JOURNAL OF MAGNETIC RESONANCE **133,** 358–363 (1998)
ARTICLE NO. MN981481

COMMUNICATIONS

Model-Free Analysis of Mixtures by NMR Using Blind Source Separation

D. Nuzillard,* S. Bourg,† and J.-M. Nuzillard†,[1]

*LAM, Faculté des Sciences, 51687 Reims Cedex 2, France; and †Laboratoire de Pharmacognosie, UPRESA 6013, Moulin de le Housse, 51097 Reims Cedex 2, France

Received January 6, 1998; revised April 1, 1998

The concept of blind source separation is described and examples of its use in 1D and 2D NMR spectroscopy are presented. The goal of this data processing method is to extract the spectra of components molecules when only mixtures are available. © 1998 Academic Press

Key Words: **mixture analysis; signal processing; blind source separation; second-order statistics; 2D NMR.**

There are numerous situations in which pure chemical compounds are not available for liquid state analysis by NMR spectroscopy. Because of this, new techniques which analyze mixtures need to be developed. Traditional mixture analysis addresses two main problems: the identification and the quantification of the components present in the mixture. Identification proceeds in most cases by matching the mixture's spectral information with a library of reference compounds. The analytical performance is thus strongly dependent on the library's content. A problem related to mixture analysis is the identification of components from a collection of spectral data of mixtures with unknown compositions. Such physical mixtures may be produced by a chromatographic or any other separation process. It seems intuitively reasonable that if one gets at least as many linearly independent spectra of mixtures as there are individual components, then it is possible to separate their spectra. This communication aims to show how the concept of blind source separation helps to achieve this goal. Blind source separation was developed in the context of multisources multisensors data processing. A set of sensors receives signals from sources, but with intensities depending on their relative positions. The data analysis yields the source signals and the mixing coefficients.

Various approaches to the problem were undertaken. The difficulty lies in finding a criterion of independence between the separated signals. The employed criteria are mainly based on the second- and/or fourth-order signal moments (*1*). Concepts as such neural networks (*2*), contrast functions (*3, 4*), maximum likelihood (*5*), and problem deflation (*6*) were involved in algorithms presenting various advantages as well as drawbacks. The algorithms are designed to perform either an adaptative (*7*) or a block processing. One of these algorithms, SOBI (second order blind identification), was designed to deal with temporally correlated signals. It is based on second-order statistical analysis and has proved to be robust in the context of noisy, nonstationary signals. Its principle is presented hereafter; details on performances and theoretical justifications can be found in reference (*8*).

Source separation by the SOBI algorithm resorts to properties of the covariance matrix of time-dependent vector functions. A function $\mathbf{x}_i(t)$, $1 \le i \le m$, sampled at times t_k ($0 \le k \le T-1$, $t_0 = 0$) is represented by a matrix $\mathbf{X}$ so that $\mathbf{X}_{ik} = \mathbf{x}_i(t_k)$. The covariance matrix $\mathbf{R}_\mathbf{x}(\tau)$ is defined by

$$\mathbf{R}_\mathbf{X}(\tau) = E[\mathbf{x}(t)\mathbf{x}^*(t + t_\tau)] \qquad [1]$$

or

$$[\mathbf{R}_\mathbf{X}(\tau)]_{ij} = \frac{1}{T-\tau} \sum_{k=0}^{T-1-\tau} x_{ik} x^*_{j,k+\tau} \qquad [2]$$

An autocovariance matrix $\mathbf{R}_\mathbf{X}(0)$ is denoted $\mathbf{R}_\mathbf{X}$. Let $\mathbf{C}$ be a matrix with m columns, and $\mathbf{C}^H$ its Hermitian conjugate. The property

$$\mathbf{R}_{\mathbf{CX}}(\tau) = \mathbf{C}\mathbf{R}_\mathbf{X}(\tau)\mathbf{C}^H \qquad [3]$$

will be used to handle covariance matrices.

Source mixing is considered a linear process in which no propagation lag is introduced:

[1] To whom correspondence should be addressed.

1090-7807/98 $25.00

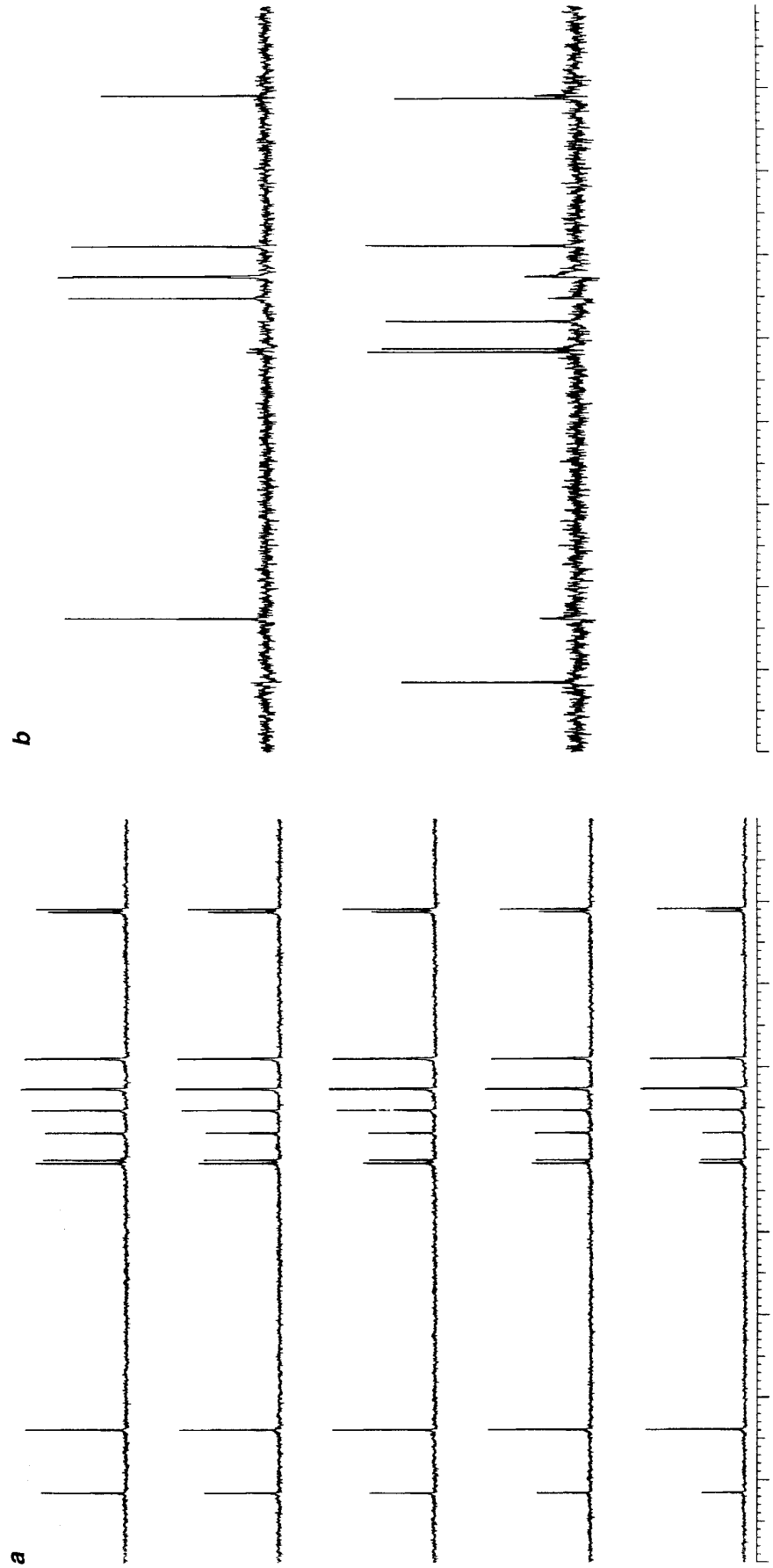

FIG. 1. (a) Five ^{13}C spectra recorded during the isomerization of α-glucose to β-glucose in D_2O. (b) The separated spectra of α-glucose (upper trace) and β-glucose (lower trace).

$$\mathbf{y}_i(t) = \sum_{j=1}^{n} a_{ij}\mathbf{s}_j(t) \quad \text{or} \quad \mathbf{Y} = \mathbf{AS} \qquad [4]$$

if n sources $\mathbf{s}_j$ are present. Signal mixtures $\mathbf{y}_i$ are detected by sensors and noise is introduced in the detected signals $\mathbf{x}_i$. Let $\mathbf{X}$ be the matrix of the measurements,

$$\mathbf{X} = \mathbf{Y} + \mathbf{N} \quad \text{or} \quad \mathbf{X} = \mathbf{AS} + \mathbf{N}, \qquad [5]$$

where $\mathbf{N}$ is the noise matrix and $\mathbf{A}$ is the mixing matrix. Signal samples, noise, and mixing coefficients are complex numbers.

Blind source separation consists in finding $\mathbf{A}$ and $\mathbf{S}$ with only $\mathbf{X}$ as input. The problem is largely underdetermined. The definition of $\mathbf{X}$ can be rewritten as

$$\mathbf{x}_i(t) = \sum_{j=1}^{n} \frac{a_{ij}}{\alpha_j}\alpha_j\mathbf{s}_j(t) + \mathbf{n}_i(t). \qquad [6]$$

The sources are thus only defined as relative values. A change of indexes j by permutation does not change $\mathbf{X}$. There is no way of labeling the sources unambiguously.

The source separation algorithm SOBI imposes constraints on the sources and the noise. The power of the sources is supposed to be normalized. This is not a restriction, as their absolute power cannot be determined. Sources are also pairwise statistically independent:

$$\mathbf{R}_\mathbf{S} = \mathbf{I}_n. \qquad [7]$$

Time-shifted source signals are statistically independent as well. This means that $\mathbf{R}_\mathbf{S}(\tau)$ is a diagonal matrix. Moreover, each source must be time correlated so that there is no diagonal term in $\mathbf{R}_S(\tau)$ which is always zero. The noise generated by the sensors must be time-uncorrelated, pairwise uncorrelated, and uncorrelated with the sources as well:

$$\mathbf{R}_\mathbf{N}(\tau) = \sigma^2\mathbf{I}_m\delta(\tau) \qquad [8]$$

$$E[\mathbf{n}(t)\mathbf{s}^*(t + t_\tau)] = 0, \qquad [9]$$

where δ is the Dirac function and σ^2 is the noise variance.

The m signals to be detected are linear combinations of the n sources ($m \geq n$). The size of the separation problem is reduced if n linearly independent combinations of the m signals are determined. Achieving this goal is possible through a judicious choice of combinations which form a set of normalized and orthogonal vectors. This approach was already proposed (*9*) in the context of the separation of signals of spin systems from homonuclear 3D spectra. Let $\mathbf{W}$ be the desired combination matrix, also named the whitening matrix:

$$\mathbf{R}_\mathbf{WY} = \mathbf{I}_n = \mathbf{WR}_\mathbf{Y}\mathbf{W}^H = \mathbf{WAR}_\mathbf{S}\mathbf{A}^H\mathbf{W}^H = (\mathbf{WA})(\mathbf{WA})^H. \qquad [10]$$

Once $\mathbf{W}$ is found, there must be a unitary matrix $\mathbf{U}$ so that $\mathbf{WA} = \mathbf{U}$. In the presence of sensor noise, only estimates $\hat{\mathbf{W}}$ and $\hat{\mathbf{U}}$ of $\mathbf{W}$ and $\mathbf{U}$ can be obtained, because the matrix $\mathbf{Y}$ is not experimentally accessible. The source separation is now split into two subproblems: the search for an estimate of the whitening matrix $\hat{\mathbf{W}}$ and for the unitary transformation $\hat{\mathbf{U}}$ so that

$$\hat{\mathbf{W}}\mathbf{A} = \hat{\mathbf{U}}. \qquad [11]$$

The whitened data matrix $\hat{\mathbf{W}}\mathbf{X}$ must fulfill

$$\mathbf{R}_{\hat{\mathbf{W}}\mathbf{X}} = \mathbf{I}_n. \qquad [12]$$

By diagonalization, $\mathbf{R}_\mathbf{X}$ can be written

$$\mathbf{R}_\mathbf{X} = \mathbf{H}\boldsymbol{\Delta}\mathbf{H}^H \qquad [13]$$

with

$$\boldsymbol{\Delta} = \text{diag}(\lambda_1, \ldots, \lambda_m) \text{ and } \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_m]. \qquad [14]$$

If $m = n$, then

$$\hat{\mathbf{W}} = \mathbf{H}\boldsymbol{\Delta}^{-1/2} \qquad [15]$$

because

$$\mathbf{R}_{\hat{\mathbf{W}}\mathbf{X}} = \boldsymbol{\Delta}^{-1/2}\mathbf{H}^H \cdot \mathbf{H}\boldsymbol{\Delta}\mathbf{H}^H \cdot \mathbf{H}\boldsymbol{\Delta}^{-1/2} = \mathbf{I}_n. \qquad [16]$$

If $m > n$, the lowest eigenvalues $(\lambda_{n+1}, \ldots, \lambda_m)$ of $\mathbf{R}_\mathbf{X}$ allow the noise power to be estimated:

$$\hat{\sigma}^2 = \frac{1}{m-n}\sum_{i=n+1}^{m}\lambda_i. \qquad [17]$$

Setting

$$\lambda_i' = \lambda_i - \hat{\sigma}^2, \quad \boldsymbol{\Delta}' = \text{diag}(\lambda_1', \ldots, \lambda_n'), \text{ and}$$
$$\mathbf{H}' = [\mathbf{h}_1, \ldots, \mathbf{h}_n] \qquad [18]$$

gives

$$\hat{\mathbf{W}} = \mathbf{H}'\boldsymbol{\Delta}'^{-1/2}. \qquad [19]$$

Another whitening method is described in reference (*10*).

The matrix $\hat{\mathbf{U}}$ is obtained by first considering that

$$\mathbf{R}_\mathbf{X}(\tau) = \mathbf{R}_\mathbf{Y}(\tau) \quad \text{for } \tau > 0 \qquad [20]$$

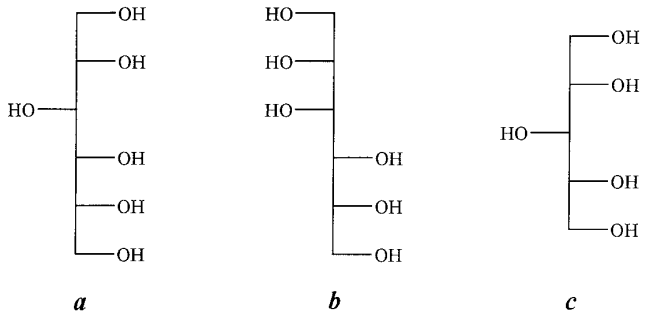

FIG. 2. The structures of sorbitol (a), mannitol (b), and xylitol (c).

because of the properties of noise. Then

$$\mathbf{R}_{\hat{\mathbf{W}}\mathbf{X}}(\tau) = \mathbf{R}_{\hat{\mathbf{W}}\mathbf{Y}}(\tau) = \hat{\mathbf{W}}\mathbf{A}\mathbf{R}_{\mathbf{S}}(\tau)(\hat{\mathbf{W}}\mathbf{A})^{H}. \qquad [21]$$

This means that $\hat{\mathbf{U}} = \hat{\mathbf{W}}\mathbf{A}$ is the unitary transformation that diagonalizes $\mathbf{R}_{\hat{\mathbf{W}}\mathbf{X}}(\tau)$ independently of τ, because $\mathbf{R}_{\mathbf{S}}(\tau)$ is diagonal. A robust estimation of $\hat{\mathbf{U}}$ is performed by jointly diagonalizing a set of $\mathbf{R}_{\hat{\mathbf{W}}\mathbf{X}}(\tau)$ matrices, by means of an algorithm related to the Jacobi diagonalization method. The matrix $\hat{\mathbf{U}}$ is built as a product of elementary complex-valued rotation matrices named Givens matrices. Details may be found in reference (*8*). The simplest choice of the τ values is 1, 2, 3, ... A better choice is possible if some prior knowledge of the time correlation properties of the sources is available.

The mixing matrix is then determined as the estimate

$$\hat{\mathbf{A}} = \hat{\mathbf{W}}^{\#}\hat{\mathbf{U}} \qquad [22]$$

where $\hat{\mathbf{W}}^{\#}$ is the pseudo-inverse of $\hat{\mathbf{W}}$. An estimation of the source signals is finally achieved, considering that

$$\hat{\mathbf{A}}\hat{\mathbf{S}} = \mathbf{X} = \mathbf{R}_{\mathbf{X}}\mathbf{R}_{\mathbf{X}}^{-1}\mathbf{X} = \hat{\mathbf{A}}\hat{\mathbf{A}}^{H}\mathbf{R}_{\mathbf{X}}^{-1}\mathbf{X}, \qquad [23]$$

and therefore

$$\hat{\mathbf{S}} = \hat{\mathbf{A}}^{H}\mathbf{R}_{X}^{-1}\mathbf{X}. \qquad [24]$$

The whole process can be summarized as follows:

1. Estimation of $\hat{\mathbf{W}}$ by diagonalization of $\mathbf{R}_{\mathbf{X}}$.
2. Computation of a set of $\mathbf{R}_{\hat{\mathbf{W}}\mathbf{X}}(\tau)$ matrices.
3. Evaluation of $\hat{\mathbf{U}}$ that jointly diagonalizes the set.
4. Estimation of the mixing matrix: $\hat{\mathbf{A}} = \hat{\mathbf{W}}^{\#}\hat{\mathbf{U}}$.
5. Estimation of the source signals: $\hat{\mathbf{S}} = \hat{\mathbf{A}}^{H}\mathbf{R}_{\mathbf{X}}^{-1}\mathbf{X}$.

The ^{13}C NMR spectroscopy provides time-domain signals possessing the required qualities to be submitted to source separation by means of the SOBI algorithm. The orthogonality constraint expressed in Eq. [7] for the sources is fulfilled when the spectra of the components to be separated do not overlap significantly. The ^{13}C resonance lines are generally narrow enough to limit the probability of exact peak superimposition. The thermal noise recorded by the spectrometers is supposed to have good properties (Eqs. [8] and [9]). The modeling of NMR time-domain signals as sums of decaying exponential functions proves their time-correlation property. The linear mixing model (Eq. [4]) is correct for physical mixtures of components if there is no significant chemical shift change due to intra- or intermolecular interactions. This last point is probably the most questionable, as discussed below.

Two applications of blind source separation to spectra of mixtures of chemicals will be presented, in 1D and 2D NMR spectroscopy. The first example deals with the isomerization of α-glucose into β-glucose in D_2O. Crystallized α-glucose (10 mg) was dissolved in 1 mL D_2O. After 90 min a series of 20 ^{13}C FIDs were acquired. The acquisition time was 5 min and was followed by a 10 min delay. Five time-domain signals were built by coadding FIDs by groups of four consecutive records. The corresponding spectra are presented in Fig. 1a. The time evolution of α-glucose into β-glucose is visible, even though the changes are not dramatic. The number n of sources was set to 2. The SOBI algorithm produces time-domain data, whose Fourier transforms are presented in Fig. 1b. The top and bottom traces are the ^{13}C NMR spectra of α-glucose and β-glucose, respectively. These spectra are difficult to obtain in a ''pure state,'' as isomerization takes place as soon as the compounds are in solution. In this context, blind source separation is used as a generalized and automatic way of performing difference spectroscopy. Some defects are visible, especially in the spectrum of β-glucose. Dispersion-like signals arise from small frequency glitches of resonance lines due to concentration effects. The SOBI algorithm takes as parameters the time lags required for the computation of covariance matrices. The default choice proposed by the authors of the algorithm was taken as $t_\tau = t_1$, t_2, t_3, and t_4. Different conditions do not bring any significantly better results.

The example in Fig. 1 shows that the separation seems to cause a degradation of signal-to-noise ratio in the separated spectra. This effect can be analyzed on a very simple example. Two signals are combined to provide two mixtures of close composition according to the mixing matrix

$$\mathbf{A} = \begin{bmatrix} 1+\epsilon & 1 \\ 1 & 1+\epsilon \end{bmatrix}. \qquad [25]$$

The detection process introduces noise of identical levels into both recorded signals. The separation is achieved by application of the matrix

$$\mathbf{A}^{-1} = \frac{1}{2\epsilon}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad [26]$$

given at the first order of approximation when ϵ is small. The noise level after separation is proportional to $1/\epsilon$, and therefore

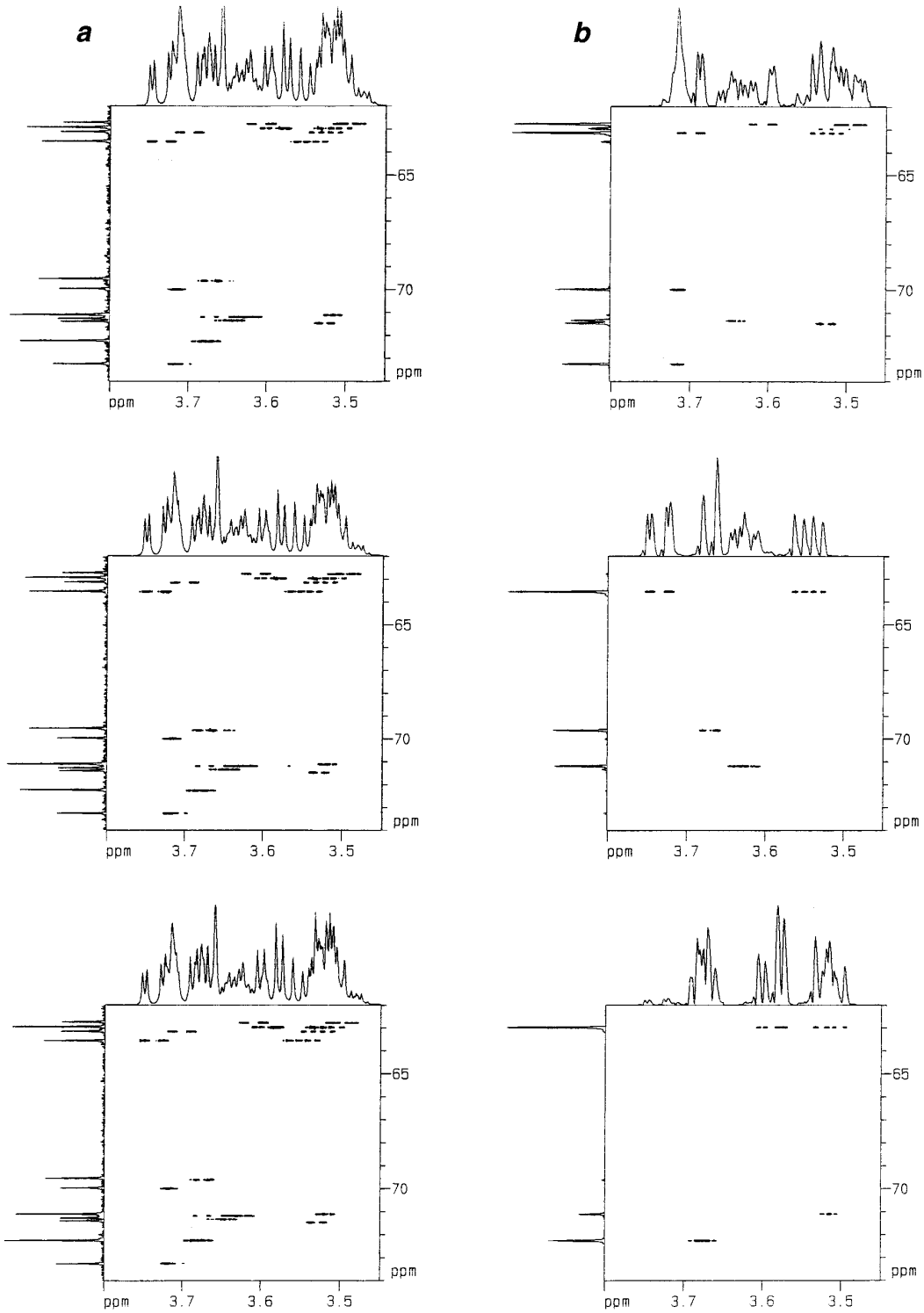

FIG. 3. (a) The HSQC spectra of three mixtures of sorbitol, mannitol, and xylitol in D_2O. Their relative proportions are about 1/1/1 (top), 1/1.2/1 (middle), 1/1/1.2 (bottom), respectively. Spectral widths are 20 ppm and 1 ppm, numbers of points in the time domain are 512 and 1024, and sizes of the spectra are 2048 and 1024 in dimensions 1 and 2, respectively. Linear prediction is applied in F_1. Four scans resulting in 90 min acquisition times were used for each spectrum. (b) The separated HSQC spectra of the components of the mixtures.

the SNR of the separated signals varies like ϵ. Thus, at the extreme limit where the mixtures have the same composition ($\epsilon = 0$), separation cannot be achieved. In Fig. 1a, the kinetics of the reaction do not make important changes in the intensities of the resonance lines. This causes the observed low SNR of the spectra in Fig. 1b.

The second application deals with the separation of the gradient enhanced HSQC (*11*) spectra of three mixtures of three compounds, sorbitol, mannitol, and xylitol, in D_2O (see Figs. 2 and 3a) at concentrations in the range 40–60 mM. The advantage of considering 2D NMR techniques for blind source separation is the spreading of information into a 2D plane instead of a 1D axis. The possibilities of signal overlap are lower and therefore the requirement of signals orthogonality is easier to meet.

The SOBI algorithm was designed to deal with 1D time domain signals. Therefore, some pre- and postprocessing must be performed. In a first attempt rectangular zones were defined around the peaks; their content was collected line by line and put all together on a single row in a precise order. The 1D pseudo spectra thus obtained were subjected to an inverse Fourier transformation to form three 1D pseudo FIDs. They contain the relevant information of the 2D spectra but in a more compact form. After separation of the pseudo FIDs, Fourier transformation, and reconstruction of three 2D data sets, the result was disappointing. A closer look at the three regular 1D spectra showed that resonance frequencies for a given nucleus may appear shifted by many times the linewidth. It is clearly impossible to achieve separations in this context.

Supplementary steps were added to the processing in order to cope with the frequency shift of the peaks. Thus, the spectra were integrated over the rectangular zones. The pseudo 1D spectra were built as vectors whose components are the integral values. The separation was then performed on the inverse Fourier transforms of these ''integral spectra.'' The Fourier transforms of the separated pseudo FIDs yield the integral values of the separated spectra. During the reconstruction of the 2D spectra, each zone receives the data of the peak from the original 2D spectrum with the highest integral. The peak is then scaled so that its integral takes the value given by the corresponding point in the integral spectrum.

The result is presented in Fig. 3b. Parameters for SOBI were set as in the 1D example. The projections of the 2D spectra illustrate the quality of the separation. In the spectrum of sorbitol (top), small unwanted peaks appear. They are mainly due to overlapping peak extensions caused by temperature fluctuations. The latter arise from the heating of the water solution by the heteronuclear decoupling pulse sequence. Projections along the F_1 axis are the sum of matrix columns. The peak intensities in these projections reflect rather well the number of attached protons: 1/1/1/1/2/2 for the nonsymmetrical sorbitol, 1/1/2 and 2/1/4 for the symmetrical mannitol and xylitol.

Spectra were recorded on a Bruker DRX 500 spectrometer. Standard acquisition programs were used: zgdc and inviedgptp for the 1D and 2D spectra, respectively. The SOBI algorithm was provided by the author of reference (*8*) as a MATLAB (The MathWorks, Inc.) code. Basic processing was performed using xwinnmr and routines written in C language.

Blind source separation provides a useful tool for mixture analysis by NMR. The concept of a mixture itself can be extended to spectral editing. For example, a series of DEPT (*12*) spectra can be considered as the ^{13}C spectra of CH, CH_2, and CH_3 subunits of a molecule, appearing with intensities modulated by the angle of the read pulse. In this context there are no frequency shift problems because all the signals come from the same sample. The same principle holds for the interpretation of DOSY spectra. These applications are under development and involve the quantitative aspect contained in the mixing matrix **A**.

ACKNOWLEDGMENT

S.B. thanks the Département de la Marne (France) for financial support.

REFERENCES

1. J.-F. Cardoso, *in* ''Proc. ICASSP,'' pp. 2109–2112 (1989).
2. C. Jutten and J. Hérault, *Signal Processing,* **24,** 1–10 (1991).
3. D. Dohono, *in* ''Applied Time-Series Analysis II,'' pp. 565–609, Academic Press, New York (1981).
4. P. Comon, *Signal Processing* **36,** 287–314 (1994).
5. M. Gaeta and J.-L. Lacoume, *in* ''Proc. EUSIPCO,'' pp. 621–624 (1990).
6. N. Delfosse and P. Loubaton, *in* ''Proc. ICASSP,'' pp. 41–44 (1994).
7. E. Moreau and O. Macchi, *in* ''Proc. IEEE SP Workshop on Higher-Order Stat.,'' Lake Tahoe, NV, pp. 215–219 (1993).
8. A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, *IEEE Trans. SP* **45,** 434–444 (1997).
9. D. Abergel and M.-A. Delsuc, *J. Mol. Struct.* **286,** 65–70 (1993).
10. A. Belouchrani, Ph.D. Thesis, ENST, Paris, France (1995).
11. W. Willker, D. Leibfritz, R. Kerssebaum, and W. Bermel, *Magn. Reson. Chem.* **31,** 287–292 (1993).
12. M. R. Bendall, D. T. Pegg, and D. M. Doddrell, *J. Magn. Reson.* **52,** 81–117 (1983).