Global Spectral Deconvolution (GSD) of 1D-NMR spectra

Carlos Cobas [1], Felipe Seoane [1], 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

MESTRELAB RESEARCH
NMR Solutions

INTRODUCTION to the GSD concept

No matter how "perfect" a typical 1D-NMR spectrum might be, it *always* contains the following "parts":

1. A set of spectral "peaks" with NMR characteristics (such as RF-pulse dependent nutation angle and phase),
2. Baseline distortions due to receiver artifacts (channels U and V offsets), probe ringing and acquisition dead time.
3. Spikes and other "environmental" RF signals entering the detector but not generated by the nuclei in the sample.
4. The ubiquitous receiver noise which, at least in first approximation, is completely independent of the sample.

We are used to cope daily with the co-presence of all these parts, even though only the one indicated as (1) carries meaningful information. Using a number of distinct algorithms, an experienced operator copes with the co-presence reasonably well when processing the data manually, but it often plays a havoc with automatic routines, such as fitting theoretical spectra of coupled-spin systems to experimental data or, for that matter, just attempting a simple integration. Hence,

the first goal of GSD is to extract from the spectrum only part (1)

The second problem one encounters regards the imperfections of part (1) itself. In particular:

a. Rather then being Lorentzian, peaks are distorted by field inhomogeneity and molecular dynamics.
b. Even for modest molecules, there are many orders more theoretical transitions than experimentally recognizable peaks. This may affect the apparent peak widths and shapes even than inhomogeneity.
c. Due to their finite linewidths, peaks overlap heavily giving rise to hump-like peak convolutions.
d. Some peaks are meaningless artifacts (rotational sidebands) and others, though physically meaningful, are undesirable (solvent and impurity peaks and, in some contexts, isotopomer signals such as ^{13}C sidebands). All such peaks should be "edited" out, something that is very difficult to do directly in an analog spectrum.

Again, a complete, automated procedure dealing with all these problems at once has so far never been tried. At best, there are separate procedures like reference deconvolution, resolution enhancement algorithms, multiplet deconvolution, etc. which address some of these problems one at a time. Hence,

the second goal of GSD is to set up a numerical Peaks List such that a synthetic spectrum generated from it differs from part (1) by less than the experimental noise.

The way the GSD algorithm works, the two goals are closely intertwined and can not be intended as separate phases. One can view GSD as an extension to the whole spectrum of the classical multiplet deconvolution (hence the name) but actually it goes far beyond it. For one thing, it is fully automatic and superior to a human in deciding the number of peaks and their starting parameters (peak recognition).

Another way of looking at GSD is as an information filter which discards the undesirable parts of the spectrum. The peaks list it produces contains **all** the desirable information present in the original spectrum but **none** of the undesirable one. All subsequent data processing tasks (such as integration, digital J-Correlator, or structure verification and/or elucidation) can work exclusively on this numeric information without bothering any more about the imperfect original spectrum. Moreover, the list is very easy to **edit** both by the User and automatically.

UNDER the HOOD

GSD operates in a fully automatic mode on either real or complex f-domain data. At present, an approximate prior phase adjustment is required. What follows is of course just a very schematic exposition of the algorithm. The individual steps will be illustrated on a spectrum of *brucine*. All evaluatins were carried completely automatically and on the whole spectrum but, for reasons of clarity, the performance is illustrated on selected narrower windows.

Step 1: Noise and mean linewidth estimate

The pre-requisite for a correct operation of GSD is a reliable estimate noise and mean linewidth. To do that, we have devised two new and amazingly robust algorithms which will be published and discussed elsewhere.

Step 2: Automatic calculation of first and second derivatives. This is done by means of the Savitsky-Golay convolution algorithm (SG). The trick, of course, is a correct setting of the SG parameters (number of points and order), based on the estimated noise and linewidth. The use of derivatives all but removes any baseline dependence. Moreover, the use of the 2nd derivative enormously enhances resolution (like in our Resolution Booster™) and converts bare shoulders into distinct peaks. The noise in each of the derivative spectra is exactly known and its value can be used for subsequent evaluation.

Figure (a) shows the 4 – 4.3 ppm portion of the spectrum (bottom) and its first (middle) and second (top) derivatives.

Step 3: Special points classification

An efficient peak-picking algorithm (again based on correct noise estimate) is applied to the original spectrum as well as to the first and the second derivative spectra. Points where recognizable local minima and maxima occur in each of the three arrays are appropriately flagged (in case of a complex spectrum, this is done for both the real and imaginary parts).

Figure (b) shows the same portion as before. Negative marks correspond to local minima, positive ones to local maxima. The 3 divisons long marks refer to the original spectrum, the shortes ones (half a division) to the 1st derivative and the one division tall marks to the negative second derivative (in case of superposition, the mark lengths are added).

Step 4: Peaks recognition

All recognizable peaks are detected and marked, using the positive maxima in the second derivative spectrum (resolution and total baseline independence) and the maxima in the original spectrum (for broad features). Peaks with estimated linewidth smaller than 2 points are discarded as spikes.

Figure (c) shows the peak recognition algorithm at work. The 0th and 2nd derivative marks are as before, but the 1st derivative mark pairs (positive-negative) are replaced by 5 divisions tall rectangles. Notice that some very small peaks (barely above noise) were correctly picked up, some 1st derivative rectangles are topped by more than one 2nd derivative peak and some by none, etc. There is not enough room on this poster to explain all the fine points of this peak recognition method.

Step5: Raw setting of each peak parameter (frequency, linewidth, height)

To avoid baseline dependence, the parameters estimate is based on the knowledge of only the special points (maxima and minima) of the 1st and 2nd derivative spectra. The result is a raw Peaks List. The latter can be used to generate a synthetic spectrum which already resembles quite well the experimental one but is void of any baseline roll and noise.

Figure (d) shows the 3.6 – 4.3 ppm portions of the experimental spectrum (up) and of the synthetic spectrum (down) computed from the raw Peaks List. Notice that the comparison, though not perfect, is certainly excellent as a starting point for the fit.

Step 6: Refinement of the parameters of all the peaks in the list

There are several modes how to do this which differ somewhat in final quality and dramatically in speed (work in progress). We will eventually choose two modes – fast and slow - among which the User of the software will be able to choose. The goal is, of course, a perfect match of the spectrum – minus the baseline and noise and spikes. Indeed, this step is not illustrated here because the fit is perfect – disturbed only by the baseline artifacts in the experimental spectrum which are absent in the synthetic one as shown in Figure (e).

What is it good for?

We have already mentioned several of the advantages of passing a spectrum through the GSD procedure and reducing it to an equivalent and clean list of peaks. Think only about integration which is replaced by summing peak areas with a total exclusion of peak overlap errors. Or think about resolution enhancement. Or, for that matter, baseline correction - this is the only method so far that can potentially find out the baseline of both real and imaginary parts.

For us, however, this is primarily a necessary step towards automated and/or computer aided molecular structure verification and elucidation. The Peaks List, rather than the spectrum, becomes The Input to further editing and analyses aimed at the determination of all compatible spin systems and, eventually, molecules.

(a)

(b)

(c)

(d)

(e)