BLIND SOURCE SEPARATION OF WATER ARTEFACTS IN NMR SPECTRA USING A MATRIX PENCIL

[1] *K. Stadlthanner,* [2] *A.M. Tomé,* [1] *F.J. Theis,* [1] *W. Gronwald* [1] *H.-R. Kalbitzer,* [1] *E.W. Lang*

[1] Institute of Biophysics, University of Regensburg, D-93040 Regensburg, Germany
[2] Dept. de Electrónica e Telecomunicações, Universidade de Aveiro, P-3810 Aveiro, Portugal
email: elmar.lang@biologie.uni-regensburg.de

ABSTRACT

Multidimensional ^{1}H nmr spectra of biomolecules dissolved in light water are contaminated by an intense water artefact. We discuss the application of the generalized eigenvalue decomposition method using a matrix pencil to explore the time structure of the signals in order to separate out the water artefacts. 2D NOESY spectra of simple solutes as well as dissolved proteins are studied. Results are compared to those obtained with the FastICA algorithm, too.

1. INTRODUCTION

Modern multidimensional NMR spectrocopy [4], [7] is a versatile tool for the determination of the native 3D structure of biomolecules in their natural aqueous environment. Proton NMR, i.e. the observation of the magnetization of the ^1H nuclei in the probe, is an indespensible contribution to this structure determination process but is hampered by the presence of the very intense water (H_2O) proton signal. Since it is the most intense signal in two-dimensional spectra it causes the most trouble with baseline distortions and t_1 noise and it can obscure weak signals lying under its skirts. Because of its intensity it also causes severe dynamic range problems hence sophisticated experimental protocols have been developed to suppress the water signal as far as possible. All these procedures introduce spectral distortions that can neither be avoided nor removed and prevent the analysis of the spectral region close to the water resonance. Hence equivalent spectra of the molecules dissolved in heavy water (D_2O) have to be taken also which raises additional problems not the least being that heavy water differs in its physical-chemical properties from light water sufficiently to doubt a direct comparison of both spectra. Hence it is interesting whether blind source separation techniques can contribute to the removal of the water artefact in such spectra without regard to any sophisticated water suppression pulse protocols except a simple presaturation to reduce the dynamic range problem [5]. It has to be noted that even a long weak pulse on the water resonance can bleach nearby solute proton resonances and can also affect other signals through cross-relaxation or chemical exchange. Concerning structure determination homonuclear 2D NOESY spectra are a must. They rely on the nuclear Overhauser effect, the change in the intensity of the resonance of one spin species upon saturation of an adjacent spin with which it has an appreciable dipole-dipole interaction. They 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 corresponds to a sum of free induction decay (FID) signals, each for a fixed evolution period $t_{1,j}$, and can be modelled by a sum of damped complex harmonic functions

$$\begin{aligned} S(t_{1,j},t_2) &= \sum_i M_i \exp(-i\Omega_{1i}t_{1,j})\exp(-\lambda_{1i}t_{1,j}) \\ &\cdot \exp(-i\Omega_{2i}t_2)\exp(-\lambda_{2i}t_2) \end{aligned} \quad (1)$$

to which Gaussian noise is superimposed. The evolution period is incremented during the experiment to yield typically 256 FIDs. Signal processing is performed by Fourier analysis, resulting in spectra made of sums of Lorentzian shaped resonance lines [4] given by

$$\begin{aligned} S(\omega_1,\omega_2) &= \sum_i M_i \left(\frac{1}{i\Delta\Omega_{1i}+\lambda_{1i}}\right) \\ &\cdot \left(\frac{1}{i\Delta\Omega_{2i}+\lambda_{2i}}\right) \end{aligned} \quad (2)$$

Statistical independence of two signals implies 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. But because of the limited range of chemical shifts, i.e. the spread of the proton resonances on the frequency scale, compared to individual resonance line widths, statistical independence is hard to assure in general. Second order blind identification techniques like the GEVD using matrix pencils discussed below as well as many others exploit some weaker conditions for the separation of sources assuming that they have temporal structure with different autocorrelation functions or equivalently different power spectra.

Blind source separation (BSS) adresses the problem of finding which signals contribute to any given sensor signals recorded. It is of interest if little or nothing is known about the source signals and the mixing process, hence the term blind. BSS has a very close relationship to a recently developed new statistical data processing technique called Independent Component Analysis (ICA). For recent authorative reviews see [8] [3], [12]. In general the problem is very ill-posed and needs to be regularized to become solvable. Two venues have been considered in the past. Either one assumes statistically independent source signals and exploits higher order correlations in the data or one exploits time correlations in the data relying on second order statistics only. In any case a linear mixing model is considered mostly.

Higher order decorrelation techniques have been intensely studied recently and many algorithms have been proposed (see [8], [3], [14]). 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 [17], [1]. Recently GEVD solutions have been presented which comprise the simultaneous diagonalization of a matrix pencil formed with the sensor signals. The matrices forming the pencil can be computed in different ways: Souloumiac [13] considers two segments of time-dependent signals with distinct energies, Lo [10] considers different embedding spaces of chaotic signals Molgedey [11] and Chang [2] compute time-delayed correlation matrices and Tomé [15] considers filtered versions of the sensor signals. Later Tomé [16] also presented an algebraic formulation of the GEVD problem using the notion of congruent matrix pencils and block matrix operations. An iterative method to compute the eigendecomposition of a symmetric positive definite pencil has also been presented. We will follow this latter approach and apply it to the separation of water artefacts from 2D NOESY NMR spectra. A higher order decorrelation technique as implemented with the FastICA algorithm [9] has been applied also and corresponding results will be shown.

2. THE GENERAL EIGENDECOMPOSITION APPROACH

For convenience we shortly review the generalized eigenvalue decomposition (GEVD) approach using congruent matrix pencils [15], [16]. Consider 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. Both pencils are considered congruent if there exists an invertible matrix $\mathbf{A}$ such that

$$\begin{aligned} \mathbf{R}_{x1} &= \mathbf{A}\mathbf{R}_{s1}\mathbf{A}^T \\ \mathbf{R}_{x2} &= \mathbf{A}\mathbf{R}_{s2}\mathbf{A}^T \end{aligned} \tag{3}$$

In BSS problems $\mathbf{A} = \{a_{ij}\}, i = 1, ..., m, \quad j = 1, ..., n$ represents the instantaneous mixing matrix. Congruent pencils possess the same eigenvalues which form the roots of the characteristic polynomials

$$\begin{aligned} \chi_x(\lambda) &= \det(\mathbf{R}_{x1} - \mathbf{R}_{x2}) = 0 \\ \chi_s(\lambda) &= \det(\mathbf{R}_{s1} - \mathbf{R}_{s2}) = 0 \end{aligned} \tag{4}$$

The GEVD of the sensor signal pencil now reads

$$\mathbf{R}_{x1}\mathbf{E} = \mathbf{R}_{x2}\mathbf{E}\mathbf{\Lambda} \tag{5}$$

$$\mathbf{A}\mathbf{R}_{s1}\mathbf{A}^T\mathbf{E} = \mathbf{A}\mathbf{R}_{s2}\mathbf{A}^T\mathbf{E}\mathbf{\Lambda} \tag{6}$$

where $\mathbf{E}$ represents the unique eigenvector matrix if the diagonal matrix $\mathbf{\Lambda}$ has distinct eigenvalues λ_i. The latter statement can be obtained easily by substituting eqn.(1) into eqn.(3).

If $\mathbf{A}$ is an $(m \times m)$ invertible matrix both sides of eqn.(6) can be multiplied by $\mathbf{A}^{-1}$ and using $\mathbf{E}_s = \mathbf{A}^T\mathbf{E}$ the corresponding GEVD statement of the source signal pencil results

$$\mathbf{R}_{s1}\mathbf{E}_s = \mathbf{R}_{s2}\mathbf{E}_s\mathbf{\Lambda} \tag{7}$$

where $\mathbf{E}_s$ represents its eigenvector matrix. Concerning the BSS problem eqn.(6) shows that the generalized eigenvector matrix 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. This is encountered if the source signal pencils are both diagonal.

In summary the GEVD approach to BSS problems is feasible if the congruent source signal pencils are formed with uncorrelated source signals yielding the identity matrix or a permutation matrix only.

3. COMPUTING THE EIGENDECOMPOSITION OF SYMMETRIC PENCILS

A very common approach to compute the eigenvalues and eigenvectors of a matrix pencil is to reduce the GEVD statement

$$\mathbf{R}_{x2}\mathbf{E} = \mathbf{R}_{x1}\mathbf{E}\mathbf{\Lambda} \tag{8}$$

to the standard EVD problem which is of the form

$$\mathbf{C}\mathbf{Z} = \mathbf{Z}\mathbf{\Lambda} \tag{9}$$

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

$$\mathbf{R}_{x1} = \mathbf{S}\mathbf{D}\mathbf{S}^T = \mathbf{S}^{1/2}\mathbf{D}^{1/2}\mathbf{S}^T\mathbf{S}\mathbf{D}^{1/2}\mathbf{S}^T = \mathbf{W}\mathbf{W} \tag{10}$$

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

$$\mathbf{W}^{-1}\mathbf{R}_{x2}\mathbf{W}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{\Lambda} \tag{11}$$

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

$$\mathbf{W}^{-1} = \mathbf{S}\mathbf{D}^{-1/2}\mathbf{S}^T \tag{12}$$

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

4. RESULTS AND DISCUSSION

4.1. EDTA spectra

First 2D NOESY spectra of simple solute molecules like EDTA (Ethylen diamine-N,N,N',N'-tetraacidic acid - $(HO_2CCH_2)_2NCH_2CH_2N(CH_2CO_2H)_2$) have been analyzed. Presaturation of the water resonance has been applied in all cases. FID's $S(t_1, t_2)$ recorded at fixed evolution times t_1 were sampled over timespans t_2 and have been Fourier transformed with respect to both time domains to obtain corresponding spectra $S(\omega_1, \omega_2)$ which could be corrected for any phase distortions. Data matrices have then been formed with one row representing one single spectrum $S(\omega_2, t_1)$ corresponding to a fixed evolution time t_1. The final matrix then contained as many rows as there were different evolution times t_1 according to the experimental protocol. Typically 512 evolution periods have been considered and 2048 data points were sampled of each

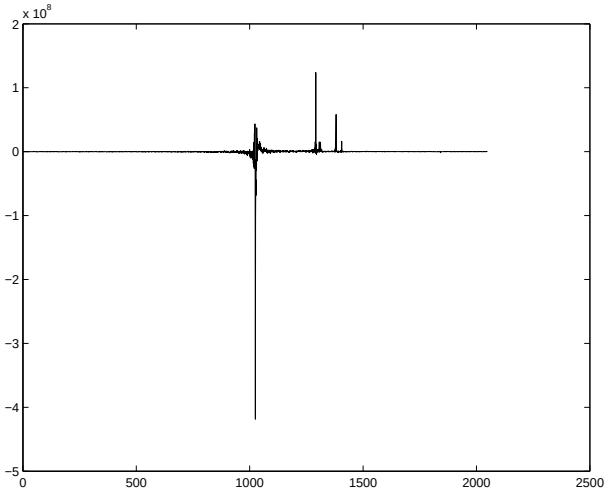

Figure 1: 1D slice of a 2D NOESY spectrum of EDTA in aqueous solution corresponding to the shortest evolution period t_2. The chemical shift ranges from $10.759ppm$ (left) to $-1.206ppm$ (right). Only the real part of the complex quantity $S(\omega_2, t_1)$ is shown.

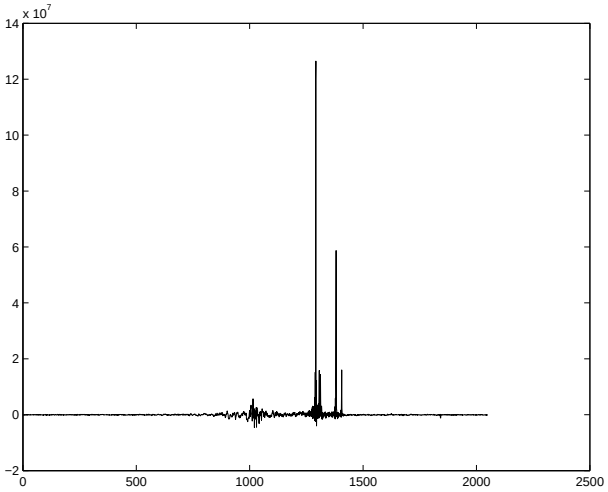

Figure 2: Reconstructed EDTA spectrum obtained according to procedure a) with the water artefact removed

spectrum. However due to phase cycling only 128 spectra can be considered at most but with more than 25 EDTA spectra no further improvement could be observed hence data matrices of size (30×2048) have been used finally. Assuming as many sources as there were sensor signals a (30×30) demixing matrix has been understood.

A matrix pencil $(\mathbf{R}_1, \mathbf{R}_2)$ of zero mean data comprises two correlation matrices $\mathbf{R}$ of the data where

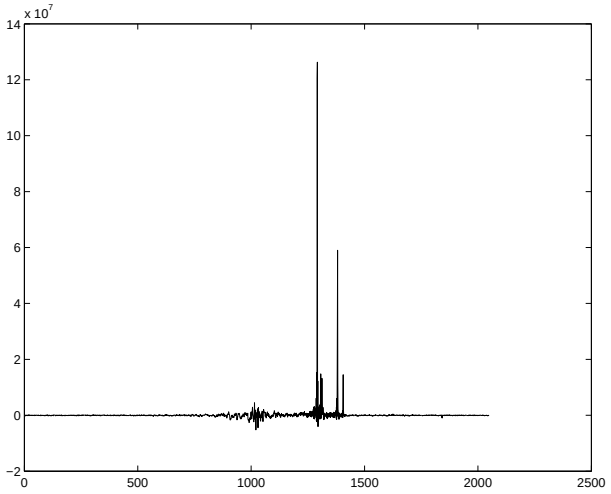

Figure 3: Reconstructed EDTA spectrum with the water artefact removed as obtained with the FastICA algorithm

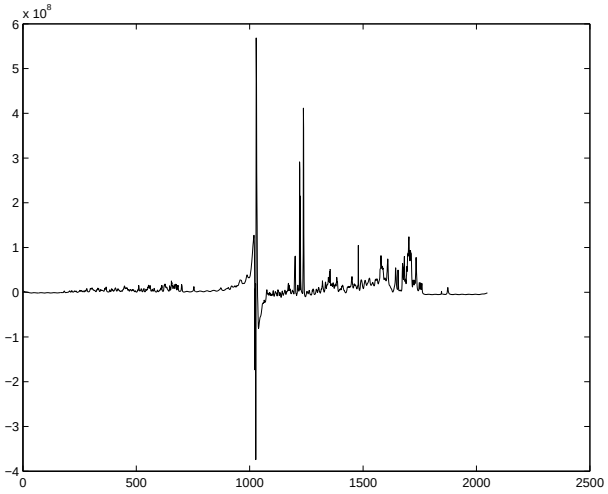

Figure 4: 1D slice of a 2D NOESY spectrum of the protein RALGEF in aqueous solution corresponding to the shortest evolution period t_1. The chemical shift ranges from $10.822ppm$ to $-1.189ppm$, i.e. one digit corresponds to a shift of 5.864E-3 ppm

the second correlation matrix $\mathbf{R}_2$ represents a delayed or filtered version of $\mathbf{R}_1$. The latter are of dimension 30×30 and the expectations have been estimated according to

$$< x_i x_j >= \frac{1}{N} \sum_{n=1}^{N} x_i(n) x_j(n) \qquad (13)$$

with $N = 2048$ representing the number of samples in the ω_2 domain in case of $\mathbf{R}_1$.

The second correlation matrix $\mathbf{R}_2$ of the pencil has been obtained in two different ways:
a) first by collecting spectral data at frequencies below the water resonance, i.e. only data points between 1285 - 2048 have been used to calculate the expectations in the covariance matrix R_2 of the pencil. That amounts to lowpass filtering the whole spectrum. It should be noted that any smaller frequency shifts did not yield reasonable results, i.e. a successful separation of the water and the EDTA resonances could not be obtained then. Note that the entries to the correlation matrices consisted of samples taken in the frequency domain.
b) A second procedure was more elaborate. First the time domain FIDs have been Fourier transformed to the frequency domain to apply an appropriate phase correction. Afterwards a bandstop filter centered at the water resonance has been applied. The spectra then have been converted to the time domain with an inverse Fourier transform and corresponding correlation matrices have been calculated with time domain data for both correlation matrices of the pencil. Note that even in case of $\mathbf{R}_1$ the data had to be Fourier transformed first to be able to effect a phase correction to the spectra which then were subjected to an inverse Fourier transform to obtain suitably corrected time domain data.

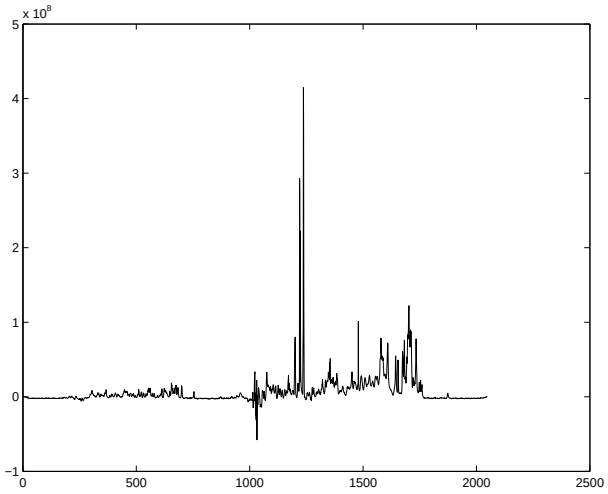

Figure 5: Reconstructed protein spectrum corresponding to Fig.1 using the matrix pencil in the time domain with the water artefact removed

The matrix pencil thus obtained has been treated in the manner given above to estimate the idependent components (ICs) of the EDTA spectra and the corresponding demixing matrix. Independent components showing spectral energy only in the frequency range of

the water resonance have been related with the water artefact. It turned out that roughly 25 ICs were related with the water signal and 4-5 ICs with the EDTA signal. Because of the many distortions the water signal shows this high number of ICs was necessary. To effect a separation of the water artefact and the EDTA spectra these water related IC's 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 still intense water artefact around sample point 1050 corresponding to a relative frequency shift (known as chemical shift in nmr) of 4.8 ppm. Fig.(2) presents the reconstructed spectrum with the water artefact removed. The small distortions remaining are due to baseline artefacts caused by truncating the FID due to limited sampling times.
To see whether the use of higher order statistics could perform better the data set has also been analyzed with the FastICA algorithm [9] which also uses a prewhitening stage. Again 30 data points in each column of the (128×2048)-dimensional data matrix $\mathbf{X}$ have been used. Again independent components related to the water artefact have been nulled in the reconstruction procedure and the result is shown in Fig.(4). Visual inspection shows a comparable separation quality of both methods in the case of 2D NOESY EDTA spectra.

4.2. Protein spectra

As a second data set 2D NOESY spectra of the protein RALGEF (infact only the RAS binding domain of RALGEF which represents a 87 aminoacids long part of the C-termus of the protein) [6] have been analyzed, too. This case is more difficult as the water resonance appears in between the protein resonances, hence overlapps considerably with part of the protein resonances and even hides some protein resonances completely. The data again have been analyzed with the matrix pencil method as described above. This time both correlation matrices had dimension (128×128) and all 2048 time domain data points have been used to estimate the expectations within the correlation matrices. Again the second correlation matrix $\mathbf{R}_2$ of the matrix pencil corresponded to a bandstop filtered version of correlation matrix $\mathbf{R}_1$. Fig. (4 - 5) show an original protein spectrum with the prominent water artefact and its reconstructed version with the water artefact separated out.
It is to be noted that an equally good separation of the water artefact could be obtained if both correlation matrices have been formed with the frequency domain data and the correlation matrix $\mathbf{R}_2$ has been calculated by estimating the corresponding expectations with the low frequency samples, those with shifts below the water resonance, of the spectrum only as is shown in Fig.(6).

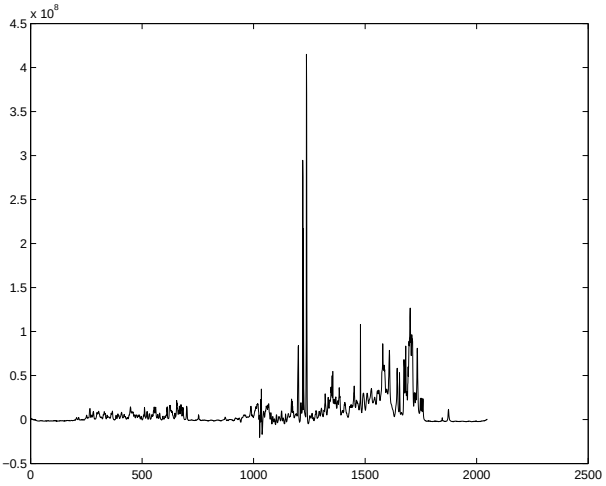

Figure 6: Reconstructed protein spectrum obtained with the GEVD algorithm using a matrix pencil in the frequency domain.

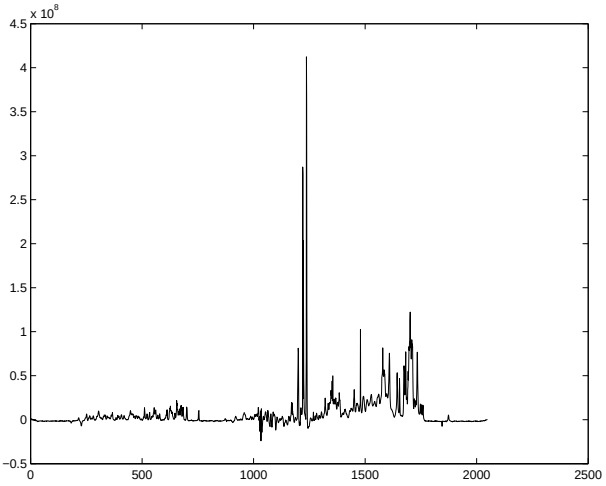

Figure 7: Reconstructed protein spectrum with the water artefact removed as obtained with the FastICA algorithm

Again the data have been analyzed with the FastICA algorithm as well yielding comparable results as is shown in Fig.(7). But it is to be noted, though hardly visible on the figures presented, that the FastICA algorithm introduced some spectral distortions that have not been observed in the analysis with the GEVD method using a matrix pencil. This is of course

an important issue concerning an automatized water artefact separation procedure, as any spectral distortions might result in false structure determinations using these 2D NOESY data.

5. CONCLUSIONS

Proton 2D NOESY spectra represent an indispensable ingredient to any determination of the three-dimensional conformation of native proteins, which forms the basis for understanding their function in living cells. Water is the most abundant molecule in biological systems, hence proton protein spectra are generally contaminated by large water resonances causing severe dynamic range problems. We have shown that ICA methods can be useful to separate these water artefacts out and obtain largely undistorted pure protein spectra. Generalized eigenvalue decompositions using a matrix pencil represent an exact and easily applied second order technique to effect such arefact 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 spetral distortions in the frequency range of the removed water resonance. It is important to note that no noticable spectral distortions have been introduced farther away from the water artefact contrary to the FastICA algorithm which introduced distortions also in other parts of the spectrum. It is to be noted further that baseline artefacts due to the intense water resonance can be alleviated also to a large extent with this procedure. 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. Also methods to suitably quantify a comparison of different separation results need to be developed.

6. REFERENCES

[1] 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)

[2] Chang, Ding, Yau, Chan, A matrix pencil approach to blind source separation of colored nonstationary signals, IEEE Trans. Signal Processing **48**, 900-907, (2000)

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

[4] R.R.Ernst, G.Bodenhausen, A.Wokaun, Principles of nuclear magnetic resonance in one and two dimensions, Clarendon Press, Oxford, (1987)

[5] R. Freeman, Spin Choreography, Spektrum Academic Publishers, Oxford, (1997)

[6] M. Geyer, Ch. Herrmann, S. Wohlgemuth, A. Wittinghofer, H.-R. Kalbitzer, Structure of the RAS-binding domain of RALGEF and implications for RAS binding and signalling, Nature Structural Biology, **4**, 694-699, (1997)

[7] K.H.Hausser, H.-R.Kalbitzer, NMR in Medicine and Biology, Springer Verlag, Berlin, (1991)

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

[9] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Computation **9**, 1483 - 1492, (1996)

[10] Lo, Leung, Litva, Separation of a mixture of chaotic signals, Proc. Int. Conf. Accoustics, Speech and Signal Processing, Detroit, USA, p.1798-1801, (1996)

[11] L.Molgedey, H.G.Schuster, Separation of a mixture of independent signals using time delayed correlatiions, Phys.Rev.Lett. **72**, 3634-3637, (1994)

[12] St.Roberts, R. Everson
Independent Component Analysis / Principles and Practice
Cambridge University Press, Cambrige, U.K., 2001

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

[14] F.J.Theis, A.Jung, C.G.Puntonet, E.W.Lang, Linear Geometric ICA: Fundamentals and Algorithms, Neural Computation, **15**, 1-21, (2002)

[15] A.M.Tomé, Blind source separation using a matrix pencil, Int. Joint Conf. on Neural Networks (IJCNN), Como, Italy, (2000)

[16] 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)

[17] 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)