Automatic Concurrent Phase & Baseline Correction in 1D NMR spectra

Ester Maria Vasini [3], Carlos Cobas [1] and Stanislav Sýkora [2]

[1] Mestrelab Research, Xosé Pasín 6, Santiago de Compostela, 15706 Spain; carlos@mestrec.com
[2] Extra Byte, Via Raffaello Sanzio 22/C, Castano Primo (Mi), Italy I-20022; sykora@ebyte.it
[3] University of Bologna, DICAM dept., Via Terracini 28, 40139, Bologna, Italy; estermaria.vasini2@unibo.it

MESTRELAB RESEARCH
NMR Solutions

INTRODUCTION

The algorithmic problem of phase correction (PC), and that of baseline correction (BC), of 1D NMR spectra have been both tackled many times by many authors over the last half a century.

There are many algorithms which emulate manual procedures. Basically, they all optimize (fit) the parameters which describe either the phase (ph0, ph1) or the baseline (according to various parametric models) so as to maximize some quality assessment of the corrected spectrum. Historically, the employed *quality functions* [1, 2] included features such as peak heights, negative peak lobes, symmetry, DISPA patterns [3], selected baseline points, peak ablation, and others. Here we propose a radically different type of 'quality function' Q to be optimized. It is based on the histograms of the spectrum (real and imaginary parts) which turn out to be very sensitive to both phase and baseline distortions. This permits us to:

➤ Carry out the phase and baseline corrections *simultaneously*. In traditional approaches these are always done as separate evaluation steps and carried out sequentially (first phase correction and then baseline correction), even though all NMR spectroscopists know that the two corrections interfere with each other. We think that our approach overcomes this problem.

➤ Carry out the corrections (not just phase) on *both* the in-phase and the out-of-phase parts of a spectrum. So far, the baseline correction has been always done only on the in-phase part. Since this leaves the baseline in the out-of-phase part uncorrected, it has various adverse effects on other evaluation tasks.

➤ Enhance the *objectivity* of the corrections, especially considering that in practice one often faces situations with multiple and/or ill-defined acceptable 'solutions'. Moreover, the manual corrections in current use (especially the phasing procedure) often differ depending of the personal habits of each spectroscopist.
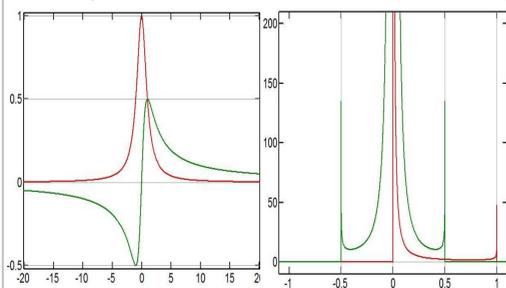
We wish to describe here a novel algorithm developed for the stated purpose and illustrate some of the results, even though this is more a feasibility study and work in progress than a finished product.
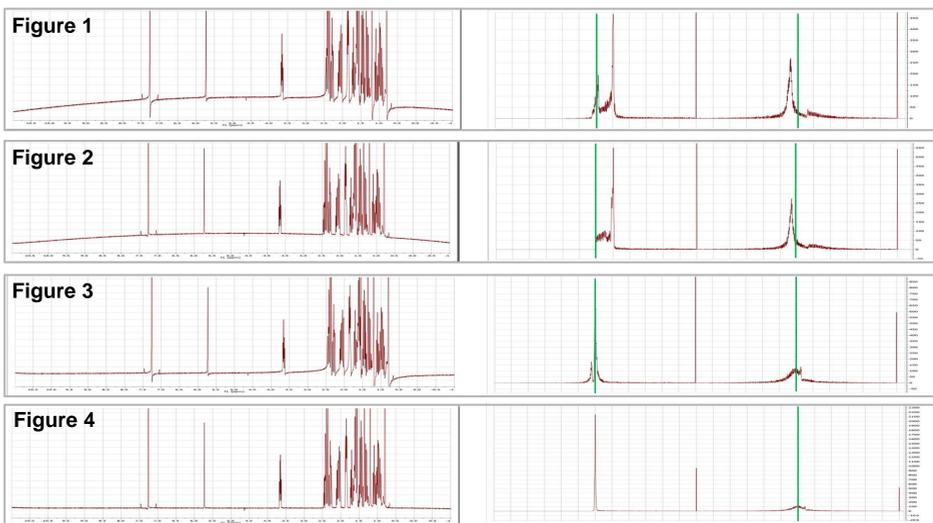
Pilot Idea: Let us use the Histograms

When an NMR spectrum is well conditioned, its absorption mode (real) part contains flat baseline stretches of noise with no peaks. In a histogram of such a data set this implies a sharp peak in the neighborhood of zero (note: histogram $H(\zeta)$ is a diagram of data points counts when binned according to their height; we keep the size of the bins normalized so that there are about 5000 of them per Rms(h) average of the heights of all data points (the horizontal scales of all histogram plots appearing in this poster are normalized so that $\zeta = \pm 1.0$ corresponds to the height $\pm$ Rms(h)).

If we compare the histograms of an experimental spectrum which is badly conditioned (Fig.1) with those of the same spectrum manually properly phased (Fig.2), or baseline corrected (Fig.3), or with both corrections (Fig.4), we see how their real parts tend to progressively peak around zero. In the final, properly conditioned spectrum, the histogram has practically no intensity at negative co-ordinates which exceed the noise, and just a small-intensity 'tail' at positive co-ordinates.

For the histograms of the imaginary part of a properly conditioned spectrum, one notices that it is symmetric with respect to zero and much smaller in maximum amplitude than the histogram of the real part.

It is also interesting to compare this with the histogram of an ideal Lorentzian line as shown on the left (with red traces showing the real parts and green traces the imaginary parts. The left plot shows a normalized Lorentzian peak, while the other one shows the respective theoretical histogram. Note that the sharp asymptotes in the histogram become sharp peaks once an experimental noise is added (convolution with the noise probability function).

Figure 1
Figure 2
Figure 3
Figure 4

Example of histograms:
Left column shows the spectra, while the dual right column shows the corresponding histograms of the real (left side) and imaginary (right side) part of the spectra. The zero position is highlighted by vertical green lines. For more information, see the text on the left.

Modelling the baseline

For the fitting, apart from the two classical phase parameters ph0 (constant) and ph1 (linear) that need to be adjusted, we model the baseline correction by means of a linear combination of a pre-defined number (N) of low-frequency harmonic functions or, alternatively, low-indexed Chebyshev polynomials. In either case, this introduces 2*N fittable coefficients, because the baseline corrections for the real and the imaginary parts are considered totally independent.

The quality function

One then needs to construct a quality function Q(p) of all the 2*N+2 fittable parameters p which, when maximized, forces the histogram to possess the principle characteristics corresponding to well-conditioned spectra. We define Q(p) as *the integral of a properly weighted histogram $H(\zeta)$ of a spectrum corrected using the current parameters* p. Actually, two such integrals are computed, $Q_r(p)$ for the real part of the spectrum and $Q_i(p)$ for the imaginary part, with two different weight functions, $w_r(\zeta)$ and $w_i(\zeta)$, respectively, and then added together, $Q(p) = Q_r(p) + Q_i(p)$. The fact that this simple scheme is based on an integral over the histogram (and therefore over all experimental data points) is very important because it represents, as an extra bonus, a concurrent noise-suppressing filter.
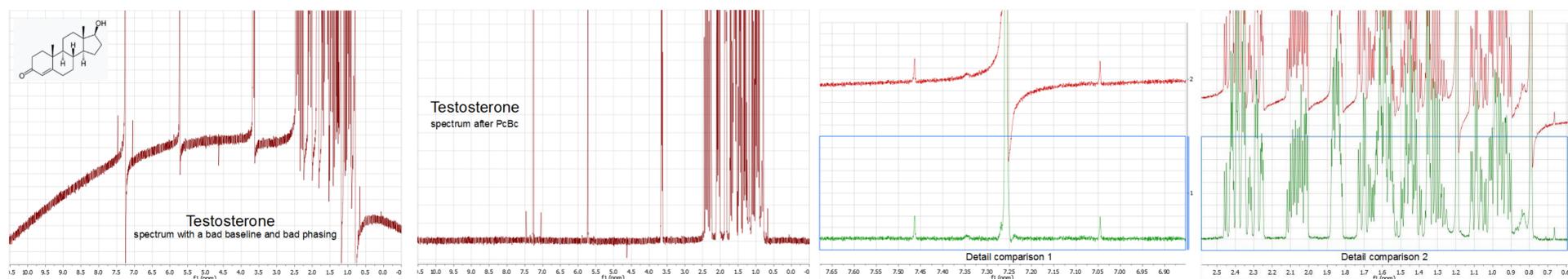
The weight functions

Modelling efficient weight functions suitable for our purposes is not an easy task because there is no rigorous mathematical theory which might help. For the imaginary part, we for now base it just on a-symmetry, making sure that for perfectly symmetric histograms the result is zero, while any a-symmetric shape generates a negative value. For the real part, it is evident that the area around $\zeta = 0$ should have the largest height, any negative values should be discouraged (negative penalty), and discouraged more than positive values for which the adjusted histograms still maintain some intensity, albeit small. There are several other considerations affecting the modelling of the weight functions which, however, exceed the scope of this poster. The rather simple real-part weight function we have used for the preliminary results shown below is plotted on the left side of this paragraph.

The flowchart

Consequently, in extreme synthesis, the flowchart to compute the quality functions Q(p) follows these steps:

1. <take the experimental spectrum> ==>
2. <apply PC + BC corrections defined by the parameters p> ==>
3. For $Q_r(p)$: <compute the histogram of the resulting spectrum> ==>
4. For $Q_r(p)$: <evaluate the weighted integral of the histogram>

3'. For $Q_i(p)$: <compute the integral over -[imag(H(ζ) - H(-ζ))]²> (this is maximum (0.0) when H(ζ) is symmetric and negative otherwise)
4'. For $Q_i(p)$: <multiply the result by a suitable weight constant>

Some preliminary results using 5 harmonic coefficients for the baseline correction

Testosterone
spectrum with a bad baseline and bad phasing

Testosterone
spectrum after PcBc

Detail comparison 1

Detail comparison 2

It seems to work! (though some more work is still needed)

References

1. Cobas C., Bernstein M.A., Martin-Pastor M., Tahoces P.G., 2006. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. Journal of Magnetic Resonance 183, 145-151;
2. Binczyk F., Tarnawski R., Polanska J., 2015. Strategies for optimizing the phase correction algorithms in Nuclear Magnetic Resonance spectroscopy. Biomedical Engineering On line 14 (S2);
3. Cobas C. and Sykora S., 2007. Extended DISPersion-Absorption (eDISPA) approach to Automatic Phasing of HR-NMR Spectra. 48th ENC Conference, 2007, DOI: 10.3247/SL2Nmr07.004.
4. Zorin V., Bernstein M.A. and Cobas C., 2017. a robust, general automatic phase correction algorithm for high-resolution NMR data, Magnetic Resonance in Chemistry 55, 738-746.