ON THE USE OF INDEPENDENT COMPONENT ANALYSIS TO REMOVE WATER ARTIFACTS OF 2D NMR PROTEIN SPECTRA


*- Institute of Biophysics, University of Regensburg, D-93040, Germany
+ Dep. Electrónica e Telecomunicações/IEETA, Universidade Aveiro, 3810-Aveiro, Portugal

### ABSTRACT

Multidimensional 1 H nmr spectra of biomolecules dissolved in light water are contaminated by an intense water artifact. Independent Component Analysis (ICA) is used to extract a set of signals out of a set of measured or sensed signals without knowing how the mixing process is carried out. Hence it is interesting to apply ICA techniques to the removal of water artifact in such spectra. Experimental 2D NOESY spectra of proteins are studied.

1. **INTRODUCTION**

Modern multi-dimensional NMR spectroscopy is a versatile tool for the determination of the native 3D structure of biomolecules in their natural aqueous environment [1]. Proton NMR is an indispensable contribution to this structure determination process but is hampered by the presence of the very intense water (H\textsubscript{2}O) proton signal. The latter causes severe dynamic range problems, the most trouble with baseline distortions and t\textsubscript{1} noise and can obscure weak signals lying under its skirts. Sophisticated experimental protocols have been developed to suppress the water signal as far as possible but introduce spectral distortions that prevent the analysis of the spectral region close to the water resonance. 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 presaturation to reduce the dynamic range problem [1]. 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.

A two-dimensional NMR time domain signal S(t\textsubscript{1}, t\textsubscript{2}) corresponds to a sum of free induction decay (FID) signals sampled for fixed evolution periods t\textsubscript{1} and extending over a sampling time interval of duration t\textsubscript{2},. The evolution period t\textsubscript{1} is incremented during the experiment to yield typically 512 FIDs. Signal processing is performed by Fourier analysis, resulting in spectra S(ω\textsubscript{1}, ω\textsubscript{2}) made of sums of Lorentzian shaped resonance lines [1].

ICA techniques extract a set of signals out of a set of measured signals without knowing how the mixing process is carried out [2], [3]. Hence it is interesting whether ICA can contribute to the removal of the water artifact in such spectra. Considering that the set of measured spectra S is a linear combination of a set of independent components X, i.e., S = AX. The goal is to estimate the inverse of the mixing matrix A, using only the measured spectra, and then compute the independent components. Then, the spectra are reconstructed using the mixing matrix A and the independent components, X, not related with the water artifact.

2. **COMPUTING INDEPENDENT COMPONENTS**

The techniques to compute the independent components in a signal are based on the second order and/or the higher order statistics of the data [2]. In what concerns the second order statistics methods, the generalized eigenvalue decomposition (GEVD) of a matrix pencil comprising a pair of matrices will be formulated using the definition of congruent pencils [4]. The generalized eigendecomposition approach to independent component analysis considers a matrix pencil, (R\textsubscript{x1}, R\textsubscript{x2}), computed on the mixed signals. The matrices of the pencil have the following property

\[
R_{x1} = A \Lambda_{x1} A^T \quad \text{and} \quad R_{x2} = A \Lambda_{x2} A^T \tag{1}
\]

where A is the instantaneous mixing matrix and \( \Lambda_{x1}, i \in \{1, 2\} \), are diagonal matrices. Then, the eigenvector matrix of the GEVD of the mixed pencil provides an estimate of the inverse of the mixing matrix. This solution is possible because the mixed pencil, (R\textsubscript{x1}, R\textsubscript{x2}), and the independent components pencil, (\( \Lambda_{x1}, \Lambda_{x2} \)), are related as described by eqn (1). In particular if A is an invertible matrix the two pencils are called congruent pencils [5]. Congruent pencils have identical eigenvalues as can be easily shown writing...
the characteristic polynomial of the mixed pencil
\[
\chi(\lambda) = \det(R_{x1} - \lambda R_{x2}) = \det(A) \det(\Lambda_{s1} - \lambda \Lambda_{s2}) \det(A^T) = 0
\] (2)
which has the same roots as the characteristic polynomial of the source matrix pencil \((\Lambda_{s1}, \Lambda_{s2})\).

Using the GEVD statement of the mixed pencil
\[
R_{x1}E = R_{x2}ED
\] (3)
where \(E\) is the eigenvector matrix and it will be an unique matrix (with the columns normalized to unity length) if the diagonal matrix \(D\) has distinct eigenvalues, \(\lambda_i\). Otherwise the eigenvectors which correspond to the same eigenvalue might be substituted by their linear combinations without affecting the previous equality. So, supposing that the diagonal elements of \(D\) are all distinct, equation (3) can be written as
\[
\Lambda_{s1} A^T E = \Lambda_{s2} A^T E D
\] (4)
if \(A\) is an invertible matrix, we can multiply both sides of the equality by \(A^{-1}\) and setting
\[
E_s = A^T E
\] (5)
leads to the eigendecomposition of the source pencil because the eigenvalues are the same as proved in eqn 2
\[
\Lambda_{s1} E_s = \Lambda_{s2} E_s D
\] (6)
where \(E_s\) is the eigenvector matrix. In equation (5) each column of \(E_s\) is related to a column of \(E\) by the transpose of the mixing matrix. Then, the normalized eigenvectors for a particular eigenvalue are related by \(e_s = \alpha A^T e\) where \(\alpha\) is the constant that normalizes, to unit length, the eigenvectors. Then eigenvector matrix \(E\) will be an approximation to the inverse of the mixing matrix, if the \(E_s\) is the identity matrix (or a permutation). This is a fact when the matrices \(\Lambda_{si}\) are both diagonal as it was assumed. Consequently the independent components can be computed multiplying the measured signals by the transpose of the eigenvector matrix. Those results can also be proved when the mixing matrix is \(m \times n\), with \(m > n\), i.e. when the number of measured signals is higher than the number of independent components, as described in[4].

The matrix pencil can be calculated using distinct strategies (see for example [6],[7],[8]). We will use the data and a linear filtered version of the data to compute correlation matrices in a very similar way as described in [9] and [10].

3. RESULTS AND DISCUSSION

FID’s \(S(t_{1,j}, t_2)\) recorded at fixed evolution times \(t_{1,j}\) were sampled over time spans \(t_2\) and Fourier transformed is computed. Data matrices have been formed with one row representing one single spectrum corresponding to a fixed evolution time \(t_{1,j}\). The final matrix, \(S(\omega_2, t_1)\), then contained as many rows \(j\) as there were different evolution times \(t_{1,j}\) according to the experimental protocol. Typically \(j = 512\) evolution periods have been considered and \(N = 2048\) data points of each spectrum were sampled. However due to phase cycling only 128 spectra have been considered at most hence data matrices of size \((128 \times 2048)\) have been used finally.

Two data sets will be considered in the following comprising either a 2D NOESY spectrum of a simple solute (EDTA) dissolved in water or a corresponding spectrum of the protein P11. Presaturation of the water resonance has been applied in all cases.

A matrix pencil is first computed from the data matrices. The demixing matrix is estimated then and used to estimate the independent components (ICs). Those ICs showing spectral energy in the frequency range of the water resonance have been related with the water artifact and have been set to zero deliberately. Then the remaining spectrum has been reconstructed with the estimated inverse of the demixing matrix and the corrected matrix of estimated source signals.

3.1. Computing the matrix pencil

The matrix pencil \((R_{x1}, R_{x2})\) of zero mean data comprises two correlation matrices of the data. The first matrix is computed as follows:
\[
R_{x2} = \frac{1}{N} S(\omega_2, t_1) S^{H}(\omega_2, t_1)
\] (7)
with \(N = 2048\) representing the number of samples in the \(\omega_2\) domain and \(S^{H}\) the conjugate transpose of the matrix \(S\). The second correlation matrix \(R_{x1}\) of the pencil has been computed after filtering each single spectrum (each row of \(S(\omega_2, t_1)\)) by a function, \(h(\omega_2)\) that modifies the spectra shapes. Then if \(H\) is a matrix with all rows equal to the samples of \(h(\omega_2)\), the second matrix can be computed as
\[
R_{x1} = \frac{1}{N} [S(\omega_2, t_1) \diamond H](H^{T} \diamond S^{H}(\omega_2, t_1)]
\] (8)
where \(\diamond\) is the Hadamard product. The mixed pencil has the property described in eqn 1 assuming a linear mixing model, i.e, \(S(\omega_2, t_1) = AX(\omega_2, t_1)\), where \(X(\omega_2, t_1)\) are the independent components.

3.2. EDTA Spectra

First 2D NOESY spectra of simple solute molecules like EDTA (Ethylene diamine - N,N,N',N' - tetraacetic acid) have been analyzed as a test case. The matrix pencil constructed as described and in particular the second matrix of the pencil has been formed with signals that have been filtered using the following simple filter
10
8
6
4
2
0
-1
-2
-3
-4
-5
0
2
4
6
8
10
12
14 x 10^7

Fig. 1. 1D slice of a 2D NOESY spectrum of EDTA in aqueous solution corresponding to the shortest evolution period \( t_1 \). \( \delta \) denotes a normalized frequency shift, called chemical shift, relative to the resonance of a standard compound. This chemical shift ranges from \(-1 \text{ ppm}\) to \( 10 \text{ ppm}\).

\[
h(\omega_2) = \begin{cases} 
0 & \text{if } 10.759 \leq \delta \leq 3.250 \text{ ppm} \\
1 & \text{else} 
\end{cases}
\]  

i.e. that part of the spectrum containing the water resonance has been removed completely. The demixing matrix is the transpose of the eigenvector matrix of the pencil and the independent components are then estimated. It turned out that roughly 25 ICs had to be assigned to the water resonance and 4-5 ICs to the EDTA signals. A typical 1D EDTA spectrum \( S(t_1, \omega_2) \) is shown in Fig. (1) illustrating the still intense water artifact around a normalized frequency shift of 4.8 ppm relative to the resonance frequency of the standard. Fig.(2) presents the reconstructed spectrum with the water artifact removed as discussed above. The small distortions remaining are due to baseline artifacts caused by truncating the FID due to limited sampling times.

An equally good result can be obtained when working in the time domain. To do so the FIDs were first Fourier transformed into the frequency domain to effect a phase correction then the data have been filtered by a Gaussian band-pass filter \( h(\omega_2) \) centered at the water resonance with a half width of \( \sigma = 2 \). Then the data have been inversely Fourier transformed back into the time domain and the matrices of the pencil have been formed with the original data and the filtered versions of them. A simultaneous diagonalization of the matrix pencil generated quite comparable results to those obtained with the frequency domain data.

### 3.3. Protein spectra

Next the algorithm has been applied to a 2D NOESY spectrum of an aqueous solution of a synthetic polypeptide P11 which is identical to the H11 helix of the human Glutathione reductase and consists of 24 amino acid residues [11]. The data have been treated in the frequency domain. The second correlation matrix of the pencil has been created with filtering the complex valued spectra with a Gaussian band-pass filter \( h(\omega_2) \) centered at the water resonance at 4.8 ppm. The half width of the filter turned out not to be critical and has been chosen to \( \sigma = 1 \). All 128 spectra \( S(t_1, \omega_2), i = 1, \ldots, 128 \) have been considered hence the demixing matrix had dimension \( 128 \times 128 \). Of the 128 estimated ICs again roughly 25 components had to be assigned to the water resonance. Setting these ICs deliberately to zero during the reconstruction process an almost perfect separation of the water artifact resulted. Also all baseline artifacts could be removed as well which means that these distortions also could be separated into different independent components.

### 4. CONCLUSIONS

We have shown that independent components analysis can be an useful tool to remove water solvent artifacts in pro-
tein spectra. In particular methods based on the GEVD of a matrix pencil proved to be fast and efficient (analogous results were achieved with other protein solutions). Other algorithms were also used (FastICA and SOBI) to compute the independent components and their results were less effective than the ones discussed here. Further investigations will have to answer the question if solute resonances hidden underneath the water resonance can be made visible with these or related techniques. Methods to automatically identify the components related to the water resonance need to be studied also. Further quantitative measures to compare different separation results need to be developed as well.

5. REFERENCES


