QMSA Letters, 3(2025)

Prior Knowledge* and Constraints

Reino Laatikainen, Pekka Laatikainen and Henri Martonen
Spin Discoveries Ltd., Kuopio, Finland

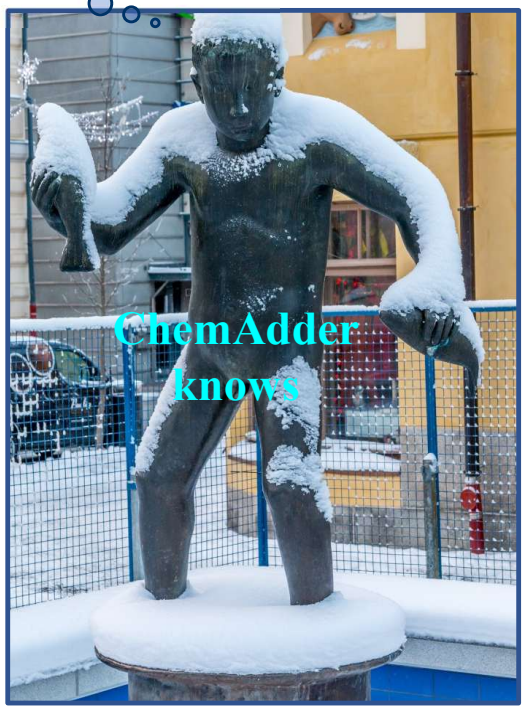
Last update Feb. 6th, 2025

*WIKIPEDIA: Prior knowledge refers to what a learner already knows before learning new information. That is, it's the information and educational context already present before new instruction. Prior knowledge is important as it serves as a foundational building block for new knowledge. … (not WIKIPEDIA) the phrase has also an intimate meaning!

IBM 360, Yess!

IBM 360 was one of the computers used with LAOCOON and NUMARIT

We know, fuzzy!



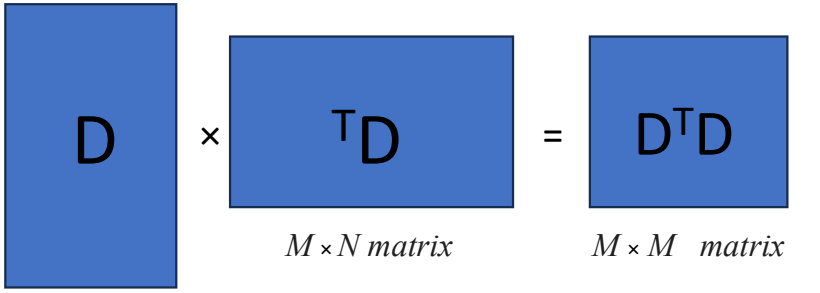
From fuzzy to truth

Normal Equations

The objective of QMSA is to minimize the sum of squares between the observed and calculated NMR spectra:

$$\text{sum of squares} = \sum (I(n)_{obs} - I(n)_{calc})^2$$

The iterative protocol is based on differentials $\partial I(n)_{calc}/\partial P(m)$, where (I(n) = spectral intensity at point m, (P(m) = spectral parameter (chemical shifts, coupling constants, line widths, line shape, baseline, …), for details see QMSA Letters. The task can be written into form of group of N non-linear equations, having M unknowns (where N and M can be 1-512000 and 1-4000):



The equations can be written into form of an N*M matrix. Because there is far more spectral points (N) than unknowns (M), the equations can be solved only into criterium of least-square, via **normal equations (D^TD matrix)**, which can be solved using the standard matrix algebra.

The equations can be used to get improvements to trial parameters. Usually, an iterative procol in which the trial parameters are adjusted toward the least-square fit is needed.

The elements of the matrix D are differentials: $D(n,m) = \partial(I(n)_{obs} - I(n)_{calc})/\partial P(m)$

D^TD is a symmetric sparse matrix: most of its elements are 0. The improvements of the parameters (dP) are obtained from a matrix equation: $dP = (D^TD)^{-1} D^T dI$ where dI is the observed-calculated intensity vector

Prior Knowledge and Constraints

If the trial parameters are obtained from a database (ASL) and the spectrum is measured in similar conditions (solvent, pH, concentrations, temperature and same reference), the trial chemical shifts are usually obtained within a range of a few Hz. If they have been obtained from using spectral parameter prediction and the structure is uncommon, the **range** is seldom better that 0.1-0.2 ppm. See that a range 0.01 ppm means 6 Hz at 600 MHz, while it is only 0.8 Hz at 80 MHz.

ChemAdder offers two ways of adding the prior knowledge to the iteration:

- In hard constraint, iterator does not allow parameter to exit the DEFAULT±RANGE
- As soft constraint, the iterator constraints the parameters to the DEFAULT values in the least-square way, by manipulating the **Normal Equations**. The strength of the constraint can be defined by FORCE. In ChemAdder FORCE = 1.0 means that the constraint is strong as the spectral information about the parameter. The force can be > 1, and it can be gradually decreased if the iteration converges.
- Examples of typical constraints:
  - A spectral parameters are forced to their predicted values and ranges.
  - All the linewidths are forced to the same value. Useful in beginning of the menu.
  - The linewidths of a compound are forced to the same value. Useful if the population of the compound is low.
  - All the response factors or within a compound are forced to 1.0.
  - The baseline is forced linear and/or positive.
  - ….

Principal Component Regression (PCR)

Sometimes a spectral parameter has no/only a small effect on the outlook of the spectrum, for example if the information is hiding under a major compound spectrum. This is seen from the corresponding diagonal element of the **Normal Equations**.

Two parameters may have an identical or similar effect on the outlook. We say that the parameters are then CORRELATED. The correlation is seen from the off-diagonal elements the **Normal Equations**. This is typical for strongly second-order system.

A general solution for the above is to use PCR: the original parameters are replaced by their linear-combinations, which are not-correlated – same as *orthogonal*.

The **Rank** gives the number of orthogonal (not correlated) parameters of the system. The rank can be controlled by the **Threshold** value: by setting it to 90%, the PCR uses only the (Rank) linear combinations that explain 90% of the **total variance**. The total variance describes the sensitivity of the spectrum to the parameters and is here the sum of the diagonal elements (=**Trace**) of the **Normal Equations**. Each linear combination has its eigenvalue, which reflects the sensitivity of the spectrum to the combination.

PCR in combination with Broadening offers a fast tool to a population analysis of simple samples. Unfortunately, we have not had time to explore it to details.

Equalizing, Weighting and Locking

For example, the ^{2}J coupling of CH_2-groups has only a weak effect on spectral outlook if the chemical shifts of the CH_2 protons are similar. If many CH_2-signals overlap, there is not much spectral information about the differences of the couplings. Therefore, to prevent the iteration to diverge, the couplings can be kept fixed, constrained to defaults or **equalized.** The values also follow simple rule: an electronegative substituent decreases, double bond and aromatic substituent increases the coupling.

The intensity of, for example, aromatic region in biofluid spectra is usually much lower than that of aliphatic region. If a compound has signals in both the regions, and the parameters in the aliphatic region are poorly defined, the iteration often diverges. The divergence can be prevented by overweighting the aromatic region by multiplying the region by 10. If a spectral alternatively some ranges or signals can be underweighting them by multiplying with 0.01 or 0.10.

Stepwise protocol:

1. Optimize fit for aromatic region first, and if iteration converges, **LOCK** the optimised parameters.
2. Optimize the aliphatic region.

It is not necessary optimize both the ranges together if the number of parameters grow too large (>1000) !

Conclusion … to be added