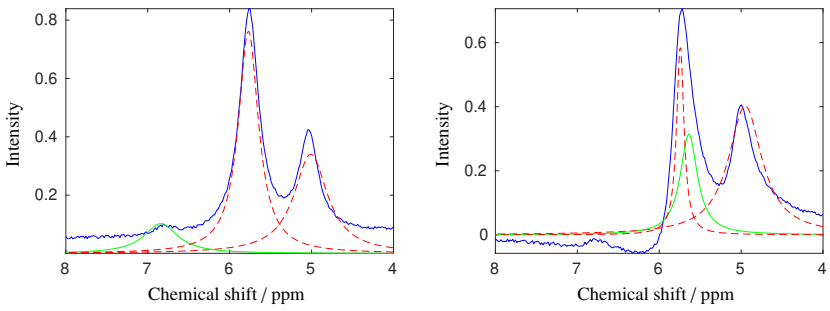


Figure 7: A baseline distorted spectrum and a phased spectrum. Left: Data in the time layer t_{36} with added baseline with a 10% amplitude. Right: Phased data with the angle $\pi/6$. The second (mis-assigned) peak is drawn in green. The optimal Gauss-Lorentz approximations are drawn as broken red lines. The phasing leads to partially negative signals so that the second peak in the negative cannot be modeled correctly by a strictly positive Gauss-Lorentz profile.

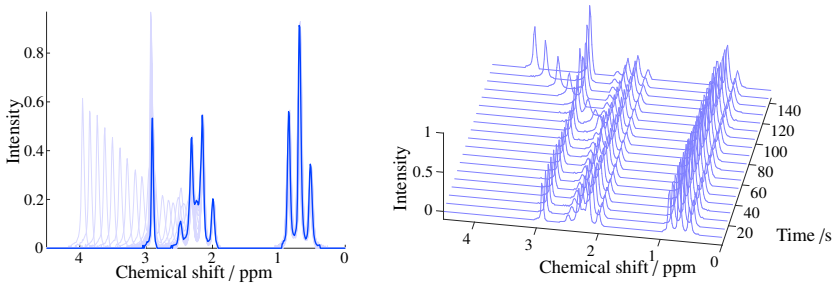


Figure 8: Data set 1. Batch distillation of methanol and diethylamine.

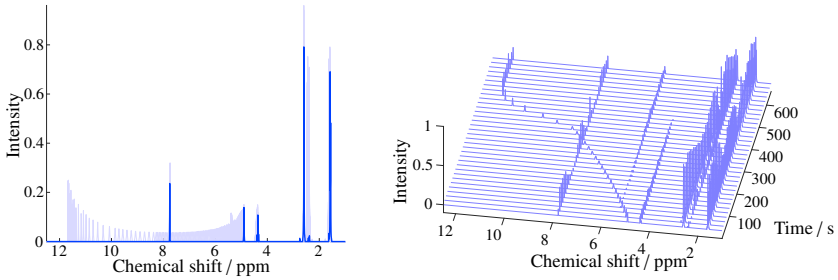


Figure 9: Data set 2. Reaction of acetic anhydride and 2-propanol in the presence of H_2SO_4 to isopropylacetate and acetic acid.

of methanol (≥99.8 mass-%, Carl Roth) and diethylamine (≥99 mass-%, Acros Organics) a glass batch distillation equipment was used. The composition of the liquid phase was analyzed online by a medium field NMR spectrometer (Spinsolve Carbon, Magritek) with a field strength of 1T corresponding to a proton Lamor frequency of 42.5 Mhz, which was connected to the batch distillation still by a sample loop. NMR spectra were recorded in intervals of 1 minute with one scan, an acquisition time of 3.2 s and a pulse angle of 90°. In total, 151 spectra with 16k data points each were recorded. The relevant part of the measurement covers about 6001 data points, thus $D_1 \in \mathbb{R}^{6001x151}$.

Data set 2. Reaction of acetic anhydride and 2-propanol in the presence of H_2SO_4 to isopropylacetate and acetic acid, see Fig. 9. An equimolar mixture of acetic anhydride (≥99 mass-%, Sigma Aldrich) and 2-propanol (≥99.9 mass-%, Merck KGaA) was prepared gravimetrically using a precision balance (Mettler Toledo, AG204) with a specified absolute uncertainty of ±0.0001 g. As inert component, benzene (≥99.8 mass-%, AppliChem) was added to the reacting mixture. The reaction was accelerated by sulphuric acid (95–97 mass-%, J.T. Baker). The reactants were mixed thoroughly, then loaded into an 5 mm NMR tube, and placed in a high-field NMR spectrometer (magnet: Ascend 400, console: Avance 3 HD 400, Bruker) equipped with a 9.4 T vertical superconducting magnet corresponding to a proton Larmor frequency of 400.25 MHz. The temperature of the reacting mixture was about 25 °C. NMR spectra were recorded with one scan, an acquisition time of 6 s and a pulse angle of 30°. A number of 363 spectra was recorded with 65k data points each. The relevant part of the measurement covers about 40k data points, thus $D_2 \in \mathbb{R}^{40000x363}$.

The three steps of the algorithm as described in Sec. 3.2 are next illustrated for data set 1. The first step of the algorithm performs a peak detection for every spectrum in the data set. The left-hand side of Figure 10 displays one spectrum (blue) from data set 1 with its found peak centers (red stars). The middle figure shows the second derivative of this spectrum (blue) with its negative minima (green stars) corresponding to the peak centers in the spectrum. After determining the maximal number of peaks $N = 9$, one base spectrum at time step $j^* = 151$ is chosen.

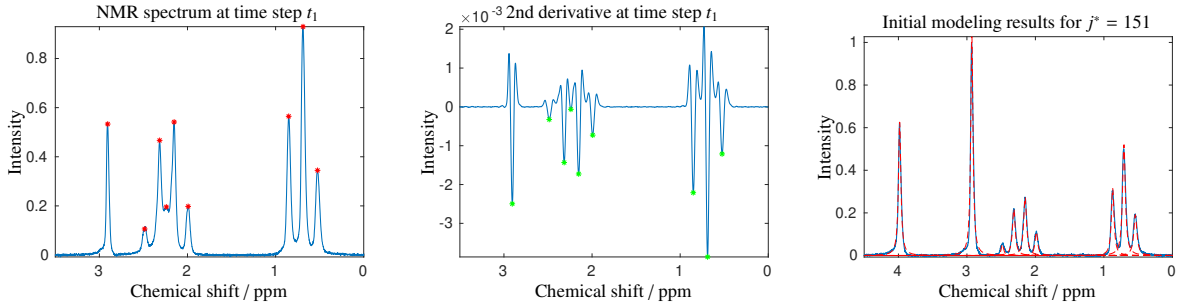


Figure 10: Peak detection for a single spectrum of data set 1. Left: the spectrum at time step t_1. The red stars represent the found peak centers. Middle: the second derivative from the spectrum on the left-hand side. The green stars represent the negative minima. Right: Model at time step t_{j^*}. The red dashed lines represent the fitted curves.

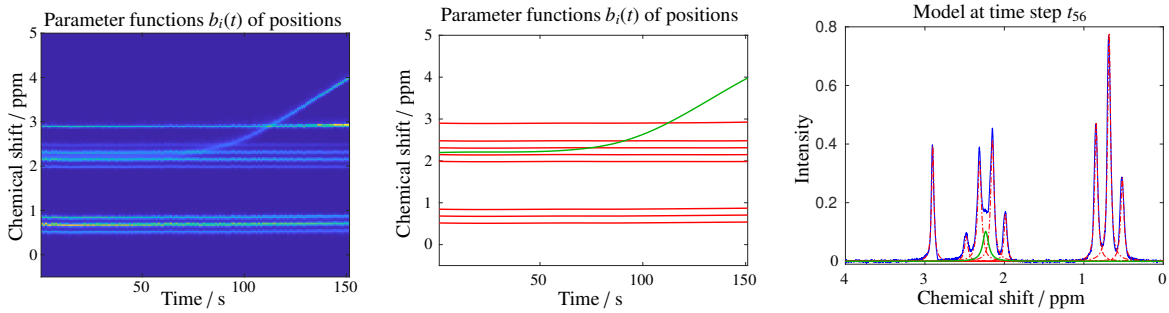


Figure 11: Spline-based approximation of the peak centers for data set 1. Left: Top view on the data set 1. Middle: The parameter functions $b_i(t)$ of the non-moving (constant position) peaks are shown in red whereas the moving peak is depicted in green. Right: Exemplary fit for the spectrum at time step t_{56} for a section of the ppm-axis. The constant peaks are shown in red and the moving peak is depicted in green. For more information see Figure 12.

In the second step this spectrum is modeled by minimizing the objective function (3). The results are shown on the right-hand side of Figure 10. The fitted curves are depicted in red. The found parameters $P^*_{:,j^*}$ are used to initialize the spline functions in the third step of the algorithm.

In the last step, the parameter functions of the peaks are approximated by spline functions. A node set $I = [t_1, t_{35}, t_{65}, t_{85}, t_{102}, t_{121}, t_{151}]$ with $q = 7$ base points is selected. The parameter functions are found by minimizing the objective function with respect to (6). Figure 11 shows a top view of data set 1 (left) and the parameter functions $b_i(t)$ for the positions (middle). As seen in Figure 11 the algorithm can trace the moving NH/OH-peak (green). An exemplary spectrum (blue) at time step t_{56} with its modeled curves (red and green) is shown on the right side. More details of the fitted curves are shown in Figure 12. We finally report the relative error of the fit for this data set

$$e_1 = \frac{\|D_1 - \tilde{D}_1\|_F}{\|D_1\|_F} = 0.1816.$$

Therein $\|\cdot\|_F$ denotes the Frobenius norm, namely the square root of the sums of squares of the matrix elements.

The Frobenius norm is not a perfect objective measurement of the performance of the algorithm. However, the numerical optimization by nonlinear least-squares minimization needs an objective function in a sum-of-squares form. Error measures that work only in certain windows are also possible. However, this would necessitate a manual pre-modeling of the data. This seems to be impractical in general situations.

Applying the algorithm to data set 2 gives comparable results. The spline-based signal tracking can correctly trace the peak centers as shown in Figure 14. Even though small errors occur due to the size of the data set, the complexity of the underlying optimization problem and the multiple crossings, see Figures 15 and 16, the algorithm still achieves

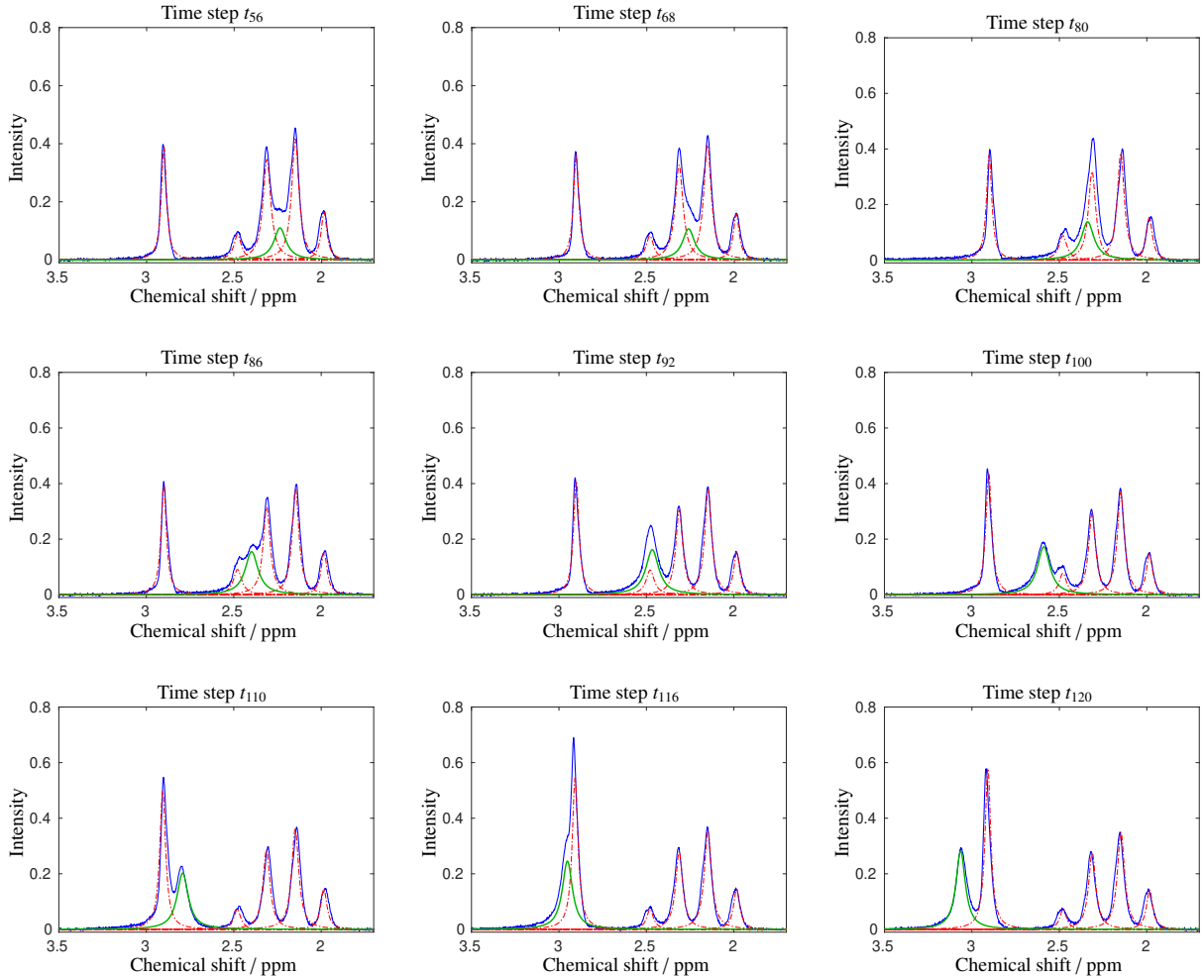


Figure 12: Excerpt from data set 1 at different time steps with a moving peak (green). The recorded NMR spectra are depicted in blue whereas the modeled peaks are shown in red and green (for the moving peak).

a relative error of

$$e_2 = \frac{\|D_2 - \tilde{D}_2\|_F}{\|D_2\|_F} = 0.1367.$$

A node set $I = [t_1, t_{37}, t_{100}, t_{171}, t_{221}, t_{254}, t_{304}, t_{351}, t_{385}, t_{418}, t_{435}, t_{452}, t_{485}, t_{518}, t_{535}, t_{552}, t_{569}, t_{604}, t_{646}, t_{679}]$ with $q = 20$ base points is selected. At time layer $j^* = 221$ the algorithm detects $N = 17$ peaks, as can be seen in Figure 13.

5. Conclusion

Efficiently processing large NMR data sets is of great interest for online-monitoring methods. The spline-based modeling of NMR time series can support an automated chemometric analysis of such data especially for spectra series with overlapping or crossing peaks. This might qualify the method for benchtop low-field NMR spectra with their often stronger overlapping and broader peaks. For such data peak tracking is a more serious problem.

The suggested technique also adds a new aspect concerning dimensionality reduction for high-dimensional NMR data. In general, (hard) models considerably reduce the data dimension and allow to extract the essential, structure-determining information. An additional benefit of spline representations for the model parameters is that they represent

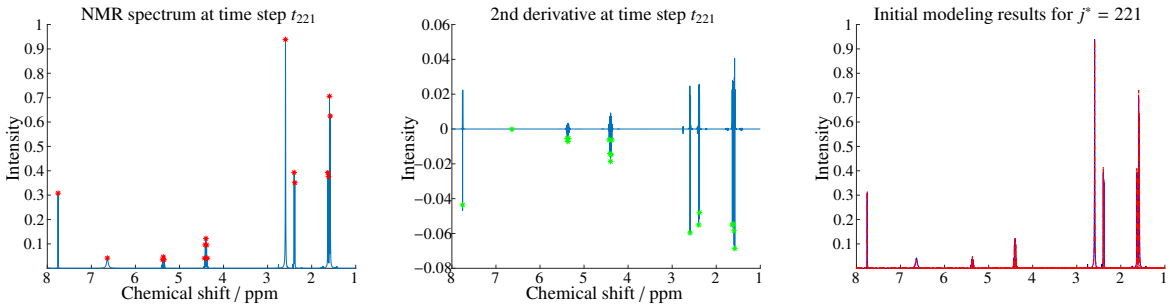


Figure 13: Peak detection for a single spectrum of data set 2. Left: the spectrum at time step t_{221}. The red stars represent the found peak centers. Middle: The second derivative of the spectrum plotted left. The green stars represent the negative minima. Right: Model at time step t_{221}. The red dashed lines represent the fitted curves.

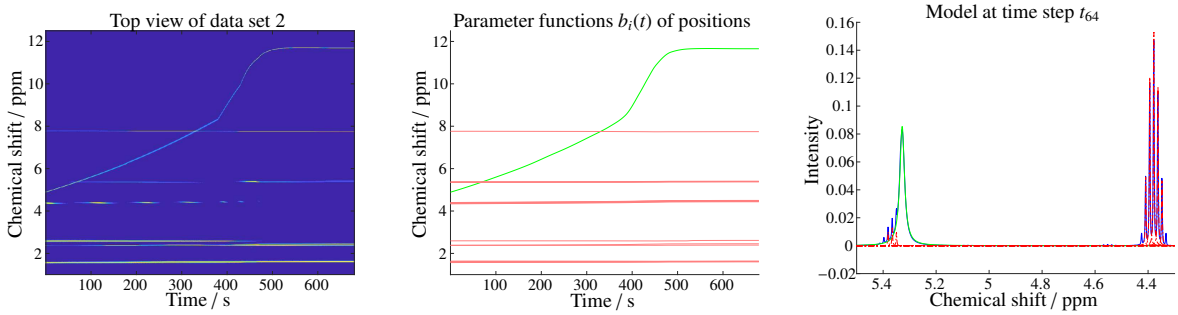


Figure 14: Spline-based approximation of peak centers from data set 2. Left: Top view on the data set 2. Middle: The parameter functions $b_i(t)$ of the non-moving peaks are shown in red whereas the moving peak is depicted in green. Right: Exemplary fit for the spectrum at time step t_{64} for a section of the ppm-axis. Again the constant peaks are shown in red and the moving peak is depicted in green.

an effective coupling of such models for (time or other) series of NMR spectra which makes even a further reduction of the problem dimension possible.

Future work can concern the combination of spline-based modeling with kinetic modeling and the usage of alternative function bases as wavelets for the model building.

References

[1] F. Alsmeyer, H.-J. Koß, and W. Marquardt. Indirect spectral hard modeling for the analysis of reactive and interacting mixtures. *Appl. Spectrosc.*, 58(8):975–985, 2004.

[2] L. Antonov and D. Nedeltcheva. Resolution of overlapping UV–Vis absorption bands and quantitative analysis. *Chem. Soc. Rev.*, 29(3):217–227, 2000.

[3] T. Coleman and Y. Li. On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Program.*, 67(1-3):189–224, 1994.

[4] T. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.*, 6(2):418–445, 1996.

[5] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Springer-Verlag, 2006.

[6] C. De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

[7] A. de Juan, J. Jaumot, and R. Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods*, 6(14):4964–4976, 2014.

[8] A. Friebel, E. von Harbou, K. Münnemann, and H. Hasse. Online process monitoring of a batch distillation by medium field NMR spectroscopy. *Chem.l Eng. Sci.: X*, 219:115561, 2020.

[9] F. Holler, D. Burns, and J. Callis. Direct use of second derivatives in curve-fitting procedures. *Appl. Spectrosc.*, 43(5):877–882, 1989.

[10] J. Keeler. *Understanding NMR spectroscopy*. John Wiley & Sons, second edition, 2011.

[11] E. Kriesten, F. Alsmeyer, A. Bardow, and W. Marquardt. Fully automated indirect hard modeling of mixture spectra. *Chemom. Intell. Lab. Syst.*, 91(2):181–193, 2008.

[12] E. Kriesten, D. Mayer, F. Alsmeyer, C. Minnich, L. Greiner, and W. Marquardt. Identification of unknown pure component spectra by indirect hard modeling. *Chemometr. Intell. Lab. Syst.*, 93(2):108–119, 2008.

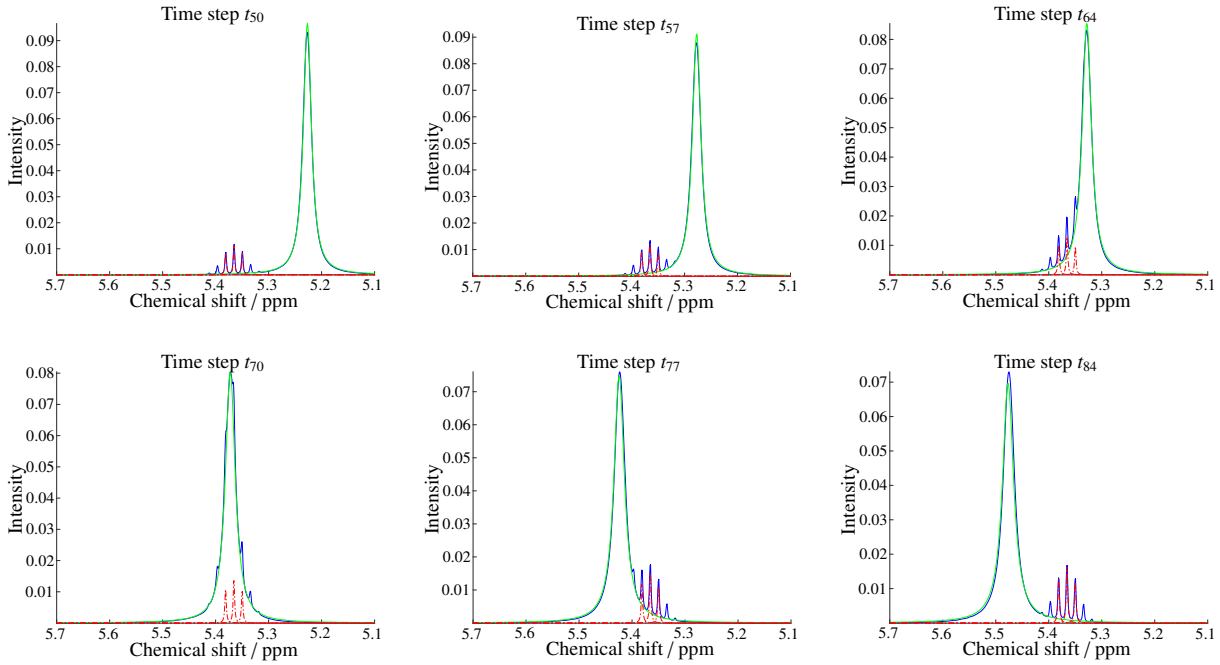


Figure 15: Excerpt from data set 2 depicting the first crossing. The moving peak is shown in green, the non-moving peaks are depicted in red and the measured spectrum is drawn in blue.

[13] M. Maeder and Y.-M. Neuhold. *Practical data analysis in chemistry*, volume 26. Elsevier, 2007.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[15] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[16] M. Sawall, E. von Harbou, A. Moog, R. Behrens, H. Schröder, J. Simoneau, E. Steimers, and K. Neymeyr. Multi-objective optimization for an automated and simultaneous phase and baseline correction of NMR spectral data. *J. Magn. Reson.*, 289:132–141, 2018.

[17] N. Zientek, K. Meyer, S. Kern, and M. Maiwald. Quantitative online NMR spectroscopy in a nutshell. *Chem. Ing. Tech.*, 88(6):698–709, 2016.

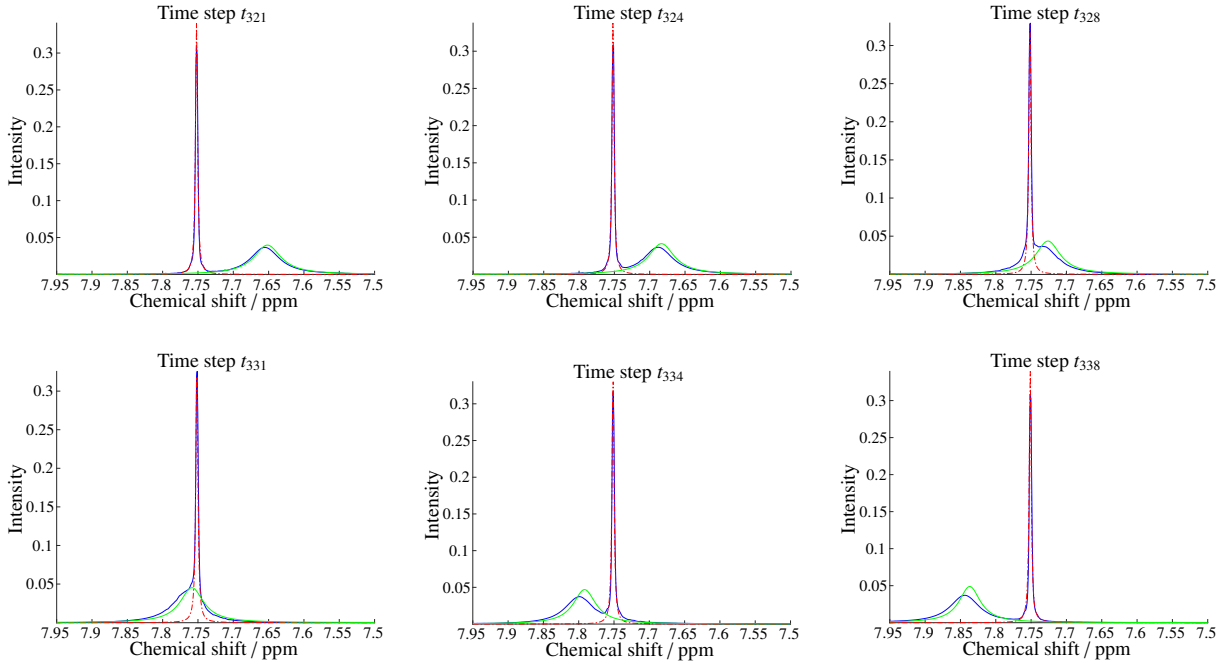


Figure 16: Excerpt from data set 2 depicting the second crossing. The moving peak is shown in green, the non-moving peak is depicted in red and and the measured spectrum is drawn in blue.