Available online at www.sciencedirect.com
ScienceDirect
Journal of Magnetic Resonance 187 (2007) 288–292
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Robust baseline correction algorithm for signal dense NMR spectra

David Chang [a,b,*], Cory D. Banack [a], Sirish L. Shah [b]

[a] Chenomx Inc., Edmonton, AB, Canada T5K 2J1
[b] Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, Canada T6G 2G6

Received 6 February 2007; revised 7 May 2007
Available online 16 May 2007

Abstract

This paper outlines a fully automated algorithm for baseline correction. Based on our experience with NMR spectra of complex mixtures, this algorithm is designed to automatically differentiate signal points from baseline points. The algorithm's strength is its ability to accurately determine baseline points in very dense spectra, without destroying the line shapes of prominent peaks. The algorithm described is implemented in Chenomx NMR Suite 4.6. It is demonstrated here using two separate spectra acquired on two different NMR spectrometers.

Keywords: Baseline correction; 1D NMR spectroscopy; Signal processing; Complex mixtures; Metabolomics

1. Background

Experimental nuclear magnetic resonance (NMR) spectra tend to contain baseline distortions artifacts which can be caused by a variety of different sources, including instrument drifts and unwanted macro molecule signals. Metabolomics applications of NMR spectra often require the identification and quantitation of metabolites found in complex mixtures, since these mixtures can give a snapshot of the state of an organism. It is important to have a flat baseline in order to accurately quantify, hence the need for a good baseline correction algorithm. Systematic baseline distortions also add unwanted correlations in spectral binning data when building correlation models.

Early development on baseline correction algorithms includes work described by Pearson [1], in which baseline correction can be broken down into three steps. The first step is to determine signal and baseline noise in the spectrum. The second is to use that information to build a model of the baseline. This model can be represented using interpolated line segments, or cubic splines if a smoother line is desired. Finally the third, and somewhat trivial step, is to "correct" the signal by subtracting the baseline model from the original signal. Further developments by Zolnai [2], Heuer and Haeberlen [3], Guntert and Wuthrich [4], Bartels [5] all follow this standard pattern and have made significant contributions to each step.

While the problem of baseline correction in the realm of NMR signals is not new and there are some good solutions already available, in our experience the available methods work best on NMR spectra that do not have a very high signal density [1,4,5]. Many existing algorithms tend to be overly aggressive, often destroying the line shapes of prominent peaks in spectra with a wide dynamic range of peak shapes and sizes. The application of many of these methods to metabolomics data is therefore problematic, since NMR spectra of complex biofluids often result in very signal dense spectra. In this paper we propose a new algorithm for baseline correction which addresses this problem in a way that does not destroy the line shapes of prominent peaks. Our algorithm is designed for more densely populated spectra, but retains good performance in sparsely populated spectra as well. The algorithm was developed

* Corresponding author. Address: Department of Chemical and Materials Engineering, University of Alberta, 536 CME Building, Edmonton, AB, Canada T6G 2G6.
E-mail address: david.chang@ualberta.ca (D. Chang).

1090-7807/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2007.05.008

based on our combined knowledge of both NMR signals and of the baseline distortions that are common in the realm of complex mixtures. This paper will focus on step one of the general three-step process: A systematic application of heuristic rules which can accurately determine the baseline points in a 1D NMR spectrum.

2. Approach

This section contains a detailed outline of our baseline correction algorithm, as implemented in Chenomx NMR Suite 4.6. The goal of the algorithm is to differentiate the regions of the spectrum (S) that are considered to be baseline noise from those that are considered to be signal. This determination is then stored in a boolean vector known as a signal map (SM). SM has the same dimensions as S; each element contains information about whether the point corresponding in S is signal (true) or baseline noise (false).

The first and most important step in the algorithm is the high pass signal identification step. The objective here is to conservatively identify regions of the spectrum that are signal by looking at a modified version of S wherein all low frequency curves and rolls have been removed. Once the signal regions are identified, everything else can be considered baseline points. In order to accurately determine what is signal, the algorithm first attempts to calculate the standard deviation of the noise in S. This is a common step in other baseline correction algorithms [1,4,5]. However, the typical method for determining the standard deviation of noise by dividing the original spectrum (S) into multiple regions is insufficient. Rolling baselines and areas of high signal make it difficult to estimate the noise in a spectrum.

To overcome this problem, the authors of this algorithm chose to first use a high pass filter on the spectrum. Specifically, a moving average filter was used. This filter is designed to pass 0.5% of the high frequency through in Nyquist frequency. The resulting signal is known as the high pass filtered spectrum (*HPFS*) and contains only the high frequency noise and signal. Fig. 1 shows a spectrum before and after the high pass filter has been applied. From Fig. 1, we can also see the resultant spectrum is highly distorted and not very useful in itself. However, the *HPFS* is still useful to obtain a good estimate of the high frequency baseline noise, because rolls in the baseline have been removed, and signal dense areas have been narrowed.

At this point the *HPFS* is divided into evenly spaced segments, and the standard deviation of each segment is calculated. A percentage (*bfraction*) of the segments with the lowest intensities are assumed to be baseline signal, and the standard deviation of only the points contained within these segments is recalculated (*stdn*). *bfraction* can be adjusted based on the spectrum signal density. A value of between 0.2 and 0.5 was found to work well for complex mixtures.

Once *stdn* has been determined, the next step is to determine what percentage of the entire spectrum is signal. We continue to use the *HPFS* and consider all points with absolute intensities greater than two times the standard deviation of the noise to be signal. The indices of these absolute intensities are now sorted based on the intensities themselves and then used in the signal windowing step.

The signal windowing step returns back to the original spectrum (S). Each signal point found in the previous steps is now used as the center of a signal window. The signal window width used is 0.2% of the total sweep width of the spectrum. Each point inside of the signal window is now also marked as signal in SM. Fig. 2 shows a spectrum overlaid with the baseline points that were found after the signal windowing step.

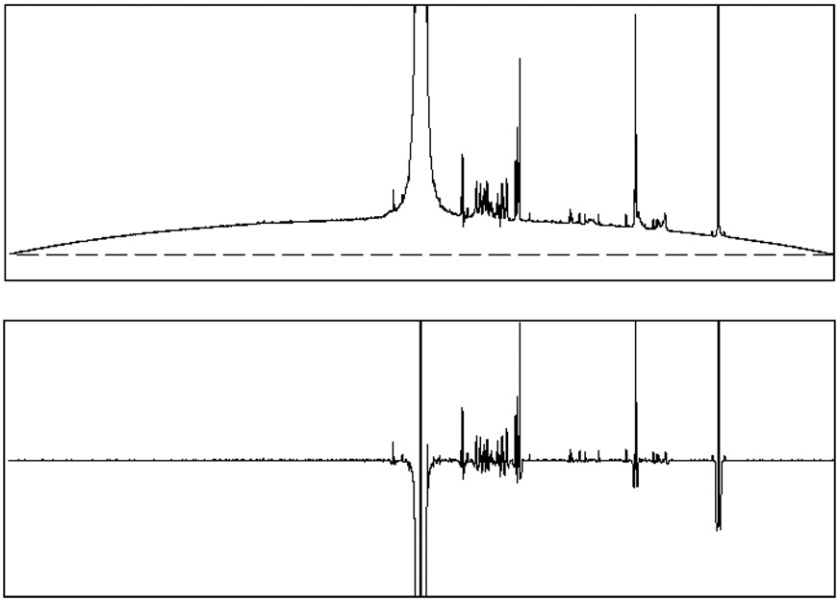

Fig. 1. Top: original spectrum (S) with a noticeable baseline distortion. Bottom: High pass filtered spectrum (*HPFS*) showing the removal of the low frequency distortions. This spectrum is of a human urine sample on a 600 MHz magnet using a presat pulse sequence.

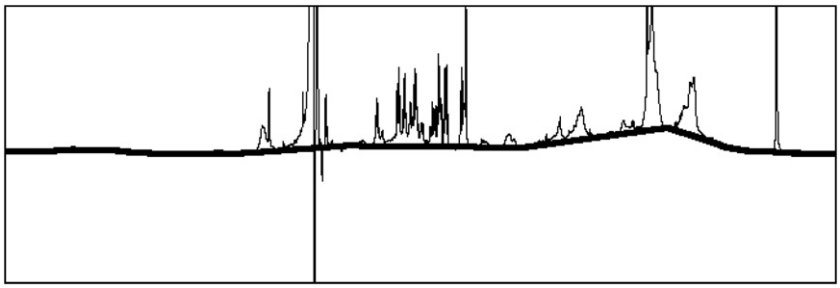

Fig. 2. Baseline points defined after signal windowing step (thick black). This spectrum is of a mouse serum sample run on a 600 MHz magnet using a presat pulse sequence.

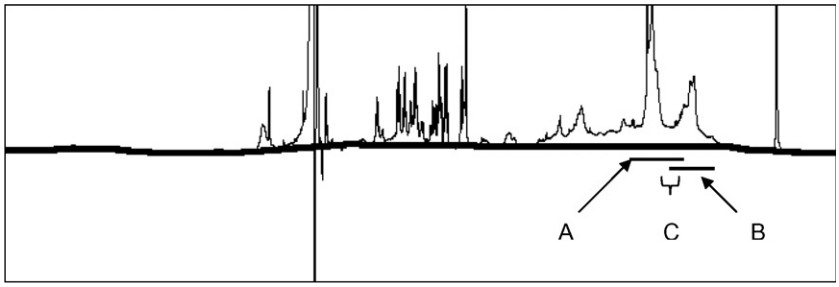

Fig. 3. Baseline points defined after correction for prominent Lorentzian peaks (thick black). (A) Region covered by three picked Lorentzian peaks. (B) Region covered by one picked Lorentzian peak. (C) Region added to *SM* to be considered as signal.

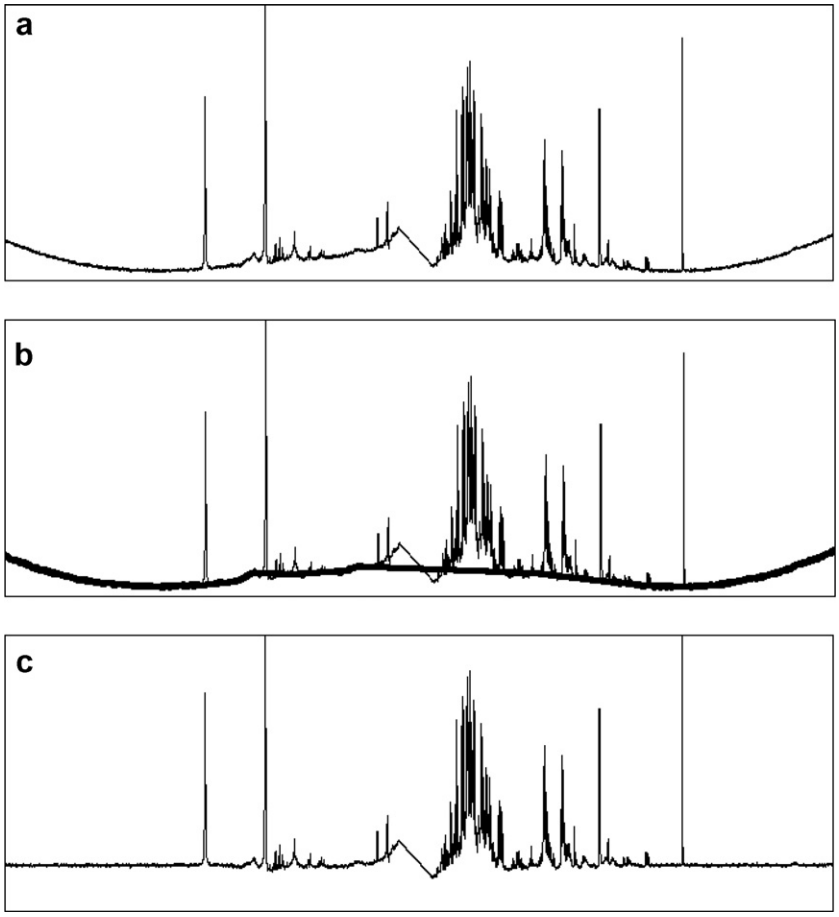

Fig. 4. (a) Original spectrum of acidic plant extract, (b) baseline distortion model, (c) spectrum after baseline correction.

The novelty of this algorithm is the use of a high pass filter. However, it is also a weakness: the high pass filter applied to very tall or large peaks in a spectrum will often misidentify the tails of these peaks as baseline in the signal map. In order to correct for this, a second step is applied. The objective of this step is to determine the most prominent Lorentzian peaks in the spectrum and guarantee that their tails are marked as signal in *SM*. This is because the tails of the most prominent peaks are often filtered out in the high pass filter, and misidentified as baseline signal due to their size relative to the signal window.

The first part of this step is to calculate the average or mean of the entire spectrum (S) in the frequency domain using only the positive values. Then, using an automatic peak picking algorithm, peaks that are twice the mean of the spectrum are located. The widths of the peaks are determined by walking halfway down both sides to find the half width of each peak. The peaks are then mathematically modeled as pure Lorentzian lineshapes and the central portion of *S* that contains 95% of their area is marked as signal in *SM*. Note that this often fixes regions that were erroneously marked as baseline in previous steps.

A 95% cutoff was needed because Lorentzian peaks have infinite tails. The algebraic model for a Lorentzian is:

$$L(x) = \frac{A \cdot w^2}{w^2 + 4 \cdot (x - c)^2} \tag{1}$$

where, for any given position x {Hz}, width w {Hz}, center c {Hz}, and amplitude A, L is the intensity of the Lorentzian at x. Once these additional "signal points" are marked in *SM*, the determination of signal and baseline points is complete. Fig. 3 shows the same spectrum as Fig. 2 with baseline points overlaid (in thick black) after correcting for prominent Lorentzian peaks. You will notice the correction of the misinterpreted baseline points (in thick black). The 95% regions from the picked Lorentzian peaks (A and B) and the region added to the *SM* (C) are also shown in Fig. 3.

We showcase the algorithm's ability to accurately determine the baseline points. To model these points in a spectrum, the original baseline points were used and a simple linearly interpolated line was used to fill in the gaps between the baseline points. A more sophisticated natural cubic spline model is used in Chenomx NMR Suite 4.6.

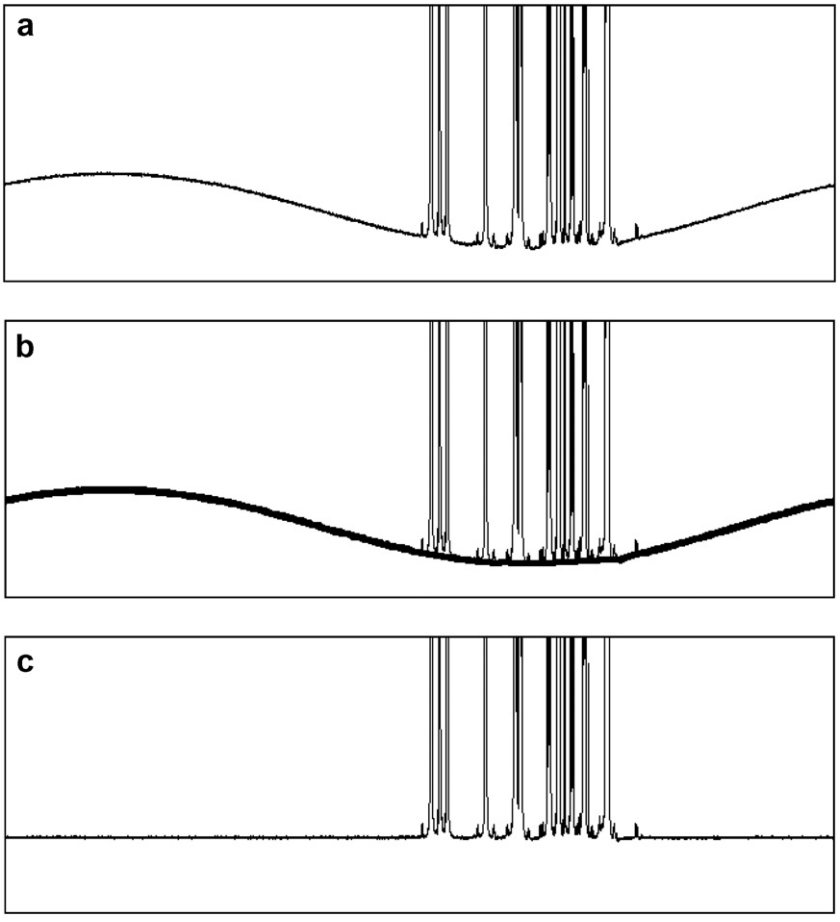

Fig. 5. (a) Original acid spectrum, (b) baseline distortion model, (c) spectrum after baseline correction.

3. Application

The performance of the algorithm is demonstrated in the following two examples, which were acquired from different NMR spectrometers and have different baseline distortion problems. As well, the first of these examples has a high signal density while the second example is sparse in signal density. These spectra were also chosen to show clearly the algorithm's ability to handle gross distortions in the baseline, while at the same time showing that it is able to non-destructively handle the more subtle baselines generated from the most advanced spectrometers today.

For our first example, we applied the algorithm to an NMR spectrum of an acidic plant extract. This sample was run through an NMR flow system on a 400 MHz Varian spectrometer using a vast1d pulse sequence. Some of the older flow systems, which make use of the ssfilter VNMR command, do not always create straight baselines. As can be seen from the black line in Fig. 4a, the original spectrum had a fairly high signal density, as well as an obvious baseline distortion. The baseline points identified by the algorithm are shown, along with linearly interpolated points in between the gaps (in thick black) in Fig. 4b. Finally, the baseline corrected spectrum (i.e. after subtraction) is displayed in Fig. 4c.

Our second example uses an acid sample acquired on an older JEOL Spectrometer, which did not have digital filtering. The lack of digital filtering is probably the cause of this spectrum's pronounced baseline roll. This spectrum was acquired on a 500 MHz magnet using a single-pulse sequence. Fig. 5a shows the original spectrum. Fig. 5b shows the baseline points identified by the algorithm with linearly interpolated points in between the gaps (in thick black). Fig. 5c shows again the high-quality spectrum after the baseline distortion has been removed.

4. Conclusions

The baseline correction algorithm outlined in this paper was designed using characteristic distortions found commonly in spectra from complex mixtures. It follows the established three-stage template and aims at ensuring the accurate determination of baseline points without indentifying too many false positives. The result is a high-quality baseline corrected algorithm that can be used in a variety of metabolomics applications.

Acknowledgments

The authors wish to express their gratitude towards Pascal Mercier and Jack Newton of Chenomx Inc. for their insights.

References

[1] G.A. Pearson, A general baseline-recognition and baseline-flattening algorithm, J. Magn. Reson. 27 (1977) 265–272.
[2] Z. Zolnai, S. Macura, J.L. Markley, Spline method for correcting baseplane distortions in two-dimensional NMR spectra, J. Magn. Reson. 82 (1989) 496–504.
[3] A. Heuer, U. Haeberlen, A new method for suppressing baseline distortions in FT NMR, J. Magn. Reson. 85 (1989) 79–94.
[4] P. Guntert, K. Wuthrich, FLATT—a new procedure for high-quality baseline correction of multidimensional NMR spectra, J. Magn. Reson. 96 (1992) 403–407.
[5] C. Bartels, P. Guntert, K. Wuthrich, IFLAT—a new automatic baseline correction method for multidimensional NMR spectra with strong solvent signals, J. Magn. Reson. 117 (1995) 330–333.