Challenges in the decomposition of 2D NMR spectra of mixtures of small molecules

Afef Cherni, Elena Piersanti, Sandrine Anthoine, Caroline Chaux, Laetitia Shintu, Mehdi Yemloul, Bruno Torrésani

Challenges in the decomposition of 2D NMR spectra of mixtures of small molecules

Afef Cherni [a], Elena Piersanti [b], Sandrine Anthoine [a], Caroline Chaux [a], Laetitia Shintu [b], Mehdi Yemloul [b], and, Bruno Torrésani [a]

Analytical methods for mixtures of small molecules requires specificity (is a certain molecule present in the mix?) and speciation capabilities. NMR has been a tool of choice for both of these issues since its early days, due to its quantitative (linear) response, sufficiently high resolving power and capabilities of inferring molecular structures from spectral features (even in the absence of a reference database). However, the analytical performances of NMR are being stretched by the increased complexity of the sample at hands, the dynamic range of the components, and the need of a reasonable turnover time. One approach that has been actively pursued for disentangling the composition complexity is the use of 2D NMR spectroscopy. While any of the many experiments from this family will increase the spectral resolution, some are more apt for mixtures, as they are capable to unveil signals belonging to whole molecules or fragments of it. Among the most popular ones one can enumerate HSQC-TOCSY[1], DOSY[2] and Maximum-Quantum (MaxQ) NMR[3]. For multicomponent samples, the development of robust mathematical methods of signal decomposition would provide a clear edge towards identification. We have been pursuing, along these lines, Blind Source Separation (BSS). Here, the un-mixing of the spectra is achieved relying on correlations detected on a series of datasets. The series could be associated to samples of different relative composition or in a classically acquired 2D experiment by the mathematical laws underlying the construction of the indirect dimension, the one not recorded by the spectrometer. Many algorithms have been proposed for BSS in NMR[4] since the seminal work of Nuzillard[5]. In this paper, we use rather standard algorithms in BSS in order to disentangle NMR spectra. We show on simulated data (both 1D and 2D HSQC) that these approaches enable to disentangle accurately multiple components, and provide good estimates for concentrations of compounds. Furthermore, we show that after proper realignment of the signals, the same algorithms are able to disentangle real 1D NMR spectra. We obtain similar results on 2D HSQC spectra, where BSS algorithms are able to disentangle successfully components, and provide even better estimates for concentrations.

1 Introduction

Nuclear magnetic resonance (NMR) is a powerful spectroscopy that provides comprehensive information on molecular structure and is well suited for the detection and identification of small molecules. The simplest NMR experiment yields to a potentially informative one-dimensional spectrum that typically results in overlapping signals in complex mixtures and hinders the identification and the quantification of components. Several developments, occurring at different stages of the NMR analysis (from pulse sequences implementation to data processing), have been proposed to differentiate between multiple peaks in extensive crowded spectra. One powerful approach to increase the information content of NMR spectra is to acquire two-dimensional data. In that regard, HSQC[6], TOCSY[7], Maximum-Quantum (MaxQ)[3,8] or Diffusion Ordered SpectroscopY (DOSY)[9] NMR experiments are very interesting for mixture analysis since they allow the direct identification of a whole molecule or fragments of it. Indeed, DOSY experiments represent a commonly used pseudo-2D NMR experiment that allows the differentiation of molecules in a mixture according to their diffusion coefficients. This technique is efficient when applied to relatively simple mixtures (less than ten molecules) leading to the extraction of the NMR spectrum of each compound. However, a limiting factor of its applicability is the requirement of a mathematical treatment capable of distinguishing molecules with similar spectra or diffusion constants. Similarly, 2D NMR experiments such as ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC are also performed routinely since they are necessary for the assignments of the mixture's molecules[10]. However, their use for the analysis of a high number of samples, as

[a] *Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France*
[b] *Aix Marseille Univ, CNRS, Centrale Marseille, iSM2, Marseille, France*

occurring in metabolomics, is difficult since the acquisition time of a 2D NMR spectrum can be extremely time-consuming, and in that case, 2D NMR experiments are only performed on representative samples. In addition, molecule assignment relies on a very exhausting and time-consuming process for which each signal of each 2D NMR spectrum must be thoroughly peak-picked and gathered according to the molecule they characterize. In order to overcome the issue of an extensive acquisition time, ultrafast and fast NMR methods such as single-scan [11] or non-uniform sampling (NUS) [12] techniques have been developed, making possible the use of 2D NMR experiments for high-throughput study of complex mixtures [13]. Regarding the signal assignments and in the case of well-studied samples such as blood plasma or cerebrospinal fluid, automated methods for the metabolite identification from 2D experiments are available online [14,15]. However, their efficiency depends on strict sample preparation protocols that limit their use for a wider range of samples. Consequently, the development of robust mathematical methods that would perform signal decomposition is of prime importance for the analytical study of complex mixtures. The mathematical "demixing" of 1D or 2D NMR spectra would thus provide a clear edge towards identification, with a non-negligible gain of time. In a previous study, our group presented a processing strategy for DOSY experiments based on the synergy of two high-performance Blind Source Separation (BSS) techniques: Non-negative Matrix Factorization (NMF) using additional Sparse Conditioning (SC), and the JADE (joint approximate diagonalization of eigenmatrices) declination of independent component analysis (ICA) [4,16,17]. Both approaches enabled to improve the processing of DOSY experiments, in cases of mildly overlapping species. For mixtures with strong overlapping signals of moieties with similar diffusion coefficients, such as a mixture of sucrose and maltotriose, NMF-SC provided a very good method for molecule separation, although not perfect and needing improvement to make it suitable for the processing of more complex mixtures [18].

In this paper, we address the un-mixing problem in a more general setting, with the aim of processing 1D as well as nD mixtures. Sticking to the family of non-negative matrix factorization approaches to BSS, we consider several algorithms and apply them to simulated and real mixture spectra. These are presented in a unified framework, that includes classical NMF approaches such as alternate least squares (ALS) [19] and sparsity-penalized versions [20], proximal approaches [21], and wavelet-based [22] variants. Interestingly enough, the framework also includes algorithms for un-mixing nD spectra. We then provide objective performance evaluations for un-mixing algorithms, using quality indices that allow assessing the quality of the estimation of source spectra and concentrations in the mixtures.

Results are given for a dataset that has been prepared on purpose, for which pure spectra and concentrations in solutions are available, which allows computing the above indices. Results on 1D simulated data (i.e. mixtures mathematically generated from pure spectra and concentrations) show that the algorithms under consideration are indeed able to recover the ground truth. Results on real 1D mixtures do not reach the same level of quality even after correcting alignment biases, which raises concerns regarding the mathematical mixture model. In the case of 2D HSQC spectra, the performances of the algorithms are again fairly good on simulated data. Results on real 2D mixtures are of weaker quality in terms of the objective performance evaluation indices. However, the increased sparsity of 2D spectra allow a good identification of the components of mixtures. In addition, concentrations appear to be better estimated than in the 1D case, which may also be interpreted as a consequence of the sparsity of 2D spectra. It is worth mentioning that the computational burden is significantly increased in the 2D case, which may be a limitation. The 2D case then represents an important challenge, this is presumably true for higher dimensional spectra.

Besides objective evaluations, visual inspection of spectra show that the mathematical un-mixing algorithms are able to identify pure compounds in the solutions under study, in several situations. This is clearly the case in simulated situations, which show that the algorithms are able to identify compounds when mixtures have been generated under a well defined model. This is also the case, to a smaller extent, in the case of ^{1}H real data, provided pre-processing steps have been carefully performed (in particular shift correction). Our results however raise a number of important questions, that include among others the validity of the mathematical mixture model, the possibility of performing some pre-processing steps simultaneously with un-mixing, but also the relevance of quantitative assessment measures in the context of NMR spectroscopy un-mixing.

2 Problem statement

2.1 Blind source separation, the Linear Instantaneous Mixture model

Blind source separation (BSS), aims at the separation of a set of *pure* signals called *sources* from a set of mixed signals, called *mixtures*, with limited information on sources or the mixing process. Sources are generically represented as an $M \times L$ matrix $S = \{s_{m\ell}\} \in \mathbb{R}^{M\times L}$, mixtures by an $N \times L$ matrix $X = \{x_{n\ell}\} \in \mathbb{R}^{N\times L}$. N is the number of mixtures, M is the number of sources, and L is the number of observations. In the context of NMR spectroscopy un-mixing, N is the number of observed spectra, M is the number of compounds and L is the number of points of the spectra. As an example, the number $x_{n\ell}$ is the ℓ-th sample of mixture n, i.e. its value at frequency l.

Among BSS problems, the simplest instance originates from the *Linear Instantaneous Mixture* (LIM) model, where the observed mixtures are linear combinations of the sources. The mixing process is then expressed mathematically as

$$X = AS + B \approx AS\,, \tag{1}$$

more explicitly

$$x_{n\ell} = \sum_{m=1}^{M} a_{nm} s_{m\ell} + b_{n\ell}\,, \qquad n = 1,\dots N\,,\ \ell = 1\dots L\,, \tag{2}$$

where $B = \{b_{n\ell}\} \in \mathbb{R}^{N\times L}$ is some residual noise, and $A = \{a_{nm}\} \in \mathbb{R}^{N\times M}$ is called the mixing matrix. The BSS problem is to identify jointly the mixing matrix A and the source matrix S from the sole

observation matrix X.

Different assumptions or models have led to different identification algorithms, among which we may mention statistics based approaches such as ICA and SOBI, or non-negative matrix factorizations (NMF), which will constitute our approach. Thorough descriptions can be found in reference textbooks[16,17], and we refer to[4] for a review of applications to NMR spectroscopy.

2.2 The LIM model for 2D spectra

In the case of 2D data such as the HSQC discussed below, observations X and pure signals S are not matrix-shaped any more, but take the form of three-way arrays: $X \in \mathbb{R}^{N\times L_1\times L_2}$ and $S \in \mathbb{R}^{M\times L_1\times L_2}$. The LIM model can still formally be written as in (1), provided the matrix×tensor product is suitably defined, in the sense

$$x_{n\ell_1\ell_2} = \sum_{m=1}^{M} a_{nm} s_{m\ell_1\ell_2} \ , \quad n = 1,\dots N, \ell_1 = 1\dots L_1, \ell_2 = 1\dots L_2 \ , \quad (3)$$

where ℓ_1 and ℓ_2 label the two spectral dimensions.

By re-organizing the ℓ_1 and ℓ_2 spectral indices into a single one (of length L_1L_2), i.e. transforming three-way arrays into matrices, one can be back to model (1) (in significantly higher dimension). We call this approach *data matricization*. However, matricization is not always suitable, as this reshaping procedure breaks the 2D structure, which is used by some algorithms.

2.3 Indeterminacies

Quite obviously, the solution of such a general problem is not unique, as for any solution (A,S) and any invertible $M\times M$ matrix Λ, one can also write $X = AS = A'S'$ where $A' = A\Lambda$ and $S' = \Lambda^{-1}S$, which produces infinitely many other solutions. Therefore, additional assumptions are necessary to solve the problem. Among these indeterminacies, the following two play a special role (and correspond to two specific types of matrices Λ):

- Scale indeterminacy (Λ diagonal): sources can only be identified up to a constant factor (in other words, multiplying a row of S by a constant and dividing the corresponding column of A by the same constant do not change X).

- Order indeterminacy (Λ a permutation matrix): estimated sources are not ordered, therefore comparison of estimated sources with reference sources has to be preceded by an ordering step.

Two solutions (A,S) and (A',S') that only differ by these two transformations are generally considered equivalent.

To overcome indeterminacy problems, additional assumptions have to be made, either on sources, mixing matrix or both. The non-negativity assumptions made in NMF approaches described below turn out to resolve a part of these problems, nevertheless scale and order indeterminacies remain. These will have to be accounted for in the NMR un-mixing algorithms, as we shall see later.

2.4 Non-negative matrix factorization (NMF)

Non-negative matrix factorization techniques address situations where both source coefficients $s_{m\ell}$ and mixing matrix coefficients a_{nm} are non-negative. In NMR spectroscopy un-mixing problems, such assumptions are relevant, since mixing matrix coefficients represent concentrations, and source coefficients represent spectrum values.

Many approaches to NMF have been proposed since early works of Paatero & Tapper[19] and Lee & Seung[23,24]. Most of them are based on so-called *variational formulations*, where numerical algorithms are used to minimize some objective function, which involves a *data fidelity term* (which forces the product AS to be close to the data matrix X) and possibly additional terms that may encode prior information on the mixing matrix and/or the source matrix. The mathematical formulation takes the form

$$\min_{A,S} F(X|A,S) \ , \qquad \text{under constraints} \quad A \geq 0,\ S \geq 0 \ . \quad (4)$$

Here F is the objective function, that depends on the data matrix X and the unknown matrices A and S. Moreover, the non-negativity constraints are imposed entry-wise, i.e. all matrix elements have to be non-negative.

The most classical choices for the objective function are penalized versions of the standard quadratic objective function

$$F(X|A,S) = \frac{1}{2}\|X - AS\|_F^2 + f_A(A) + f_S(S) \ , \quad (5)$$

where the first term, called squared Frobenius norm, is simply the sum of squares of the matrix $X-AS$, and f_A and f_S are *regularizations* that can encode prior information on A and/or S. Standard choices involve the so-called ℓ^p norms denoted by $\|\cdot\|_p$ (where $p \geq 0$), for example

$$f_S(S) = \lambda \|S\|_p^p = \lambda \sum_{m,\ell} |s_{m\ell}|^p \ . \quad (6)$$

with $\lambda \geq 0$ the regularization parameter. The case $p = 2$ is the widely used Tikhonov regularization. We shall rather use $p = 1$, which tends to enforce sparse solutions.

Alternative choices for data fidelity terms include the Kullback-Leibler divergence $F_{KL}(X|A,S)$, used for example in[23,24], which is (as well as the squared Frobenius norm) a special case of the family of so-called β-divergences[25].

2.5 Evaluation criteria

Un-mixing results on real data have to be evaluated by experts. However, performance evaluation for separation algorithms can be assessed using numerical simulations, in which cases objective assessment is possible. We briefly describe here some evaluation metrics that are routinely used in BSS problems.

In numerical simulations, one starts with pre-defined source matrix S and mixing matrix A, the mixture $X = AS + B$ can then be formed (where either $B = 0$, i.e. no noise, or B is a matrix containing random Gaussian white noise with prescribed variance). Separation algorithms yield estimates, denoted by $\hat{S}$ and $\hat{A}$, to be compared with ground truth S and A.

To assess the quality of the mixing matrix estimate, a rele-

vant quantity is the matrix product $G = A^{\dagger}\hat{A}$, with $A^{\dagger}$ the pseudo-inverse of A. G equals the identity matrix when the estimation is perfect, i.e. $\hat{A} = A$. The departure from that perfect situation may be quantified by the Amari index

$$I = \frac{1}{2M(M-1)} \sum_{m=1}^{M} \left[\frac{\sum_{m'=1}^{M} |g_{mm'}|}{\max_{m'} |g_{mm'}|} + \frac{\sum_{m'=1}^{M} |g_{m'm}|}{\max_{m'} |g_{m'm}|} - 2 \right] , \qquad (7)$$

which ranges between 0 and 1, and is equal to 0 when $\hat{A} = A$ and 1 in case $\hat{A}$ and A are maximally different. It is worth noticing that the Amari index I is insensitive to source ordering and normalization, so that no action is required to compensate for the order and normalization indeterminacies described in subsection 2.3.

To assess the quality of estimated sources, we rely on indices implemented in the software toolbox BSSEVAL[26]. In a nutshell, the estimation error $S - \hat{S}$ is split as a sum of several terms, that are used to evaluate various types of error. Of particular interest to us here are the *Signal to Distortion Ratio* (SDR), which provides a global measure of the distortion introduced by mixing and separation, and the *Signal to Interference Ratio* (SIR), which provides a quantitative evaluation of crossover terms after separation (in our case, peaks from a given source that could be completely or partially found in the estimate of another source). Both indices are graded on a logarithmic scale, and expressed in dB, as the traditional SNR (Signal to Noise Ratio). Contrary to the Amari index, these ratios are sensitive to order and normalization, therefore suitable re-ordering and normalization steps are mandatory prior to computing SDR and SIR.

2.6 Changing the representation domain, wavelets

In the above approaches, mixtures and sources are represented by point values, respectively $x_{n\ell}$ and $s_{m\ell}$. The objective functions that are optimized in NMF algorithms are separable, in the sense that for given mixing matrix A, columns of S are processed independently of each other, so that possible correlations in the spectral domain (represented by index ℓ) are not exploited. It is however possible to describe the spectral domain using a different representation, based upon an expansion on a set of L-dimensional vectors, that form a basis of the L-dimensional space, and to introduce a regularization on the corresponding coefficients rather than the source matrix itself. Denoting by $\{\psi^{(k)}, k = 1, \dots L\}$ these vectors, and concatenating them in a square matrix denoted by Ψ (columns of Ψ are the vectors $\psi^{(k)}$), it may be shown that the coefficients of the source matrix S in this basis are given by the matrix $\Gamma = \Psi^T S$ (where ^{T} stands for matrix transposition).

The variational formulations described above can be adapted to this new setting, by introducing adapted objective functions of the generic form

$$F(X|A,S) = \frac{1}{2} \|X - AS\|_F^2 + f_A(A) + f_\Gamma(\Psi^T S) , \qquad (8)$$

where f_Γ is a suitable penalty function. As before, we will choose an ℓ_1 penalization with a regularization parameter $\lambda \geq 0$, i.e. $f_\Gamma(\Gamma) = \lambda \sum_{m,\ell} |\gamma_{m,\ell}|$, which will have the effect of promoting sparse coefficient matrices, i.e. matrices having a very large number of coefficients equal or close to zero.

Among possible bases, we will use here bases of orthonormal wavelets[27]. The use of wavelets for representing NMR spectra has been advocated by several authors[28,29], the main argument being the ability of wavelet expansions to compress signals[30].

3 Algorithms

3.1 Generic algorithm

We describe in this section a generic algorithmic approach to the NMR un-mixing problem. All algorithms to be described here aim at solving Problem (4), i.e. a joint minimization problem with respect to source matrix S and mixing matrix A. This problem is addressed through an alternate optimization algorithm, i.e. we optimize alternately with respect to A and S. Various instances of the algorithm are proposed, depending on the choice of objective function, optimization strategy and source representation domain (spectral domain or wavelet domain). In all cases the generic structure is given in Algorithm 1, and differs only by the update rules for A and S, which we will generically denote by

$$\mathsf{Upd}_A : \quad (A,S) \to \mathsf{Upd}_A(A,S) \in \mathbb{R}^{N \times M} ,$$

$$\mathsf{Upd}_S : \quad (A,S) \to \mathsf{Upd}_S(A,S) \in \mathbb{R}^{M \times L} .$$

Data: X (data matrix); iter_max; ε; A_{init}; S_{init}; $\mathsf{Crit} = +\infty$;
Result: Non-negative matrix factors A and S
Initialization: $A^{(0)} = A_{\mathsf{init}}$, $k = 0$, $S^{(0)} = S_{\mathsf{init}}$;
while $k \leq$ iter_max **and** $\mathsf{Crit} > \varepsilon$ **do**
  Update of A: $A^{(k+1)} = \mathsf{Upd}_A(A^{(k)}, S^{(k)})$;
  Update of S: $S^{(k+1)} = \mathsf{Upd}_S(A^{(k+1)}, S^{(k)})$;
  Optional: normalization of A and/or S;
  Evaluation of stopping criterion $\mathsf{Crit}(k+1)$;
  Evaluation of the objective function $F(X|A^{(k+1)}, S^{(k+1)})$;
  $k = k + 1$;
end

Algorithm 1: Generic structure of alternate optimization algorithm for non-negative matrix factorization (starting by updating A is an arbitrary choice).

This generic algorithm requires additional ingredients/options, some of which are listed below

1. Initialization: initial source and mixing matrices S_{init} and A_{init} are necessary to start the iterations (some algorithms require only one of these). Usual choices include random initialization, or deterministic ones (based upon SVD, ICA or other classical methods).

2. Stopping criterion: the algorithm stops when a prescribed maximal number of iterations is reached, or preferably when some precision criterion reaches a small enough value. Possible choices include the absolute value of the objective function's gradient, or normalized norms of differences between two consecutive iterates of A and S.

3. To account for normalization indeterminacy, it is possible to normalize rows of S and columns of A at each iteration, so as to enforce a certain normalization property, without chang-

ing the product AS. This makes sense only when the objective function is itself invariant under renormalization.

4. Some algorithms require non-negative data. In such cases it is necessary to project the data matrix X accordingly, i.e. set to zero all negative matrix elements $x_{n\ell}$.

3.2 Projected alternate least squares (PALS)

The objective function is here the most classical one, i.e. the sum of squares of matrix coefficients (termed simply the squared Frobenius norm) of the discrepancy $X - AS$ between data X and the LIM model AS, and reads

$$F(X|A,S) = \frac{1}{2}\|X - AS\|_F^2 \ . \tag{9}$$

The corresponding update rules are given by

$$\mathsf{Upd}_A(A,S) = \Pi_+\left[(AS-X)S^T\right] \ , \quad \mathsf{Upd}_S(A,S) = \Pi_+\left[A^T(AS-X)\right] \ ,$$

where Π_+ denotes the operator that sets to zero all negative matrix coefficients of its argument.

3.3 Soft thresholded projected alternate least squares (STALS)

To enforce sparsity of the sources, a common practice is to add to the above quadratic objective function an ℓ_1 penalization, namely the sum of absolute values of source terms, denoted by $\|S\|_1$

$$F(X|A,S) = \frac{1}{2}\|X - AS\|_F^2 + \lambda\|S\|_1 \ , \tag{10}$$

where λ is a positive constant that tunes the strength of the penalty. A commonly used approach for solving this problem is to replace the projection Π_+ (of S) onto non-negatives with the non-negative soft thresholding operator $\mathbb{S}_\lambda^+$, which sets to zero all matrix coefficients smaller than the threshold λ (including negative values). The update rules become

$$\mathsf{Upd}_A(A,S) = \Pi_+\left[(AS-X)S^T\right] \ , \quad \mathsf{Upd}_S(A,S) = \mathbb{S}_\lambda^+\left[A^T(AS-X)\right] \ .$$

Notice that PALS coincides with STALS in the case $\lambda = 0$.

3.4 Proximal alternating linearized minimization (PALM) and pre-conditioned version (BC-VMFB)

The objective here is to deal with the cost function defined previously in (10). The idea is to propose an algorithm which intertwines the minimization of the quadratic part and the regularization part. This can be done by using either the PALM (Proximal alternating linearized minimization) algorithm[31] or its preconditioned version named the BC-VMFB (Block-Coordinate Variable Metric Forward-Backward) algorithm[32]. Both are based on a projected gradient descent algorithm and an optional preconditioning step that allows increasing the convergence speed. The update rules in this case are defined by

$$\mathsf{Upd}_A(A,S) = \Pi_+\left[A - \gamma(AS-X)S^T\right] \ ,$$

$$\mathsf{Upd}_S(A,S) = \mathbb{S}_{\lambda/\gamma}^+\left[S - \gamma A^T(AS-X)\right] \ .$$

where γ stands for the gradient descent stepsize.

3.5 Wavelet-based PALM and BC-VMFB

These algorithms address the case where sparsity is imposed on the wavelet coefficients of the spectra rather than the spectra themselves. The considered objective function is a special case of (8), namely

$$F(X|A,S) = \frac{1}{2}\|X - AS\|_F^2 + \lambda\|\Psi^T S\|_1 \ ,$$

$\|\Gamma\|_1$ being the sum of absolute values of coefficients $\gamma_{m\ell}$. The PALM and BC-VMFB algorithms can be adapted to this new setting. The wavelet-based PALM and BC-VMFB algorithms thus reduce to PALM and BC-VMFB algorithms, except that the thresholding operation is done on wavelet coefficients rather than spectrum coefficients. The update rules become

$$\mathsf{Upd}_A(A,S) = \Pi_+\left[A - \gamma(AS-X)S^T\right] \ ,$$

$$\mathsf{Upd}_S(A,S) = \Pi_+\left[\Psi\left(\mathbb{S}_{\lambda/\gamma}\left[\Psi^T\left(S - \gamma A^T(AS-X)\right)\right]\right)\right]$$

where $\mathbb{S}_\lambda$ sets to zeros only the values whose absolute value is smaller than λ.

3.6 Processing 2D spectra

In the case of 2D spectra, most algorithms described above can still be used on matricized data (see section 2.2). However, wavelet-based algorithms are not compatible with data matricization, as the latter breaks the 2D structure that is exploited by 2D wavelets. Nevertheless, the algorithms given in section 3.5 can still be used, Ψ now being a two-dimensional wavelet transform, similar to the transform used in the JPEG2000 image compression standard, which has been shown to be extremely good at compressing 2D NMR spectra[30]. The same procedure would apply to higher dimensional spectra as well.

4 Experimental results

We present and discuss in this section numerical results obtained using the algorithms described above, on real and simulated NMR mixtures.

Throughout this section, we term *real mixtures* the spectra of the solutions that have been acquired by NMR spectroscopy. By *simulated mixtures* we mean spectra that have been computed using the mathematical LIM model (1), using the spectra of pure compounds measured by NMR spectroscopy in S, and the concentrations that have been used to produce the solutions in A.

We describe the datasets before discussing results.

4.1 Data acquisition

4.1.1 Data description

Four commercially available solutions of terpenes were purchased from Sigma-Aldrich (Merck), Saint Quentin Fallavier, France: (R)−(+)− Limonene, Nerol, α-terpinolene, (−)−trans-Caryophyllene. The Initially pure compounds were dissolved in 600 μL of $CDCl_3$ at respective concentrations of 181 mM, 36.5 mM, 26.6 mM and 43.7 mM and then transferred to 5 mm NMR tubes which were sealed to prevent loss of solvent at operating temperatures. Samples were then stored at −4° C until the NMR characterization. Five synthetic mixtures of the four terpenes were prepared varying the concentrations of each compound as reported in Table 1.

	Limonene	Nerol	α−Terpinolene	β-Caryophyllene
Solution 1	23.3 mM	26 mM	8.78 mM	10.87 mM
Solution 2	17.1 mM	11.93 mM	15.5 mM	15 mM
Solution 3	9.05 mM	14.23 mM	18.89 mM	4.67 mM
Solution 4	20.99 mM	6.86 mM	13.54 mM	11.96 mM
Solution 5	4.88 mM	9.01 mM	10.81 mM	13.15 mM

Table 1 Concentrations of each component of the proposed terpenes.

4.1.2 NMR Spectroscopy

All experiments were performed on a Bruker Avance III 600 MHz spectrometer equipped with a triple resonance high-resolution probe, using a SampleJet with pre-cooling rack refrigerated at 4° C. A standard 1D pulse sequence is applied to each sample: zg ^{1}H 1D (90-Taq), with a spectral width of 6600 Hz, 96 scans, and relaxation delay of 10 s. The 90° pulse length was automatically calibrated for each sample at around 9.5 μs. Subsequently, the spectra were pre-processed: phased and baseline corrected automatically and referenced to the $CDCl_3$ at δ 7,27 ppm using the inbuilt software TOPSPIN 3.5 version (Bruker BioSpin, Germany).

2D ^{1}H-^{13}C HSQC spectra were recorded with phase sensitive sequence "hsqcetgpsi" using Echo/Antiecho-TPPI gradient selection with an INEPT delay adjusted to one-bond ^{1}H-^{13}C coupling constant of 145Hz. 256 t1 points were acquired for the indirect dimension and 32 scans for each point with TD2 = 4096. The acquisition time for the direct period was 142ms and the resolution of 7.045Hz. For the indirect period, the acquisition time was 4.2ms and the resolution of 235.80 Hz. No linear prediction was used for these experiments. Both ^{1}H and ^{13}C axes were calibrated using the chloroform peak at 7.27 and 77.2 ppm, respectively.

The spectra of the pure compounds are presented in the stacking plot, Fig. 1. The studied mixtures are well adapted to evaluate source separation algorithms. Indeed, terpenes are natural molecules present in plants and having highly crowded spectra between 1.5 and 2.5 ppm.

4.2 Algorithm validation on 1D simulated mixtures

We first report on simulation results. The goal is to validate the algorithms in a situation where a ground truth is available, in the framework of the LIM model described in section 2.1. The ground truth is provided by 1) the spectra of pure compounds, that were obtained as described in section 4.1.2, and are collected

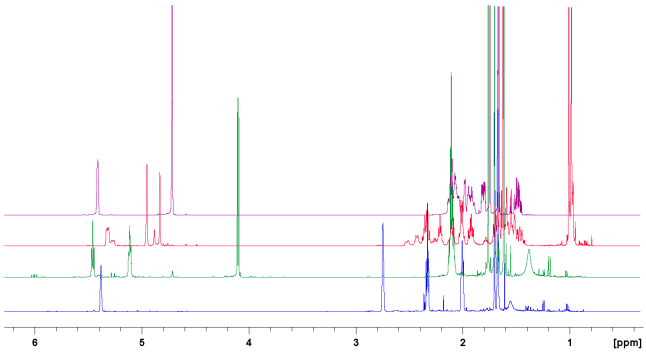

Fig. 1 ^{1}H NMR: stacked plot of the spectra of the pure compounds. Top to bottom: Limonene, β-Caryophyllene, Nerol, α−Terpinolene.

in a source matrix S, and 2) the concentrations given in Table 1, organized in a 5×4 mixing matrix A.

From these, simulated mixtures of the form

$$X_m = AS + B$$

were generated according to the LIM model (1), involving the linear mixtures given by the matrix product AS, and a zero mean Gaussian white noise B. The standard deviation σ of the latter was set to the standard deviation of experimental noise, estimated in a signal-free segment of the real mixtures X.

The six algorithms above (PALS, STALS, PALM, BC-VMFB and PALM, BC-VMFB using wavelets) were run on the simulated dataset. For the stopping criterion, we used the relative size of the objective function update from one iteration to the next, i.e.

$$\mathsf{Crit}(k) = \left| \frac{F(X|A^{(k+1)}, S^{(k+1)}) - F(X|A^{(k)}, S^{(k)})}{F(X|A^{(k)}, S^{(k)})} \right| ,$$

where $A^{(k)}$ and $S^{(k)}$ are the estimates at iteration k. The algorithms also require an initial estimate. Several choices are possible, we used here estimates obtained using the JADE ICA algorithm[33]. More precisely, running JADE on the mixture matrix yields an estimate for the un-mixing matrix, denoted by D, so that DX provides an estimate of the sources. Also, the pseudo-inverse $D^\dagger$ yields an estimate for the mixing matrix. These estimates taking both positive and negative values, we use as initialization the absolute values $S_{\mathsf{init}} = |DX|$ and $A_{\mathsf{init}} = |D^\dagger|$.

JADE only requires the number of sources to be estimated. A PCA on the observation matrix shows that only 4 of the 5 corresponding latent variables are significant, which suggests to set to 4 the number of sources to estimate (which turns out to be the actual number of terpenes present in the solutions).

STALS, PALM and BC-VMFB require choosing a thresholding parameter λ, for which five choices were tested, namely 0.01σ, 0.1σ, σ, 10σ and 100σ. Similar choices are made for the wavelet-based versions of PALM and BC-VMFB.

The un-mixing results on simulated data are globally very good for most (if not all) algorithms. The best results seem to be obtained by the STALS and PALS approaches, with various values of the thresholding parameter λ (we recall that PALS is the special case of STALS with $\lambda = 0$). The best estimate for the concentra-

	Limonene	Nerol	$\alpha-$Terpinolene	β-Caryophyllene
Solution 1	0.12 %	−0.67 %	3.82 %	0.09 %
Solution 2	−0.29 %	0.72 %	0.07 %	−0.07 %
Solution 3	1.29 %	−0.74 %	−1.93 %	0.52 %
Solution 4	−1.31 %	2.06 %	−0.38 %	−0.04 %
Solution 5	3.68 %	0.61 %	0.64 %	−0.15 %

Table 2 ^{1}H NMR spectra (simulated case): relative errors in the estimated concentrations (in %) using STALS with $\lambda = 10\sigma$. The corresponding Amari index equals 0.008.

tions (mixing matrix A) was obtained by STALS with threshold value set to 10σ (σ being the standard deviation of the noise). This corresponds to the relative errors reported in Table 2. This relative error has been computed as follows: let $\hat{A}$ be the mixing matrix estimate of A. Then the relative error (in %) is given by $(\hat{A}-A)/A*100$ where the quotient is computed element-wise.

We provide in Table 3 values of evaluation indices for all the algorithms, λ being fixed to 10σ. For STALS, the SIR and SDR values globally range between 30 dB and 55 dB, which is generally considered very good.

	Algorithms					
	PALS	STALS	PALM	BC-VMFB	PALM wav	BC-VMFB wav
Amari	.019	.008	.022	.025	.036	.031
SIR (1)	26.7	52.4	22	24.4	20.7	22.2
SIR (2)	32.5	31.2	29.8	28.5	30.1	32.5
SIR (3)	19.5	45.7	41.6	25.2	24.3	23.5
SIR (4)	47.6	29.3	19.9	24	21.4	21.7
SIR (m)	31.6	39.6	28.3	25.5	24.1	25
SDR (1)	26.7	51.4	21.7	24.4	20.4	21.3
SDR (2)	32.5	31.2	29.6	28.5	29.7	29.5
SDR (3)	19.5	44.9	40.8	25.2	24	22.8
SDR (4)	47.4	28.6	19.6	23.9	21	20.6
SDR (m)	31.5	39	27.9	25.5	23.8	23.6

Table 3 ^{1}H NMR spectra (simulated case): numerical results, using all algorithms, for $\lambda = 10\sigma$ (numbers between parentheses indicate the source number and m stands for the mean value).

The separation quality is exemplified in Fig. 2 where we can see good correspondences for the limonene spectrum while underlining some extra peaks coming from another source. As could be expected, the spurious peaks are significantly reduced for high values of the thresholding parameter λ.

A closer look at the results shows that the best results are obtained with STALS with $\lambda = 10\sigma$ for two sources (1 and 3), and PALS for the other two. This indicates that optimal thresholding parameter may be source dependent. This is further confirmed when looking at results obtained with STALS using different regularization parameters λ (see Table 4).

Note that the other algorithms also yielded good un-mixing results on simulated data.

4.3 Real 1D mixtures

The same algorithms as above were run on the real mixtures, the results were at first glance very disappointing: the algorithms failed to separate the pure compounds spectra from mixture spectra. A closer analysis revealed that the main reason was the pres-

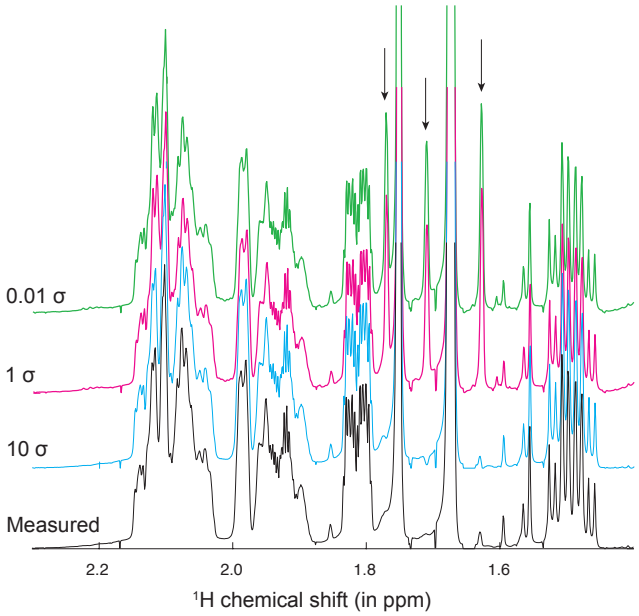

Fig. 2 ^{1}H NMR spectra (simulated case): measured spectrum of limonene and estimated spectra of the same compound with STALS for $\lambda = 0.01\sigma$, σ or 10σ. The three arrows show the presence of 3 extra peaks that are not present when $\lambda = 10\sigma$. These extra peaks are residual signals from nerol.

	λ					
	0	$\sigma/100$	$\sigma/10$	σ	10σ	100σ
Amari	0.019	0.019	0.019	0.019	0.008	0.018
SIR (1)	26.7	26.7	26.7	26.8	52.4	42.5
SIR (2)	32.5	32.3	32.2	31.9	31.2	27.6
SIR (3)	19.5	19.5	19.5	19.6	45.7	39.3
SIR (4)	47.6	45.9	44.6	42.4	29.3	18.3
SDR (1)	26.7	26.7	26.7	26.8	51.4	39.7
SDR (2)	32.5	32.3	32.2	31.9	31.2	27.3
SDR (3)	19.5	19.5	19.5	19.6	44.9	35.4
SDR (4)	47.4	45.8	44.5	42.3	28.6	17.1

Table 4 ^{1}H NMR spectra (simulated case): numerical results for the STALS algorithm, with various values of the thresholding parameter λ (numbers between parentheses indicate the source number and m stands for the mean value).

	Sol. 1	Sol. 2	Sol. 3	Sol. 4	Sol. 5
SIR	16.12	12.96	18.89	19.95	17.15
SDR	10.86	8.84	12.90	9.76	9.72

Table 5 ^{1}H NMR spectra: SIR and SDR indices (in dB) comparing measured and simulated mixture spectra.

	Limonene	Nerol	$\alpha-$Terpinolene	β-Caryophyllene
SIR	13.1	15	20.2	14.8
SDR	9.7	9.9	4.8	6.8

Table 6 ^{1}H NMR spectra (real case): SIR and SDR indices (in dB) comparing true and estimated source spectra with the BC-VMFB algorithm with $\lambda = \sigma$.

	Limonene	Nerol	$\alpha-$Terpinolene	β-Caryophyllene
Solution 1	12.34 %	25.61 %	-9.03 %	-4.33 %
Solution 2	4.05 %	-6.68 %	1.11 %	26.27 %
Solution 3	-47.70 %	1.91 %	-0.53 %	-32.53 %
Solution 4	16.36 %	-69.78 %	1.54 %	-11.15 %
Solution 5	-55.03 %	-14.94 %	4.74 %	-4.70 %

Table 7 ^{1}H NMR spectra (real case): relative errors in the estimated concentrations (in %) using BC-VMFB with $\lambda = \sigma$. The corresponding Amari index equals 0.081.

ence of spurious shifts between pure and mixed spectra, peaks of these two families of spectra were not correctly aligned. The algorithms under consideration being extremely sensitive to such issues, obtaining reliable un-mixing results without proper alignment is hopeless. We display in Figure 3 the spectra of real mixtures (i.e. the five rows of the X matrix) and simulated mixtures (the rows of $X_m = AS$). The first column displays the complete spectra, while the second and third columns zoom in specific regions. As can be seen there, there is a clear shift between peaks, that may even be significantly location dependent (third column). This could be the result of a slight variation of pH due to the decomposition of $CDCl_3$ or, of a difference of ionic strength between samples. Molecular interactions between the individual compounds could also contribute to the observed chemical shift variations.

To overcome this effect, all spectra were aligned using a standard tool. There exist several approaches of the peak alignment problem[34], based upon various approaches such as correlation analysis, least squares, dynamic time warping, parametric time warping and several others. Here the online tool *NMRProcFlow* was used to provide re-aligned spectra, on which un-mixing algorithms could be suitably tested. This alignment method is based on a least squares algorithm for which a reference spectrum (here, the average spectrum) was calculated. Each region is re-aligned by shifting it to match the reference spectrum. For the sake of further comparison, mixture spectra were aligned together with pure spectra, which will not be possible in general situations where the latter will not be available.

Re-aligned spectra are displayed in Figure 4, which again shows complete spectra and zooms in the same regions as in Figure 3. The procedure allowed to fix the alignment problems quite successfully, as exemplified by columns 2 and 3. However, the latter also exhibit slight amplitude modulations, which suggests that the departure of real mixtures from the mathematical model includes more than peak shifts. To get a quantitative assessment of the adequacy of the model, we computed values of SIR and SDR indices, that measure discrepancies between X and X_m. Results are given in Table 5. SIR values are fairly acceptable, which tends to indicate that peaks are located at the right place, SDR values are significantly lower, which we interpret mainly as a consequence of amplitude modulations. This suggests that a more adequate model for describing real mixtures should involve both shift and amplitude modulation, in combination with the linear instantaneous mixing model (1).

The un-mixing algorithms were run on aligned real mixture spectra. The resulting un-mixed spectra for the BC-VMFB algorithm with $\lambda = \sigma$ are displayed in Figure 5, and the corresponding evaluation indices are given in Table 6.

The β-caryophyllene was extracted with great accuracy while the limonene presented 4 extra peaks that belonged to the nerol. Although the major signals from the nerol spectrum were well recovered, an extra signal coming from β-carophyllene was also observed at about 1ppm. The worst result was obtained for the α-terpinolene spectrum where signals from this compound were mixed with some signals from nerol and limonene. However, despite the presence of these artifacts, the major signals of each source spectrum were accurately found, allowing the identification of the corresponding molecule without ambiguity.

We reported in Table 7 the relative errors on the mixing matrix estimate.

4.4 Results on HSQC mixtures

The algorithms were also tested on HSQC data, we report here on the results.

4.4.1 Numerically simulated HSQC mixtures

The first tests were on simulated mixtures, i.e. mixtures generated using the mathematical formulas described in section 2.2, using the measured source matrix and a mixing matrix corresponding to the true concentrations in solutions. We provide in Table 8 a summary of the results obtained using the BSS algorithms, with thresholding parameter set to $\lambda = 100\sigma$, σ being the standard deviation of the noise, measured in a signal-free part of the spectra. The results are globally very good, in particular SIR values (remember that SIR provides a measure of the cross-talk between sources, in other words the presence in an estimated source of spurious components originating from the others), which are quite high. SDR values are lower, which indicates that distortions are present in the estimated sources. Even though it is not easy to draw clear conclusions, the best results seem to be obtained with the STALS algorithm. Notice that the wavelet-based BC-VMFB algorithm yields quite good results as well. The latter algorithm turns out to be the most effective on real mixtures. We do not display here graphical comparison of real and estimated HSQC spectra, as we prefer to focus on real mixtures (see below). However, let us mention that the results are visually excellent, the fingerprint of each terpene is perfectly recovered.

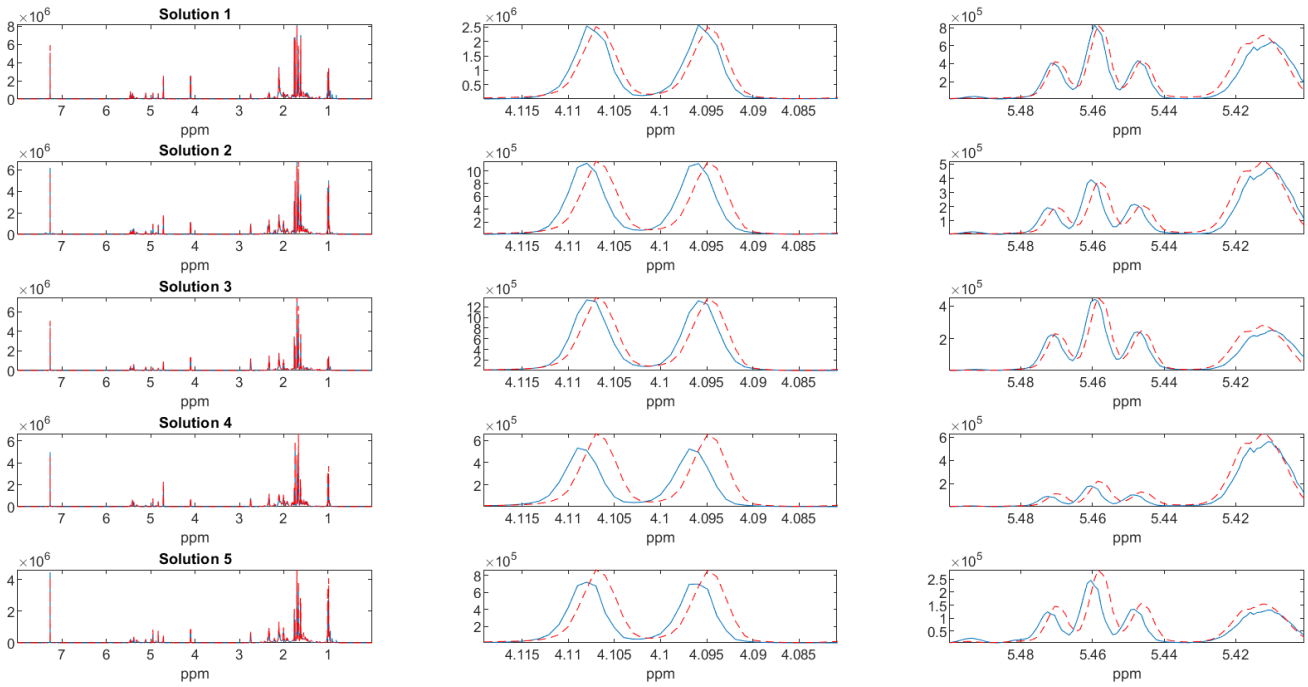

Fig. 3 ^{1}H NMR spectra: correspondence of X (blue) and X_m (red) before alignment pre-processing (whole spectra (left), expanded regions (middle,right)).

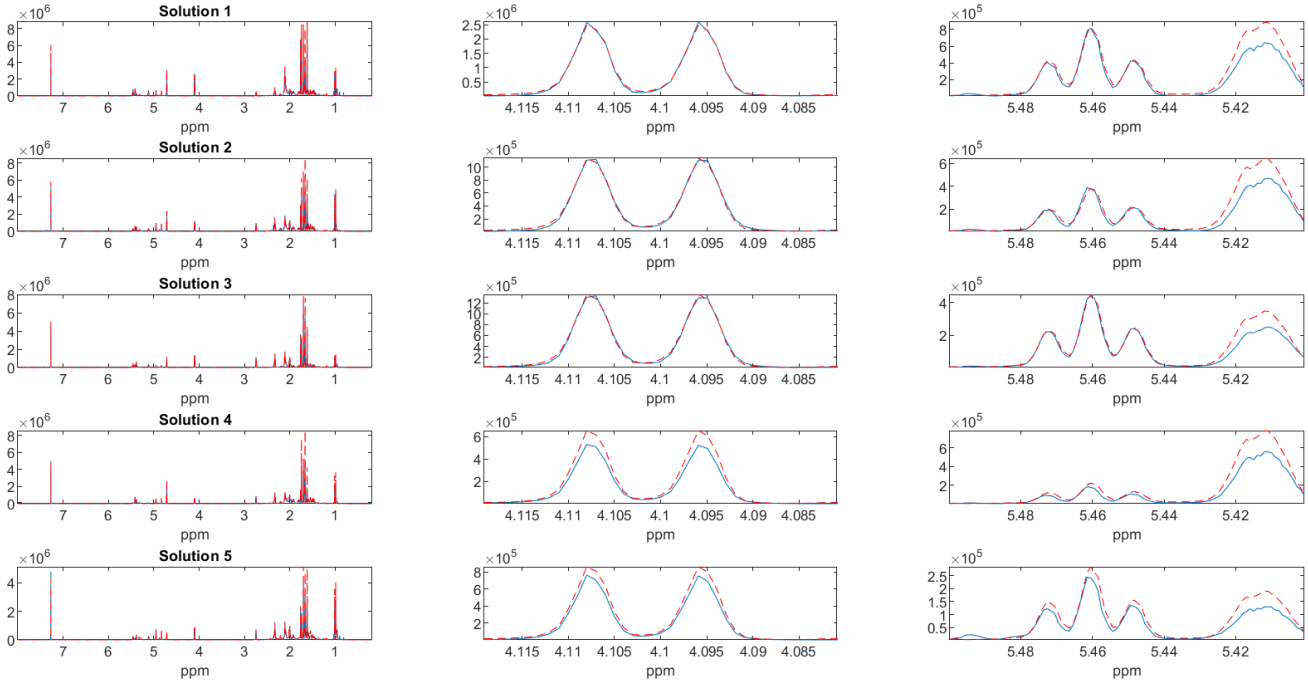

Fig. 4 ^{1}H NMR spectra: correspondence of X (blue) and X_m (red) after alignment pre-processing (whole spectra (left), expanded regions (middle,right)).

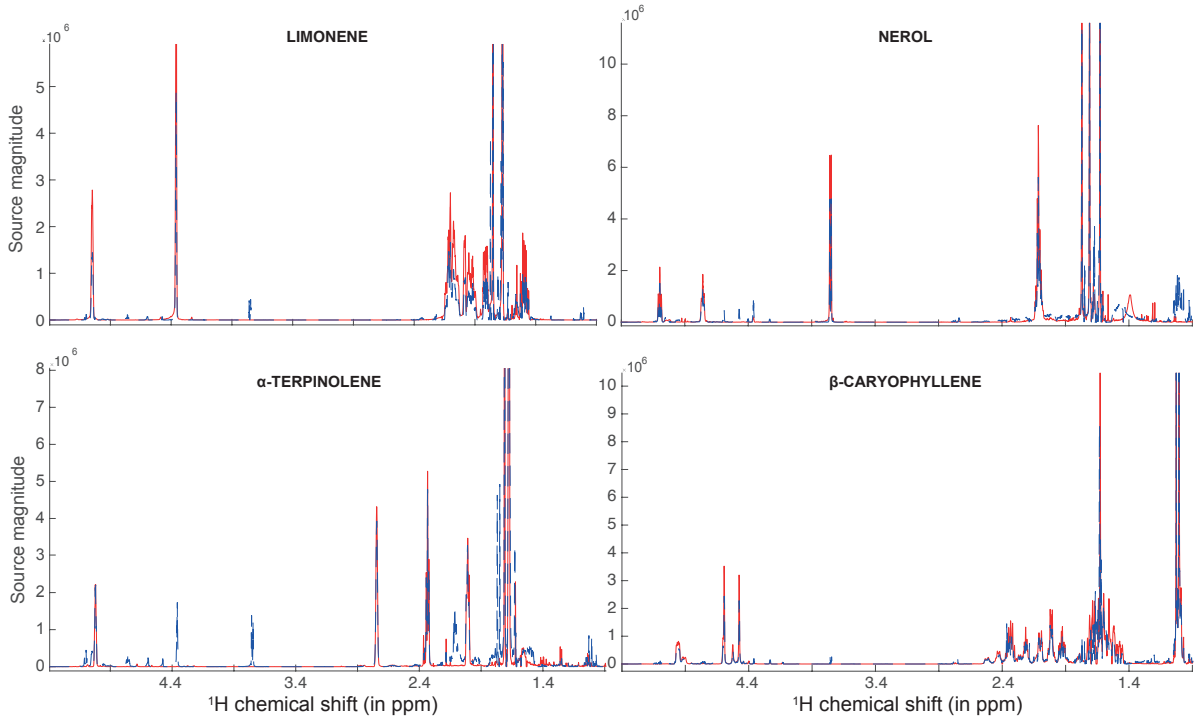

Fig. 5 ^{1}H NMR spectra (real case): spectra of the 4 sources estimated using BC-VMFB with $\lambda = \sigma$ (blue) compared to the spectra of real sources (red).

	Algorithms					
	PALS	STALS	PALM	BC-VMFB	PALM wav	BC-VMFB wav
Amari	.135	.012	.015	.11	.081	.016
SIR (1)	21.6	50.1	39.6	39.8	20.9	42.8
SIR (2)	12.4	39.1	31.7	10.3	21.8	34.6
SIR (3)	22	29	28	9.7	31.5	31.4
SIR (4)	8.3	41.4	45.4	20.8	34.2	28.6
SIR (m)	16.1	39.9	36.2	20.2	27.1	34.4
SDR (1)	20.5	25.9	25.7	24.6	19.8	24.9
SDR (2)	11.6	16	17.2	9.3	16.4	16.4
SDR (3)	15.1	13.6	13.5	8.6	14.1	13.2
SDR (4)	7.1	9.9	10.1	10.5	10.4	9.8
SDR (m)	13.6	16.3	16.6	13.3	15.2	16.1

Table 8 2D ^{1}H-^{13}C HSQC spectra (simulated case): numerical results for $\lambda = 100\sigma$ (numbers between parentheses indicate the source number and m stands for the mean value).

4.4.2 Real HSQC mixtures

The algorithms were also tested on real mixtures (the same solutions as the ones reported in the section devoted to ^{1}H NMR). The results in this case too are quite good. We do not report on the performances of all algorithms, and focus on the best performing one, that turns out to be the wavelet-based BC-VMFB (mentioned above), with a thresholding parameter set to $\lambda = 10\sigma$. The corresponding SIR and SDR indices are provided in Table 9, the Amari index equals 0.042, meaning that concentrations have been correctly estimated (see Table 10 for the relative errors on the estimated concentrations). As can be seen, SIR values are again very good (around 32dB in average), which indicates that even in crowded regions of the spectra, no significant cross-talk between sources is observed. SDR values remained significantly weaker, meaning that estimated spectra were significantly perturbed. As in the 1D situation, this may be interpreted as a departure from the 2D LIM model. Again as in the 1D case, the visual inspection of the estimated sources versus the true one in Figure 6 shows that the results are of sufficient quality to identify the four terpenes in these solutions.

Remark. As in the 1D case, the adequacy of the LIM model for HSQC spectra can be assessed by looking at quality indices. SIR and SDR indices computed from measured and simulated mixture spectra are provided in Table 11. The conclusions that can be drawn are essentially similar to the previous ones. SIR values are acceptable, at least for solutions 1-3, and lower for solutions

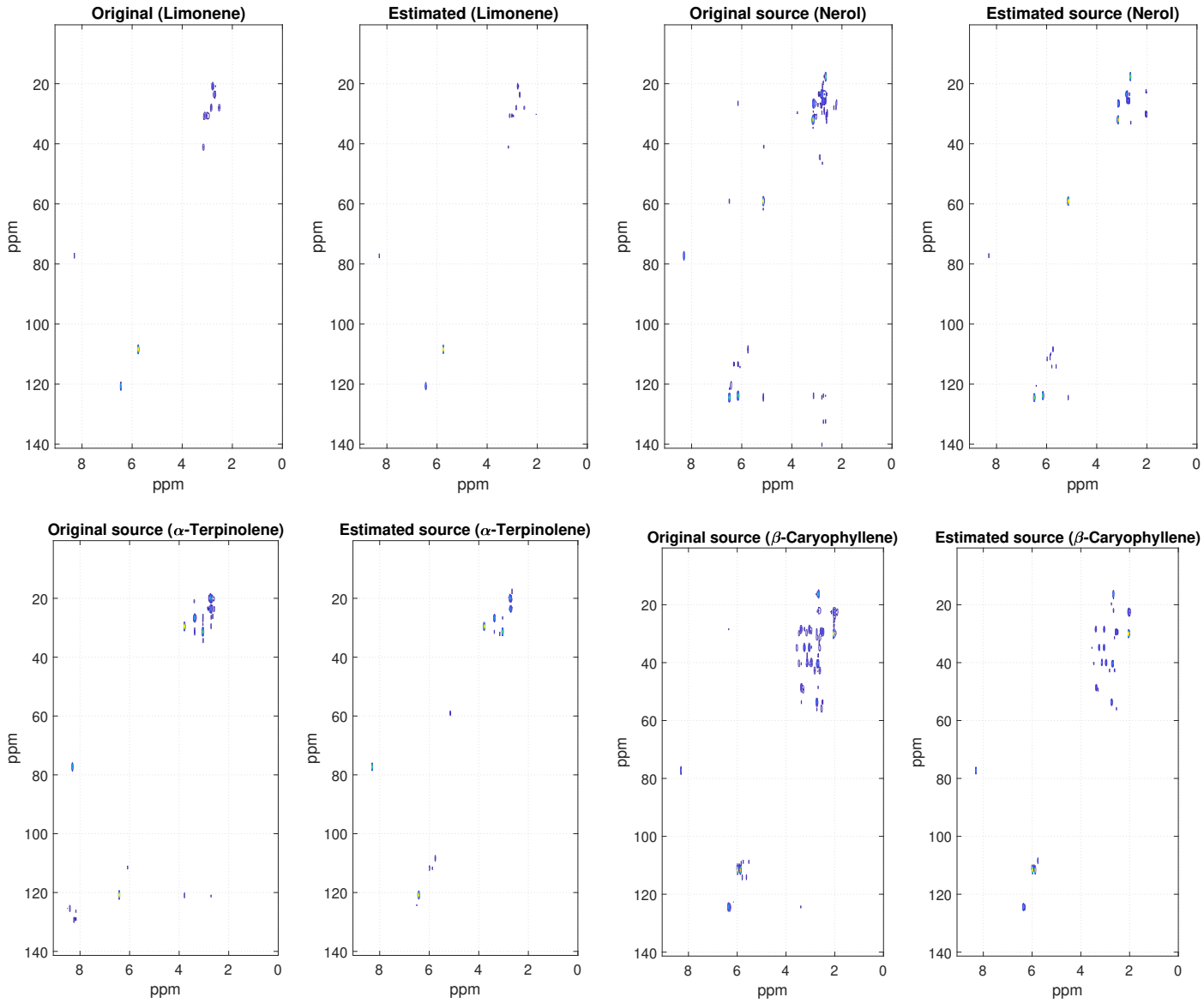

Fig. 6 2D ^{1}H-^{13}C HSQC spectra (real case): comparison of the pure HSQC spectra (left) and the HSQC spectra estimated from real mixtures, according to the two-dimensional LIM model.

	Limonene	Nerol	$\alpha-$Terpinolene	β-Caryophyllene
SIR	46.3	25.2	22.7	32.1
SDR	8.5	13.1	9.4	9.4

Table 9 2D ^{1}H-^{13}C HSQC spectra (real case): SIR and SDR indices (in dB) comparing pure and estimated source spectra using wavelet-based BC-VMFB with $\lambda = 10\sigma$.

	Limonene	Nerol	$\alpha-$Terpinolene	β-Caryophyllene
Solution 1	-0.39 %	3.16 %	-7.02 %	9.53 %
Solution 2	-8.56 %	-3.43 %	-2.66 %	10.90 %
Solution 3	5.84 %	8.42 %	2.94 %	6.68 %
Solution 4	5.74 %	-14.40 %	3.47 %	-12.84 %
Solution 5	-3.65 %	-6.89 %	0.03 %	-11.00 %

Table 10 2D ^{1}H-^{13}C HSQC spectra (real case): relative errors in the estimated concentrations (in %) using wavelet-based BC-VMFB with $\lambda = 10\sigma$. The corresponding Amari index equals 0.042.

4 and 5. SDR values are very low, which may be interpreted in terms of departures from the LIM model.

5 Discussion and conclusions

We have presented in this paper a general approach for the blind identification of compounds from solutions using NMR spectroscopy and blind source separation algorithms. It is worth recalling that these algorithms are blind in the sense that they attempt to estimate jointly pure compounds spectra and concentrations in solutions. From the mathematical point of view, we provided a general algorithmic approach, that includes as special instances a number of different algorithms, which we could test and compare. We also considered, for the sake of quantitative performance evaluation, some quantities (Amari index, various forms of signal to noise ratio), that had been introduced long ago in the blind source separation literature. Numerical tests were performed on data specifically generated for this work, including 1D as well as 2D HSQC spectra.

The results presented here show that blind source separation algorithms have the potential to perform successfully. On the studied dataset, results on simulated data range from good to excellent, depending on the algorithm and parameter values. On real data, good results could be obtained provided some important pre-processing steps could be done carefully.

However, our results also raise a number of questions, some of which we list below, that should be addressed before drawing more complete conclusions.

- Results on 1D data show that pre-processing is a crucial step. In the case considered here, alignment of spectra turned out to be fundamental. However, even after careful alignment, the Linear Instantaneous Mixture model turned out to be incompletely satisfactory, aligned data showing significant departures from that model. Therefore, this may suggest that the model is not 100% adequate, and that it would be worth considering more complex models, that could include for instance amplitude modulations as we observed on the terpene data we studied. One may also imagine including spectral shift into the model, provided the phenomenon could be sufficiently well understood and therefore modelled.

- In the same spirit, this would also suggest to modify the unmixing algorithms, to estimate amplitude modulations (and shifts) at the same time as concentrations and pure compounds spectra. There are situations in signal and image processing where one faces similar problems, it might be possible to transpose corresponding approaches to the case under consideration here.

- In the considered datasets, alignment was not problematic for the HSQC spectra, for which un-mixing algorithms could be run without pre-processing. The un-mixing results on HSQC data turned out to be of very good quality, the algorithms being able to identify clearly 2D fingerprints of the four terpenes, as well as concentrations. This is confirmed by satisfactory values for the interference index (SIR) and Amari index. Besides, the distortion index (SDR) are significantly weaker, which may indicate (as in the 1D case) that additional distortions should be taken into account in the model. However, as such indices have not been used so far (to the best of our knowledge) in NMR spectroscopy, such conclusions must be taken cautiously. More experiments are needed to validate the use of such tools in this context. It is also worth pointing out that the computational burden is significantly higher in the 2D case, and can be expected to grow fast when the dimension of spectra increases. This will be an important problem to address if one wants to proceed to higher dimensional spectra, where sparsity is expected to be higher and facilitate further the separation.

- Whatever the models, blind separation problems are always ill-posed problems, and in the framework of variational formulations, result on non-convex minimization problems. This means that the objective function to be optimized can have (and as a matter of fact, has) several (and often, many) local minima, and algorithms are extremely sensitive to initialization. We have stuck here to a simple choice, that was advocated in [18] for different algorithms. It is not clear that this choice is the most relevant for the family of approaches studied in this paper, this point clearly deserves an in-depth study. Very much in the same spirit, the fact that different algorithms that aim at optimizing the same objective function actually yield significantly different results and performances raises questions, even though there is no guarantee that they should give the same result, given the non-convexity of the problem. Again, further investigations are necessary at this point.

	Sol. 1	Sol. 2	Sol. 3	Sol. 4	Sol. 5
SIR	15.83	14.04	19.01	13.93	11.29
SDR	8.42	9.33	8.02	8.36	7.11

Table 11 2D ^{1}H-^{13}C HSQC spectra: SIR and SDR indices (in dB) comparing measured and simulated mixture spectra.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The project leading to this publication has received funding from the Excellence Initiative of Aix-Marseille University - A*Midex, a French "Investissements d'Avenir" program.

The research presented in this article is part of a project launched in collaboration with our friend and colleague Stefano Caldarelli, who passed away late 2018. His knowledge, ideas and enthusiasm have been at the heart of this work, and we miss him very much. This article is dedicated to his memory.

Notes and references

1 D. Marion, P. C. Driscoll, L. E. Kay, P. T. Wingfield, A. Bax, A. M. Gronenborn and G. Clore, *Biochemistry*, 1989, **28**, 6150–6156.

2 C. S. Johnson, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 1999, **34**, 203 – 256.

3 M. G. N. Reddy and S. Caldarelli, *Anal. Chem.*, 2010, **82**, 3266–3269.

4 I. Toumi, S. Caldarelli and B. Torrésani, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2014, **81**, 37 – 64.

5 D. Nuzillard, S. Bourg and J.-M. Nuzillard, *J. Magn. Reson.*, 1998, **133**, 358 – 363.

6 R. D. Boyer, R. Johnson and K. Krishnamurthy, *Journal of Magnetic Resonance*, 2003, **165**, 253–259.

7 A. Bax and D. G. Davis, *Journal of Magnetic Resonance (1969)*, 1985, **65**, 355–360.

8 G. N. M. Reddy and S. Caldarelli, *Chem. Commun.*, 2011, **47**, 4297–4299.

9 K. F. Morris and C. S. Johnson, *J. Am. Chem. Soc.*, 1992, **114**, 3139–3141.

10 O. Beckonert, H. C. Keun, T. M. D. Ebbels, J. Bundy, E. Holmes, J. C. Lindon and J. K. Nicholson, *Nature Protocols*, 2007, **2**, 2692.

11 L. Frydman, A. Lupulescu and T. Scherf, *J. Am. Chem. Soc.*, 2003, **125**, 9204–9217.

12 M. Mobli, M. W. Maciejewski, A. D. Schuyler, A. S. Stern and J. C. Hoch, *Phys. Chem. Chem. Phys.*, 2012, **14**, 10835–10843.

13 A. L. Guennec, P. Giraudeau and S. Caldarelli, *Anal. Chem.*, 2014, **86**, 5946–5954.

14 J. Xia, T. C. Bjorndahl, P. Tang and D. S. Wishart, *BMC Bioinformatics*, 2008, **9**, 507.

15 K. Bingol and R. Brueschweiler, *Analytical Chemistry*, 2011, **83**, 7412–7417.

16 A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, Inc., 2002.

17 P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 1st edn, 2010.

18 I. Toumi, B. Torrésani and S. Caldarelli, *Anal. Chem.*, 2013, **85**, 11344–11351.

19 P. Paatero and U. Tapper, *Environmetrics*, 1994, **5**, 111–126.

20 P. O. Hoyer, *J. Mach. Learn. Res.*, 2004, **5**, 1457–1469.

21 P. L. Combettes and J.-C. Pesquet, *Fixed-point algorithms for inverse problems in science and engineering*, Springer Verlag, 2010, pp. 185–212.

22 C. Chaux, P. L. Combettes, J.-C. Pesquet and V. R. Wajs, *Inverse Problems*, 2007, **23**, 1495–1518.

23 D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788–791.

24 D. D. Lee and H. S. Seung, Proc. Ann. Conf. Neur. Inform. Proc. Syst., 2001, pp. 556–562.

25 C. Févotte and J. Idier, *Neural Comput.*, 2011, **23**, 2421–2456.

26 E. Vincent, R. Gribonval and C. Févotte, *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2006, **14**, 1462–1469.

27 S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, Academic Press, 3rd edn, 2008.

28 I. Kopriva, I. Jeric and V. Smrecki, *Anal. Chim. Acta*, 2009, **653**, 143 – 153.

29 X. Shao, H. Gu, J. Wu and Y. Shi, *Applied Spectroscopy*, 2000, **54**, 731–738.

30 J. C. Cobas, P. G. Tahoces, M. Martin-Pastor, M. Penedo and F. J. Sardina, *J. Magn. Reson.*, 2004, **68**, 288–295.

31 J. Bolte, S. Sabach and M. Teboulle, *Mathematical Programming*, 2014, **146**, 459–494.

32 E. Chouzenoux, J.-C. Pesquet and A. Repetti, *Journal of Optimization Theory and Applications*, 2014, **162**, 107–132.

33 J.-F. Cardoso, *Proc. IEEE*, 1998, **86**, 2009–2025.

34 T. Vu and K. Laukens, *Metabolites*, 2013, **3**, 259–276.