Application of the Blind Source Separation to the mixture analysis by NMR, toward the demixing of the 3D-DOSY experiments

E. Piersanti[a], A. Cherni[b], S. Anthoine[b], C. Chaux[b], M. Campredon[a], B. Torrésani[b], L. Shintu[a] and M. Yemloul[a]
[a]Aix Marseille Univ, CNRS, Centrale Marseille, iSM2, Marseille, France. 52 Avenue Escadrille Normandie Niemen - 13013 Marseille.
[b]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France. Chateau Gombert – 39 rue F. Joliot Curie – 13453 Marseille.
elena.piersanti@etu.univ-amu.fr

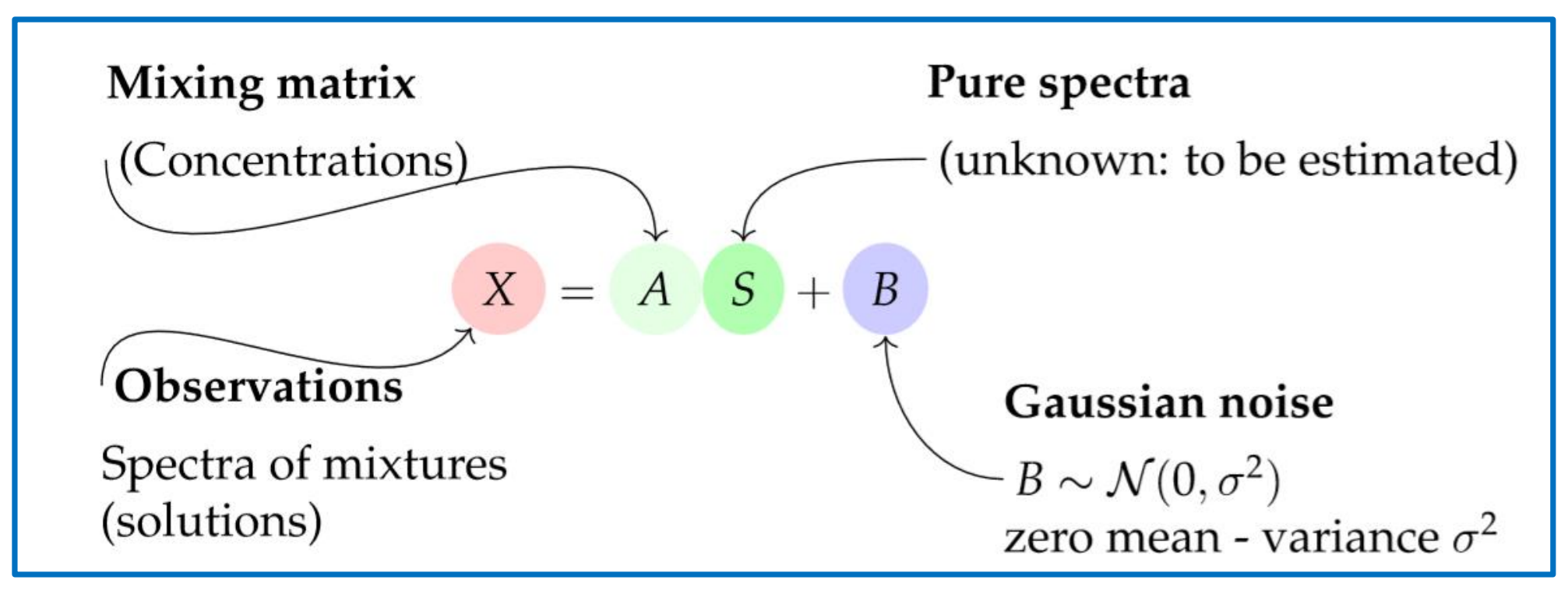
General Background: Despite the development of NMR methods to increase spectral resolution, the growing complexity of the samples leads to crowded spectra that compromise the analytical performances of this technique. The association of new mathematical methods for signal processing with the methodological developments in NMR is a promising alternative in this field.

We present the application of Blind Source Separation (BSS)[1] algorithms to NMR data. This source separation technique, originally used for disciplines such as acoustics and audio signal processing,[2] has shown its effectiveness for the demixing of 1D and 2D NMR spectra[3]. In this case, spectra deconvolution is performed using correlations (essentially concentration variations) detected over a series of data sets, which allows the pure spectra of each of the mixture constituents to be extracted.

BSS Approach: BSS - based methods aim at the separation of a set of *pure* signals (*sources* = S spectra) from a set of complex mixed signals (*mixtures* = X spectra):

X = AS + B ≈ AS where: X = observation matrix (nD observed spectra), A = mixing matrix, S = pure spectrum of each compound to be estimated, B = the residual noise, where the mean gaussian matrix is usually set to zero *(Fig. 1)*.

Among BSS problems, the simplest instance originates from the Linear Instantaneous Mixture (LIM) model where the observed mixtures are linear combinations of the sources. These algorithms are blind because they have to estimate unknown matrices A (the concentrations in solutions, represented by A coefficients) and S, from X.



$$\underset{\mathbf{A},\mathbf{S}}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{X}-\mathbf{AS}\|_F^2 + \Psi_{\mathbf{A}}(\mathbf{A}) + \Psi_{\mathbf{S}}(\mathbf{S})$$

$$\begin{cases} \Psi_{\mathbf{A}}(\mathbf{A}) & = \iota_+(\mathbf{A}) \\ \Psi_{\mathbf{S}}(\mathbf{S}) & = \lambda_1 \ell_1(\mathbf{WS}) + \lambda_2\,\mathrm{Ent}(\mathbf{S}) \end{cases}$$

Under constraints A≥0, S≥0

Fig. 1. Mathematical tools: modelling.

Previous work[3, 4] on 1D DOSY spectra → **BSS** → **It works well!**

Variable Gradient Strenght

To seek the spectral separation, several approaches based on different assumptions have been evaluated to lead to different identification algorithms. In particular, we used the Independent Component Analysis **(ICA-JADE)**, which assumes that the sources are statistically independent, and the Non-Negative Matrix Factorization **(NMF)**, where A≥0 and S≥0.

Evaluation criteria: To assess the quality of the estimated sources we used Signal to Distortion Ratio (SDR) (in dB), which provides a global measure of the distortion introduced by mixing and separation, and Signal to Interference Ratio (SIR), which supplies a quantitative evaluation of crossover terms after separation (peaks from a given source that could be completely or partially found in the estimation of another source). The higher these ratios, the closer to the original is the estimated source S.
For the matrix A, its estimation quality can be evaluated with the Amari index: 0 (good estimation) ≤ Amari ≤ 1 (bad estimation).

Studied samples

Five synthetic mixtures of four terpenes ((R) - (+) - Limonene, Nerol, α-Terpinolene, (-) – transCaryophyllene) *(Fig. 2)* were prepared by varying the concentrations of each compound in 600 µl of $CDCl_3$ and sealed tubes. M mixtures ≥ N sources are used to estimate N from M. Terpenes are natural molecules found in plants with highly crowded spectra between 1.5 and 2.5 ppm.

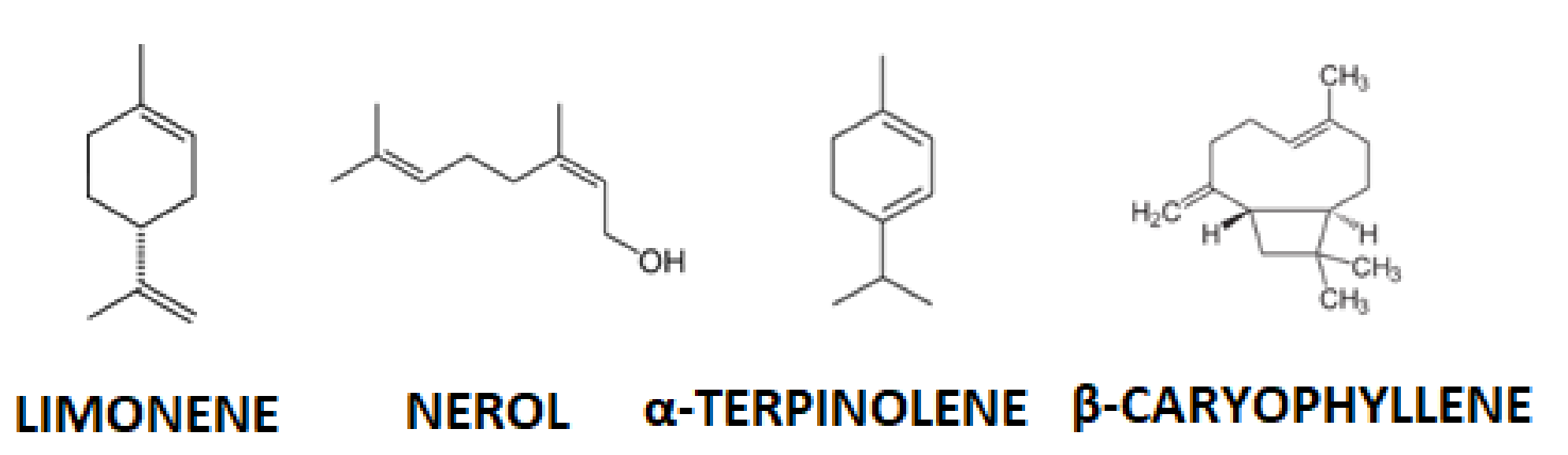


Fig. 2. Chemical structure of the four terpenes.

Separation Results

In the case of 2D HSQC spectra (*Fig. 4, 5*), the performances of the algorithms are fairly good on simulated data. The results on real 2D mixtures are of weaker quality in terms of the objective performance evaluation indices. However, the increased sparsity of 2D spectra allows a good identification of the components of mixtures. In addition, concentrations appear to be better estimated than in the 1D case (*Fig. 3)*, which may also be interpreted as a consequence of the sparsity of 2D spectra. The computational burden is significantly increased in the 2D case, which may be a limitation. The best estimated concentrations are obtained using BCVMFB with wavelets with λ = 10 σ.

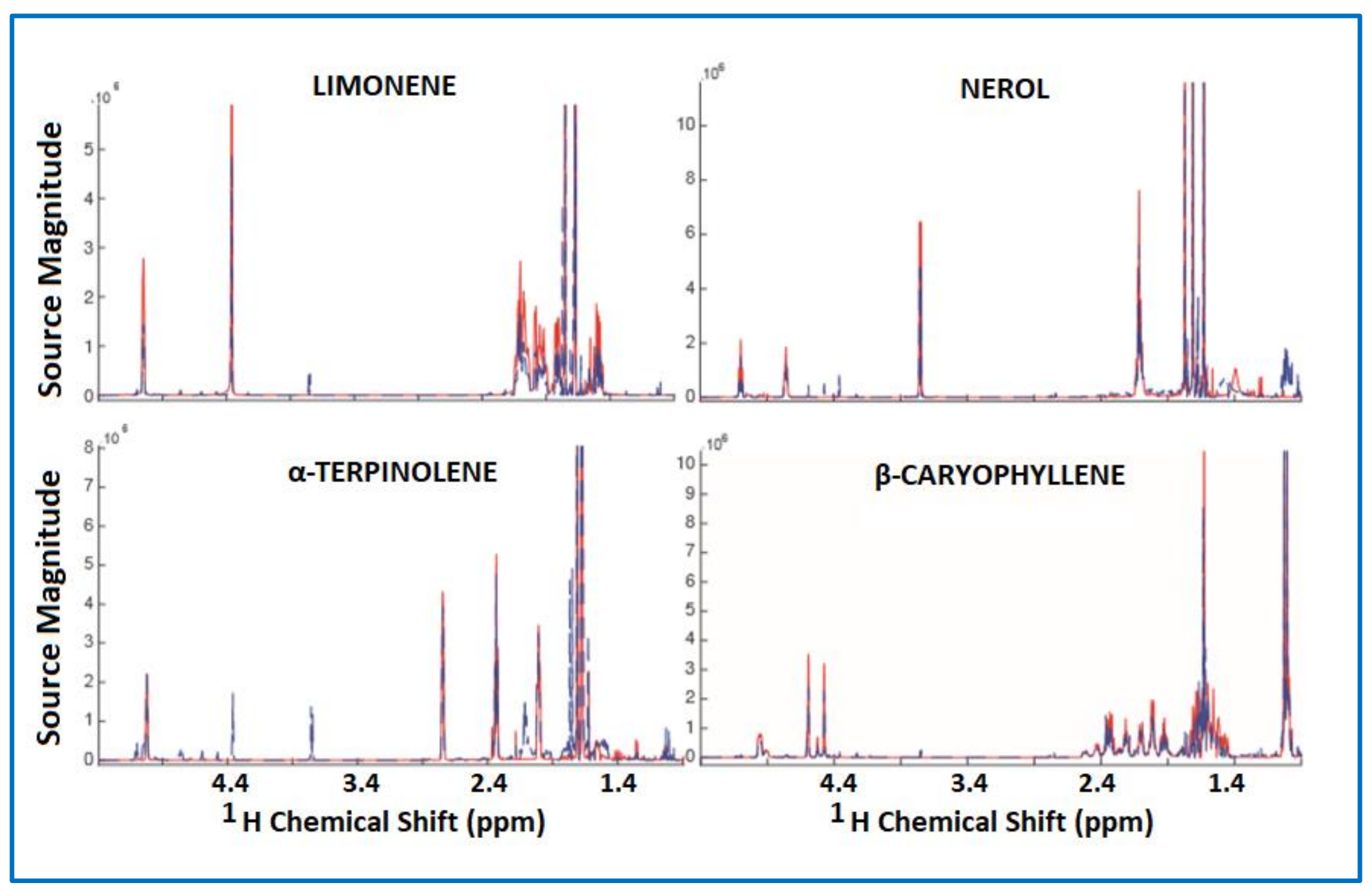


Fig. 3. Results on ^{1}H NMR spectra. Estimated sources versus real sources of the 5 mixture spectra with variable concentrations of the 4 terpenes.

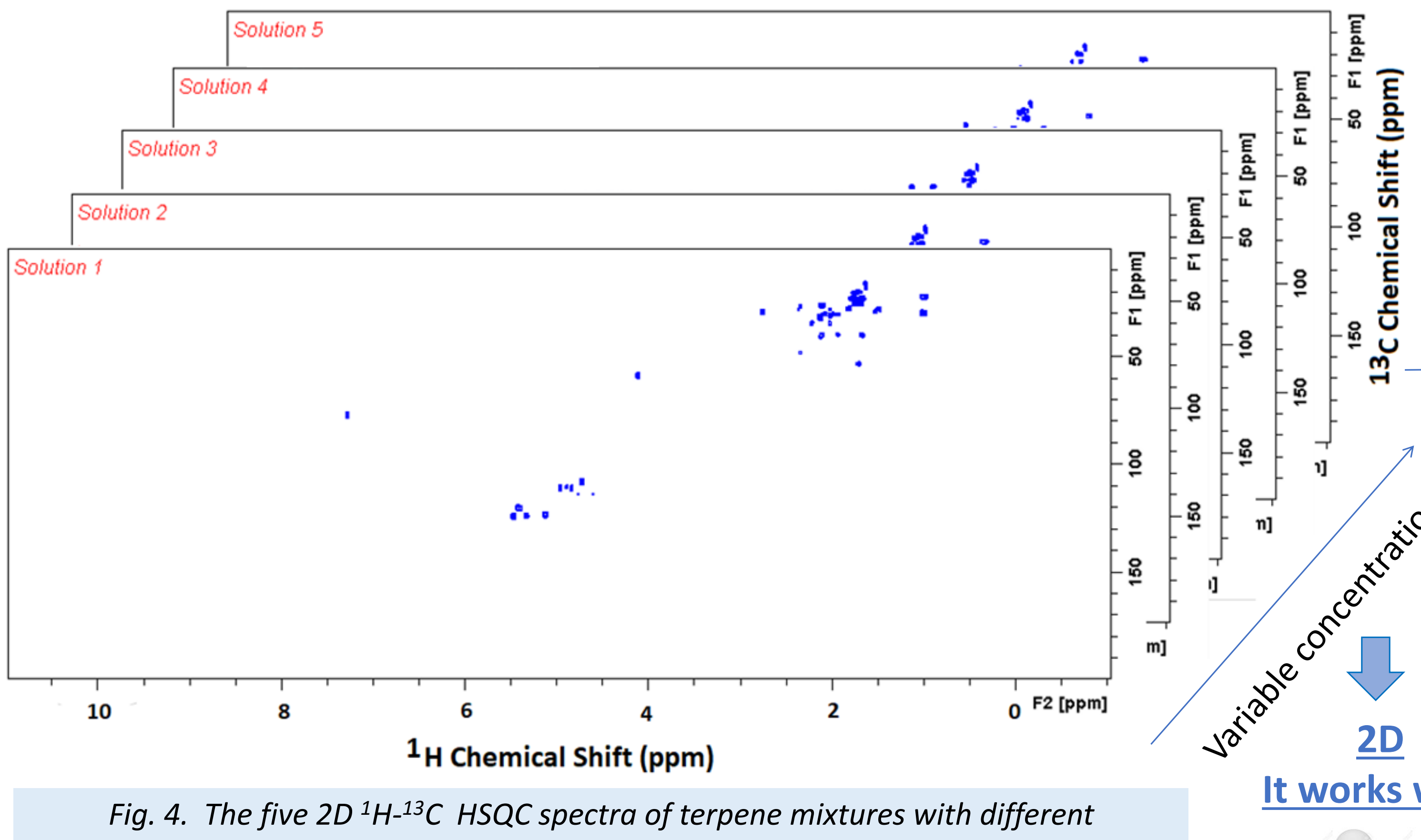


Fig. 4. The five 2D ^{1}H-^{13}C HSQC spectra of terpene mixtures with different concentrations of each pure coumpound, that means different intensities of the spots.

BSS

Variable concentrations

2D It works well!

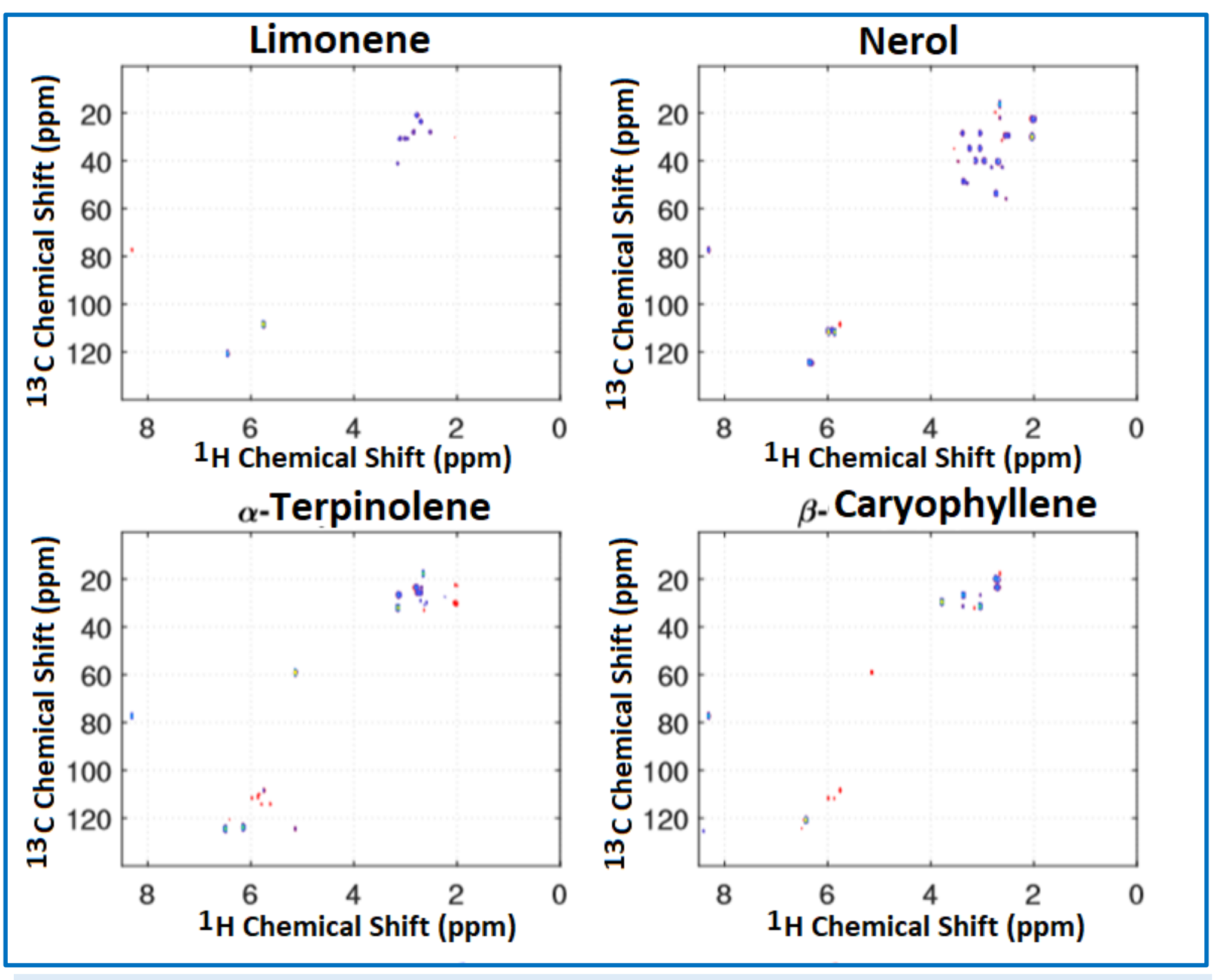


Fig. 5. Results on 2D ^{1}H-^{13}C HSQC spectra. Estimated sources versus real sources.

New results on 1D 1H spectra → **It works well!**

Work in progress & Perspectives

In a recent study[4], our group presented a strategy for processing DOSY experiments based on the synergy of two blind high-performance (BSS) techniques: the Non-Negative Matrix Factorization (NMF) and the joint diagonalization of the clean matrices (JADE, that can be used to estimate A_0 and S_0 for the initialization step). These approaches improved the processing of DOSY experiments for mixtures with strong overlaps in the spectra. As an outgrowth of this work, we will evaluate the efficiency of the BSS algorithms for the extraction of pure 2D-HSQC spectra from 3D DOSY-HSQC experiment data obtained on terpene and amino acid mixtures.
3D spectra of new amino acid complex mixtures are in progress: Proline, Lysine, Tyrosine, Histidine, Phenylalanine, Tryptophan.

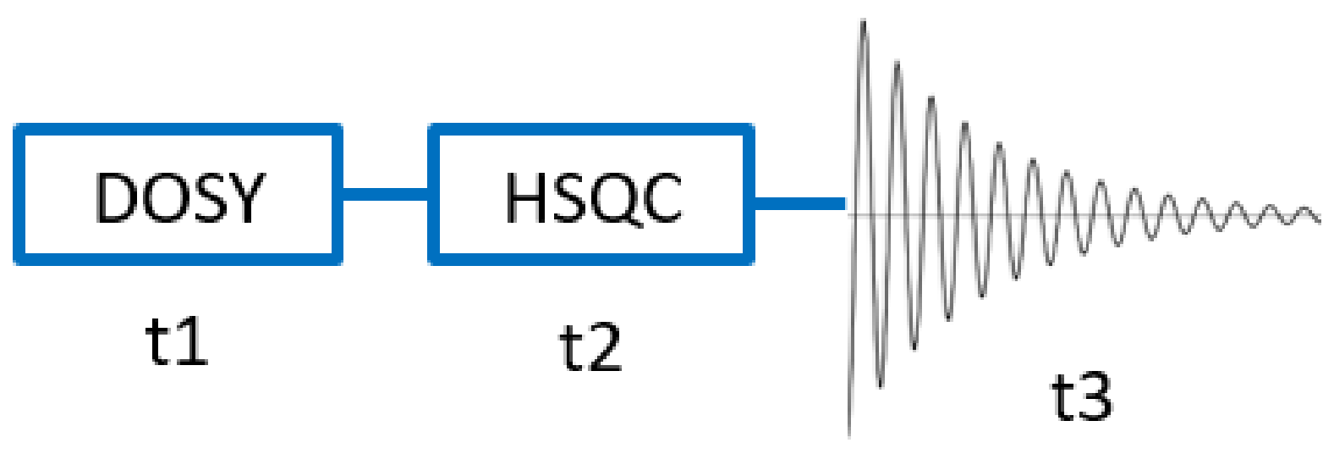


*Fig. 6. The convection compensated **3D DOSY-HSQC** was obtained by concatenation of dstbpgp3s and hsqcetgpsi pulse sequences. The INEPT delay adjusted to a ^{1}H-^{13}C coupling constant of 145Hz. Implementation of quantitative HSQC is essential for better estimation of the concentrations.*

This approach would be well suited for with only one available sample. By applying the "M mixtures ≥ N sources rule" only five variable gradients could be enough for the terpene sample and six for the amino acid sample.

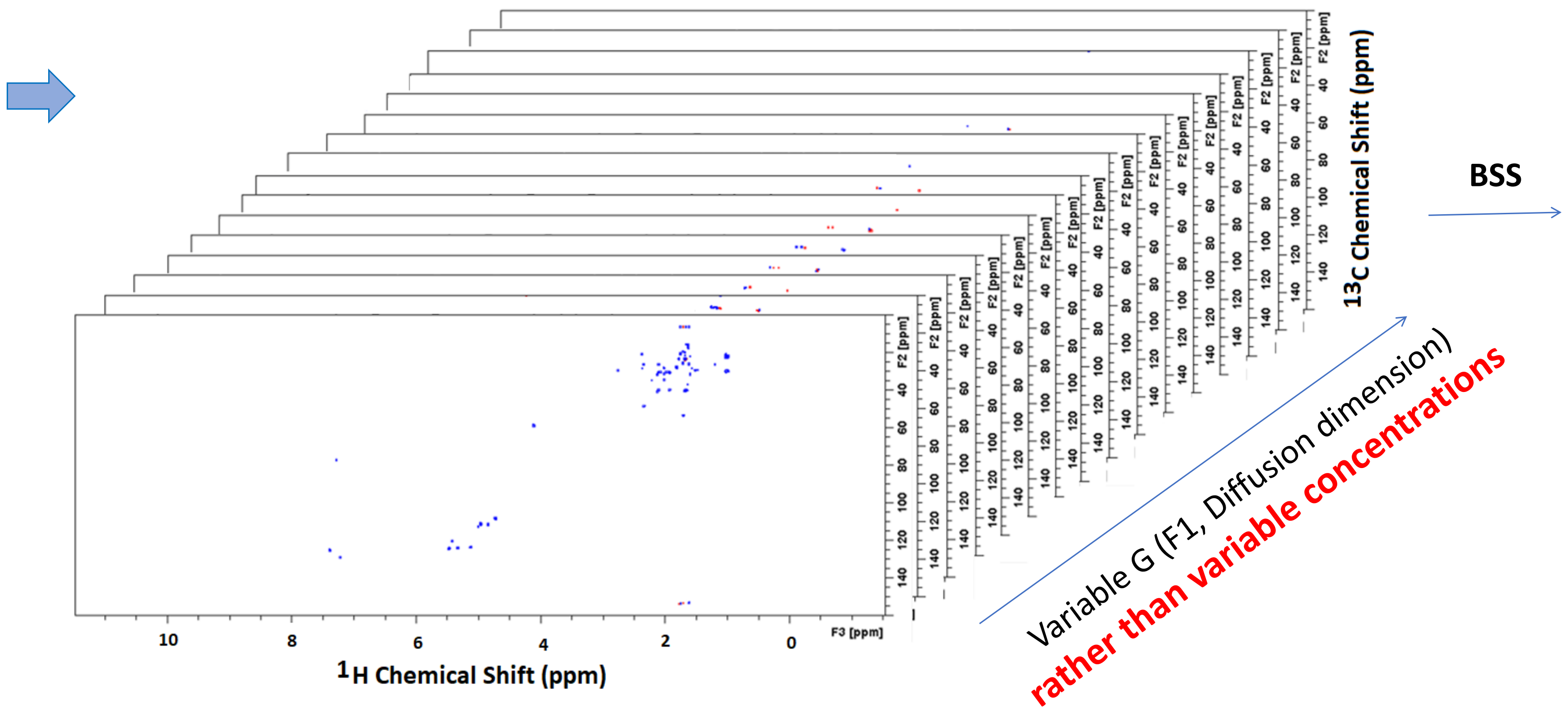


Fig. 7. The sixteen 2D 1H-13C HSQC spectra (F1=16) of terpene mixtures from the 3D spectrum with different Gradient strenght (G/cm) as a function of diffusion that means different intensities of the spots.

BSS

3D Work in Progress

Extraction of pure 2D-HSQC spectra from 3D DOSY-HSQC experimental data

Conclusions:
The results presented here show that BSS algorithms are able to perform successfully. Algorithms are extremely sensitive to initialization. In nD spectra in general, the dimensionality increases complexity and computational burden. This may be considered paradoxical, as the structure of 2D spectra seems simpler or at least much sparser than 1D spectra.

References:
1. C. Chaux, P.L. Combettes, J. C. Pesquetand, V. R.Wajs, *Inverse Problems*, 23, 1495 – 1518 (2007)
2. P. Comon and C. Jutten, Handbook of Blind Source Separation: Independent component analysis and applications, *Academic press*, (2010)
3. I. Toumi, S. Caldarelli and B. Torrésani, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 81, 37 – 64 (2014)
4. I. Toumi, B. Torrésani and S. Caldarelli, *Anal. Chem.*, 85, 11344 – 11351 (2013)

http://www.i2m.univ-amu.fr/project/bifrost/
BIFROST
Blind Identification, Filtering & Restoration On Spectral Techniques