LETTER

A Matrix Pencil Approach to the Blind Source Separation of Artifacts in 2D NMR Spectra

K.Stadlthanner, F.J.Theis, E.W.Lang
Institute of Biophysics, Neuro-and Bioinformatics Group
University of Regensburg, D-93040 Regensburg, Germany
E-mail: elmar.lang@biologie.uni-regensburg.de

A.M.Tomé
Departamento de Electrónica e Telecomunicações
Universidade de Aveiro, P-3810 Aveiro, Portugal
E-mail: ana@ieeta.pt

W. Gronwald, H.-R. Kalbitzer
Institute of Biophysics, NMR Spectroscopy Group
University of Regensburg, D-93040 Regensburg, Germany

(Submitted on September 12, 2003)

Abstract– Multidimensional proton nmr spectra of biomolecules dissolved in aqueous solutions are usually contaminated by an intense water artifact. We discuss the application of the generalized eigenvalue decomposition (GEVD) method using a matrix pencil to solve the blind source separation problem of removing the intense solvent peak and related artifacts. 2D NOESY spectra of simple solutes as well as dissolved proteins are studied.

Keywords– blind source separation, independent component analysis, generalized eigendecomposition, matrix pencil, 2D NOESY spectra

1. Introduction

Blind source separation addresses the problem of finding which signals contribute to any given sensor signal recorded. Blind source separation has a very close relationship to a recently developed new statistical data processing technique called Independent Component Analysis (ICA). For recent authorative reviews see [4] [3]. In ICA either one assumes statistically independent source signals and exploits higher order correlations in the data or one exploits the time correlations in signals relying on second order statistics only. In any case a linear mixing model is considered generally.

Second order techniques exploit the temporal structure of the source signals. The blind identification of the mixing model can be converted to standard (EVD) or generalized (GEVD) eigenvalue decomposition and simultaneous or joint diagonalization (SD) problems. In algorithms like AMUSE and EFOBI [10] a standard EVD is performed of a matrix derived from fourth-order cumulants or time-delayed correlations. Algorithms like SOBI [2] instead perform a joint approximative diagonalization of a set of delayed covariance matrices of whitened data to extract their average eigenstructure. Recently GEVD solutions have been presented which comprise an exact simultaneous diagonalization of a matrix pencil formed with the sensor signals [8], [9].

We will follow the matrix pencil approach and apply it to the separation of the intense water resonance and related artifacts from two-dimensional Nuclear Overhauser Enhancement Spectroscopy Nuclear Magnetic Resonance (2D NOESY NMR) spectra.

2. The generalized eigendecomposition approach

For convenience we shortly review the generalized eigendecomposition approach using congruent matrix pencils. The matrix pencil $(\mathbf{R}_{s1}, \mathbf{R}_{s2})$ formed with the source signals and the matrix pencil $(\mathbf{R}_{x1}, \mathbf{R}_{x2})$ formed with the sensor signals are considered congruent if there exists an *invertible* matrix $\mathbf{A}$ such that

$$\begin{aligned} \mathbf{R}_{x1} &= \mathbf{A}\mathbf{R}_{s1}\mathbf{A}^T \\ \mathbf{R}_{x2} &= \mathbf{A}\mathbf{R}_{s2}\mathbf{A}^T \end{aligned} \tag{1}$$

In BSS problems $\mathbf{A} = \{a_{ij}\}, i = 1, ..., m, \quad j = 1, ..., n$ represents the instantaneous mixing matrix. If, A is an invertible matrix, two pencils related as described by the eqn.(1) are called congruent pencils and possess identical eigenvalues.

The inverse or pseudo-inverse of the mixing matrix can be estimated from the sensor signal pencil if the eigenvector matrix $\mathbf{E}_s$ of the source signal pencil is diagonal. The generalized eigen-decomposition of the sensor signal pencil now reads

$$\mathbf{R}_{x1}\mathbf{E} = \mathbf{R}_{x2}\mathbf{E}\mathbf{\Lambda} \tag{2}$$

where $\mathbf{E}$ represents the *unique* eigenvector matrix if the diagonal matrix $\mathbf{\Lambda}$ has *distinct* eigenvalues λ_i. The corresponding eigen-decomposition statement concerning the source signal pencil can be obtained easily by substituting eqn.(1) into eqn.(2) yielding

$$\mathbf{R}_{s1}\mathbf{E}_s = \mathbf{R}_{s2}\mathbf{E}_s\mathbf{\Lambda} \tag{3}$$

where $\mathbf{E}_s$ represents its eigenvector matrix and the normalized eigenvectors corresponding to a particular eigenvalue are related by

$$\vec{e}_s = \alpha\mathbf{A}^T\vec{e} \tag{4}$$

with α is a normalizing constant.

The eigenvector matrix $\mathbf{E}$ forms an estimate of the inverse of the mixing matrix $\mathbf{A}$ if the matrix $\mathbf{E}_s$ corresponds to the identity matrix or a simple permutation matrix as is the case if the source signal pencils are both diagonal.

In summary the GEVD approach to BSS problems is feasible if the congruent source signal pencil is diagonal with distinct relations among diagonal entries, i.e. with distinct eigenvalues.

3. Computing the eigen-decomposition of symmetric pencils

A very common approach to compute the eigenvalues and eigenvectors of a matrix pencil is to reduce the GEVD statement, eqn.(2) to the standard EVD problem which is of the form

$$\mathbf{C}\mathbf{Z} = \mathbf{Z}\mathbf{\Lambda} \tag{5}$$

The strategy that we will follow is first to solve the eigen-decomposition of the matrix $\mathbf{R}_{x1}$ giving

$$\mathbf{R}_{x1} = \mathbf{S}\mathbf{D}\mathbf{S}^T = \mathbf{S}^{1/2}\mathbf{D}^{1/2}\mathbf{S}^T\mathbf{S}\mathbf{D}^{1/2}\mathbf{S}^T = \mathbf{W}\mathbf{W} \tag{6}$$

Substituting this result into the GEVD statement and defining $\mathbf{Z} = \mathbf{W}\mathbf{E}$ yields the transformed equation

$$\mathbf{W}^{-1}\mathbf{R}_{x2}\mathbf{W}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{\Lambda} \tag{7}$$

which is of the standard EVD form of a real symmetric matrix $\mathbf{C} = \mathbf{W}^{-1}\mathbf{R}_{x2}\mathbf{W}^{-1}$ if the matrix $\mathbf{R}_{x2}$ is also symmetric positive definite and the transformation matrix $\mathbf{W}^{-1}$ is obtained as

$$\mathbf{W}^{-1} = \mathbf{S}\mathbf{D}^{-1/2}\mathbf{S}^T \tag{8}$$

While the eigenvalues of the matrix pencil are available from the solution of the EVD of the matrix $\mathbf{C}$ the corresponding eigenvectors are obtained via $\mathbf{E} = \mathbf{W}^{-1}\mathbf{Z}$.

4. NMR spectra

Modern multi-dimensional NMR spectroscopy [5] is a versatile tool for the determination of the native 3D structure of biomolecules in their natural aqueous environment. Proton NMR is an indispensable contribution to this structure determination process but is hampered by the presence of the very intense water (H_2O) proton signal. Hence it is interesting whether blind source separation (BSS) techniques can contribute to the removal of the water artifact in such spectra without regard to any sophisticated water suppression pulse protocols except a simple pre-saturation to reduce the dynamic range problem.

Concerning structure determination homonuclear 2D NOESY spectra are a must. They relay on the nuclear Overhauser effect [5] and provide information about cross-relaxation rates which for protons mainly depend on magnetic dipolar interactions. The latter vary with distance as r^{-6} hence allow distances to neighboring nuclei to be determined. Loosely speaking one can consider it an atomic ruler which allows the 3D-structure to be determined if enough NOE's are available experimentally.

A two-dimensional NMR time domain signal, called free induction decay (FID), is modelled by a sum of damped complex harmonic functions

$$S(t_1, t_2) = \sum_i M_i \exp(-i\Omega_{1i} t_1) \exp(-\lambda_{1i} t_1) \exp(-i\Omega_{2i} t_2) \exp(-\lambda_{2i} t_2) \tag{9}$$

to which Gaussian noise is superimposed. Signal processing is routinely performed by Fourier analysis, resulting in spectra made of sums of Lorentzian shaped resonance lines given by

$$f(\omega_1, \omega_2) = \sum_i M_i \left(\frac{1}{i\Delta\Omega_{1i} + \lambda_{1i}} \right) \cdot \left(\frac{1}{i\Delta\Omega_{2i} + \lambda_{2i}} \right) \tag{10}$$

Statistical independence of two signals requires their scalar product to be zero both in the time domain or in the frequency domain. Therefore non-overlapping resonance lines should be reasonably independent [6]. But because of the limited range of chemical shifts, i.e. the spread of the proton resonances on the frequency scale is rather limited compared to individual resonance line widths, statistical independence is hard to assure in general. Second order techniques like the GEVD using matrix pencils discussed above exploit some weaker conditions for the separation of sources assuming that they have temporal structure with different autocorrelation functions or equivalently different power spectra.

5. Results and Discussion

Pre-saturation of the water resonance has been applied in all cases. FID's $S(t_{1,j}, t_2)$ recorded at fixed evolution times $t_{1,j}, \quad j = 1, ..., m$ were sampled over time spans t_2 and have been Fourier transformed to obtain corresponding 1-dim spectra $S(\omega_2, t_{1,j})$ corresponding to the j-th increment of the evolution time $t_{1,j}$. The final $m \times N$ matrix $\mathbf{X}$ contained as many rows as there were different evolution times $t_{1,j}$ according to the experimental protocol with each line representing a single 1D spectrum. A matrix pencil $(\mathbf{R}_{x,1}, \mathbf{R}_{x,2})$ comprises two correlation matrices $\mathbf{R}_x$ of zero mean data where the second correlation matrix $\mathbf{R}_{x,2}$ is formed with delayed or filtered data forming the matrix $\mathbf{R}_{x,1}$. This latter matrix is computed as follows

$$\mathbf{R}_{x,1} = \frac{1}{N} \mathbf{S}(\omega_2, t_1) \mathbf{S}^H(\omega_2, t_1) \tag{11}$$

with $N = 2048$ representing the number of samples in the ω_2 domain and $\mathbf{S}^H$ the conjugate transpose of the matrix $\mathbf{S}$. The second correlation matrix $\mathbf{R}_{x,2}$ of the pencil has been computed after filtering each single spectrum (each row of $\mathbf{S}(\omega_2, t_1)$) with a bandpass filter of Gaussian shape centered on the water resonance with a variance in the range of $1 \leq \sigma^2 \leq 4$. Both matrices of the pencil are of dimension 128×128 as we assume a symmetrical situation hence we consider as many source signals as there are sensor signals.

5.1. Toy spectra

5.1.1. Artificial 2D nmr spectra

These artificially generated spectra represent test cases in the sense that the proton 2D nmr spectra comprise only few well separated solute and solvent resonances. Artificially generated FIDs are of the

form

$$s_k(t) = A_k \exp(-\frac{t}{T_{2,k}}) \exp(2\pi i f_k t + i\Phi_k) \tag{12}$$

with parameters given in Table (5.1.1):

	A	$f[Hz]$	$T_2[s]$	Φ
$s_1(t)$	0.0892	0.1	2.5	$\pi/4$
$s_2(t)$	0.2208	3	0.4	0
$s_3(t)$	0.2208	8	0.4	0

Signal $s_1(t)$ represents a water resonance with a phase shift usually introduced by the pre-saturation pulse and the other two signals represent solute signals with considerably shorter spin-spin relaxation times T_2. Note that $s_1(t)$ and $s_2(t)$ overlap while $s_3(t)$ is well separated in the Fourier transformed spectrum (see Fig. (1). The FIDs have been digitized to 2048 points with a spacing of $0.01s$ yielding signals $\vec{s}_i(t)$ with unit norm. The Nyquist frequency of the Fourier transformed signals amounts to $f_N = 50Hz$. With these source signals corresponding sensor signals have been generated according to the general ICA model $\vec{x}(t) = \mathbf{A}\vec{s}(t)$ or $\hat{x}(\omega) = \mathbf{A}\hat{s}(\omega)$ using the randomly generated mixing matrix

$$\mathbf{A} = \begin{bmatrix} -0.1893 & 0.0121 & -0.4360 \\ -0.5256 & 0.7028 & -0.3267 \\ -0.1020 & -0.6196 & 0.6833 \end{bmatrix} \tag{13}$$

These sensor signals will be analyzed with the matrix pencil algorithm discussed above. The performance of the BSS can be assessed by calculating Amari's cross-talking error (CTE)[1].

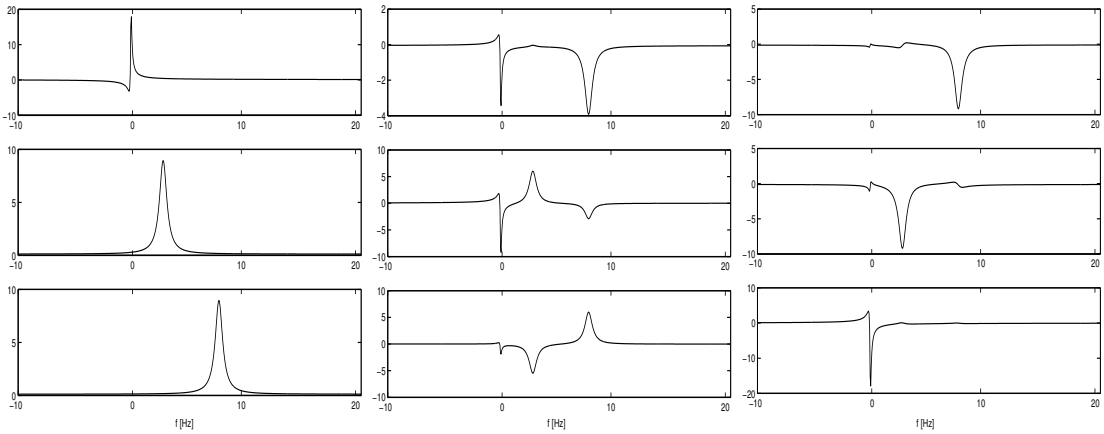

Figure 1. *left:* artificially generated source signals *middle:* artificially generated sensor signals *right:* Reconstructed source signals obtained with the matrix pencil algorithm

To apply the matrix pencil algorithm a filtered version of the sensor signals is needed. The latter has been generated by applying a Gaussian bandpass filter centered at the the water resonance

$$H(\omega) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\omega - 2\pi a)^2}{2\sigma^2}\right) \tag{14}$$

with $a = 0.1Hz$. The CTE varies considerably with the width parameter σ of the bandpass filter as is shown in Fig.(2) and Fig.(1) also shows the estimated source signals. Note that results obtained in the time domain are far less convincing than those in the frequency domain. Note further that the matrix pencil algorithm is very quick producing results within less than a second.

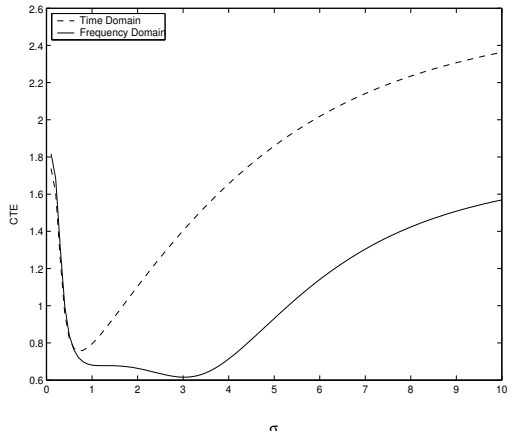

Figure 2. Dependence of the cross-talking error (CTE) on the width parameter σ of the Gaussian bandpass filter

5.2. Experimental spectra of simple solutes

Next 2D NOESY spectra of simple solute molecules like ethylen diamine-N,N,N',N'-tetra-acetate (EDTA) have been analyzed. In the simple case of EDTA $m \times N = 30 \times 2048$ data matrices turned out to be sufficient. The second correlation matrix $\mathbf{R}_{x,2}$ of the pencil has been obtained again with bandpass filtered signals in the frequency domain.

The matrix pencil thus obtained has been treated in the manner given above to estimate the independent components of the EDTA spectra and the corresponding demixing matrix. Roughly 25 independent components showing spectral energy only in the frequency range at 4.8 ppm have been assigned to the highly distorted water resonance. To effect a separation of the water resonance and the EDTA spectra these water related independent components have been set to zero deliberately. Then the whole EDTA spectrum could be reconstructed with the estimated inverse of the demixing matrix and the corrected matrix of estimated source signals.

A typical 1D EDTA spectrum is shown in Fig.(1) illustrating the intense water resonance at 4.8 ppm. Also shown is the reconstructed spectrum obtained with the matrix pencil algorithm. The small distortions remaining are due to baseline artifacts caused by truncating the FID due to limited sampling times.

5.3. Spectra of the protein *Tm*CSP

Next we present experimental 2D NOESY spectra of the cold shock protein of the bacterium *Thermotoga maritima*(*Tm*CSP) which contain many protein resonances. Note that the water resonance overlaps considerably with part of the protein spectrum with some protein resonances very close to or even hidden underneath the solvent resonance. Fig.(4) shows an original *Tm*CSP protein spectrum with the prominent water resonance and its reconstructed version with the water resonance separated out by applying Tomé's GEVD algorithm using a matrix pencil in the frequency domain. Both correlation matrices had dimension (128×128) and all 2048 data points have been used to estimate the expectations within the correlation matrices. Again the data used to form the second correlation matrix $\mathbf{R}_{x,2}$ of the matrix pencil corresponded to a bandpass ($\sigma = 1$) filtered version of the original data. Some remnants of the intense water resonance are still visible in the reconstructed spectrum indicating that a complete separation into independent components has not been achieved yet. This is due to the limited number of independent components that could be estimated with the data available.

6. Conclusions

We have shown that ICA methods can be useful to separate water resonances and concomitant baseline distortions from solute resonances and obtain largely undistorted solute spectra. Generalized eigenvalue

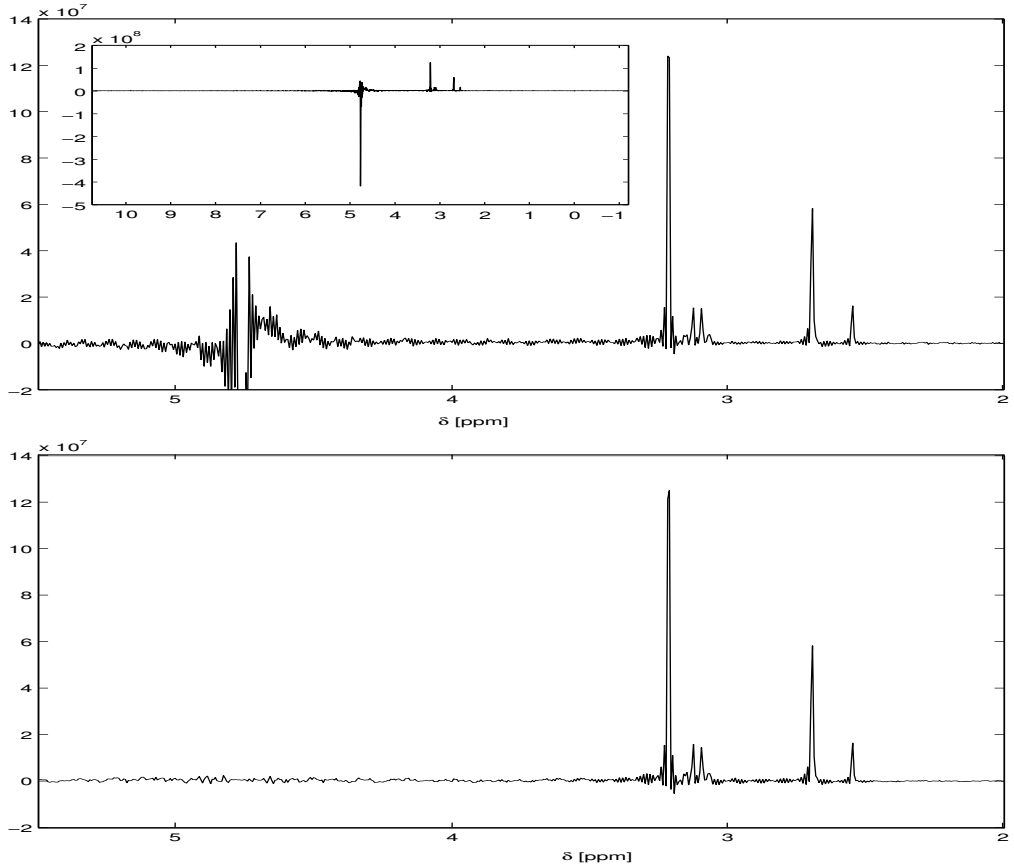

Figure 3. *top:* 1D slice of a 2D NOESY spectrum of EDTA in aqueous solution corresponding to the shortest evolution period t_2. The chemical shift ranges from $-1.206ppm$ to $10.759ppm$, *bottom:* Reconstructed EDTA spectrum with the water artifact removed with the matrix pencil algorithm

decompositions using a matrix pencil represent an exact and easily applied second order technique to effect such artifact removal from the spectra. We have tested this method with simple EDTA spectra where no solute resonances appear close to the water resonance. Application of the method to protein spectra with resonances hidden in part by the water resonance showed a good separation quality with only little remaining spectral distortions in the frequency range of the removed water resonance. The reconstructed spectra also show very convincingly that any baseline distortions stemming from the intense water artifact can be automatically cured as well. In summary the GEVD approach using congruent matrix pencils is an algebraically exact, very fast and easy to implement algorithm. Further investigations will have to improve the separation quality even further and will have to answer the question if solute resonances hidden underneath the water resonance can be made visible with these or related methods.

7. References

[1] S.Amari, A.Cichocki, H.H.Yang, "A new learning algorithm for blind signal separation", *NIPS* **8**, 1129 - 1159, (1995)

[2] A.Belouchrani, K.Abed-Meraim, J.-F.Cardoso, E.Moulines, "A blind source separation technique using second-order statistics", *IEEE Trans. Signal Processing* **45**, 434-444, (1997)

[3] A. Cichocki, S. Amari, *Adaptive Blind Signal and Image Processing*, Wiley and Sons, New York, USA, 2002

[4] A. Hyvärinen, J.Karhunen, E.Oja, *Independent Component Analysis*, Wiley and Sons, New York, USA, 2001

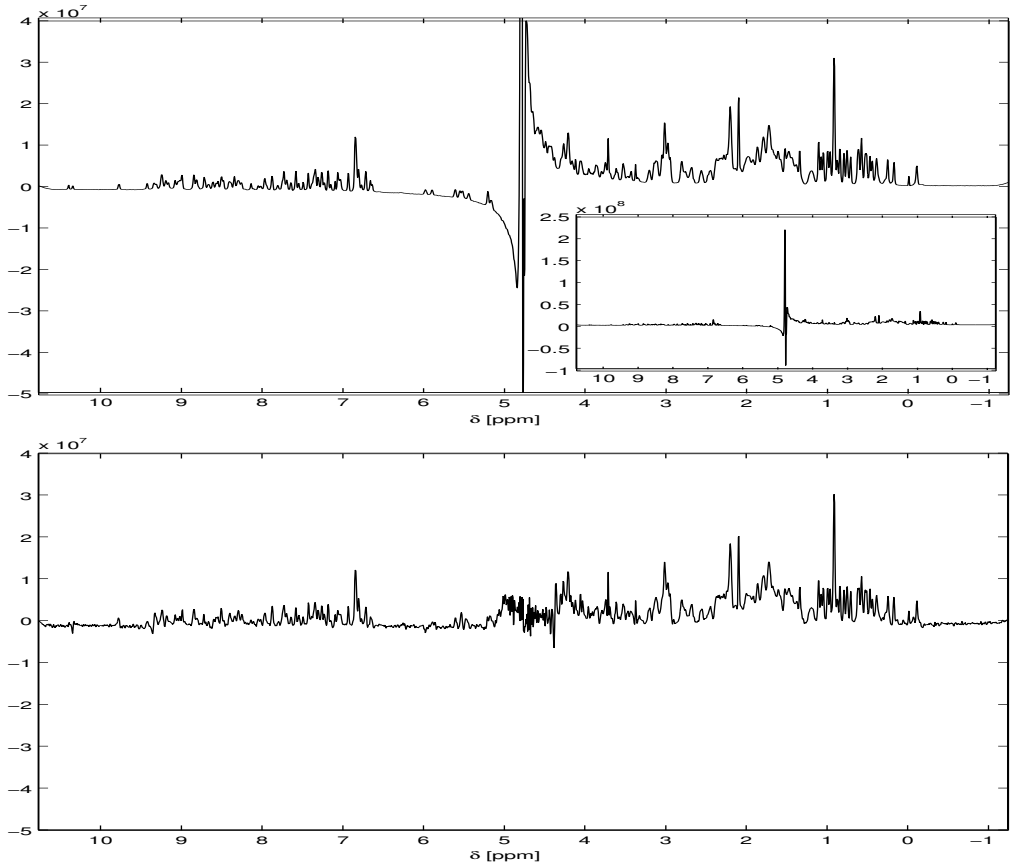

Figure 4. *top:* 1D slice of a 2D NOESY spectrum of the protein *Tm*CSP in aqueous solution corresponding to the shortest evolution period t_1. *bottom:* Reconstructed *Tm*CSP protein spectrum obtained with the matrix pencil algorithm

[5] K.H.Hausser, H.-R.Kalbitzer, *NMR in Medicine and Biology*, Springer Verlag, Berlin, 1989

[6] D.Nuzillard, J.-M.Nuzillard, "Application of Blind Source Separation to 1D and 2D Nuclear Magnetic Resonance Spectroscopy", *IEEE Signal Processing Lett.* **5**, 209-211, (1998)

[7] A.Souloumiac, "Blind Source Detection Using Second Order Non-Stationarity", *Proc. Int.Conf.Acoustics, Speech and Signal Processing*, Detroit, USA, p.1912-1916, (1995)

[8] A.M.Tomé, "An iterative eigendecomposition approach to blind source separation", *Proc. 3rd Int.Conf. on Independent Component Analysis and Signal separation*, San Diego, USA, p.424-428 (2001)

[9] A.M.Tomé, E.W.Lang, "Approximative Diagonalization Approach to Blind Source Separation with a Subset of Matrices", *Proc. 7^{th} Int. Symp. Signal Processing and Applications* (ISSPA'2003), Paris, France, (2003)

[10] L. Tong, R. Liu, V.C.Soon, Y.F.Huang, "Indeterminacy and identifiability of blind identification", *IEEE Trans. Circuits and Systems* **38**, 499-509, (1991)

Kurt Stadlthanner graduated in physics from the University of Regensburg and is doing his PhD as a stipendiate of the graduate college "Nonlinearity and Nonequilibrium in condensed matter" of the faculty of physics of the University of Regensburg. His research interest are signal processing, independent component analysis and biomedical applications.

Fabian J. Theis received a PhD in Physics from the University of Regensburg and a PhD in Computer Science from the University of Granada. Currently he works as postdoctoral researcher at the

Neuro- and Bioinformatics group of the University of Regensburg. His research interests include statistical signal processing, linear and nonlinear independent component analysis and overcomplete blind source separation based on sparse component analysis.

Elmar W. Lang received his PhD in Physics from the University of Regensburg and currently is a member of the Institute of Biophysics of the University of Regensburg and head of the Neuro- and Bioinformatics group. His current research interests include Biomedical Signal and Image Processing, Blind Source Separation, Independent Component Analysis and Biomedical Applications.(home page: http://www.biologie.uni-regensburg.de/Biophysik/Lang/index.html)

Ana M.Tomé - PhD from the University of Aveiro in 1990, and currently a member of the Department of Electronics and Telecommunications/ IEETA of the University of Aveiro. Her research interests are Digital Signal Processing, Blind Source Separation, Independent Component Analysis and Applications.

Wolfram Gronwald graduated from the University of Braunschweig in 1991, and currently is a lecturer in biophysics. His research interests include development and application of computational methods for three-dimensional protein structure determination. (Home page: http://www-nw.uni-regensburg.de/ grw28475.biophysik.biologie.uni-regensburg.de/)

Hans Robert Kalbitzer graduated in medicine (1976) and physics (1981) at the University of Heidelberg and is currently professor for biophysics at the University of Regensburg. His research interest includes structural biology, NMR spectroscopy and software development. (Home page:http://biologie.uni-regenburg.de/Biophysik/Kalbitzer).