Kernel-PCA Denoising of Artifact-free Protein NMR Spectra

K. Stadlthanner , E.W. Lang
P. Gruber, F. J. Theis,
Institute of Biophysics
University of Regensburg
D - 93040 Regensburg, Germany,
email: elmar.lang@biologie.uni-regensburg.de

A.M. Tomé, A. R. Teixeira
Dept. de Electrónica e
Telecomunicações / IEETA,
Universidade de Aveiro
P - 3810 - 193 Aveiro, Portugal,
email: ana@ieeta.pt

C. G. Puntonet
Dept. Arquitectura y
Técnologia de Computadores
Universidad de Granada / ESII
E - 18071 Granada, Spain
email: carlos@atc ugr.es

***Abstract*— Multidimensional [1]H NMR spectra of biomolecules dissolved in light water are contaminated by an intense water artifact. Generalized eigenvalue decomposition methods using congruent matrix pencils are used to separate the water artefact from the protein spectra. Due to the statistical separation process, however, noise is introduced into the reconstructed spectra. Hence Kernel - based denoising techniques are discussed to obtain noise- and artifact - free 2D NOESY NMR spectra of proteins.**

I. 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 [7], [6]. 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. The latter causes severe baseline distortions and obscures weak signals lying under its skirts.

A two-dimensional NMR time domain signal $\tilde{X}(t_{1,j}, t_2)$ corresponds to a sum of free induction decay (FID) signals sampled for fixed evolution periods $t_{1,j}$ and extending over a sampling time interval of duration t_2 . The evolution period $t_{1,j}$ is incremented during the experiment to yield typically 512 FIDs. Standard signal processing is performed by Fourier analysis, resulting in spectra $X(\omega_1, \omega_2)$ made of sums of Lorentzian shaped resonance lines [6], [5].

ICA techniques extract a set of signals out of a set of measured signals without knowing how the mixing process is carried out [9],[4]. Considering that the set of measured spectra X is a linear combination of a set of independent components S, i.e , $X = AS$. 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 those independent components, S , not related with the water artifact. It has been shown [10], [11] that 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 [6]. Second order techniques using the time structure of the signals turn out to be superior as they do not rely on the assumption of statistically independent source signals.

Note that second order techniques using the time structure of the signals do not need to assume statistically independent source signals but only uncorrelated signals. This is a much weaker assumption which is more appropriate to deal with NMR spectra. With the latter the independence assumption is hard to justify in general. Hence blind source separation algorithms based on a generalized eigenvalue decomposition using matrix pencils [16] were used to effect the separation of the water artifact [11].

Unfortunately the statistical separation process by necessity introduces additional noise not present in the original spectra. Hence denoising as a postprocessing of the artifact-free spectra is necessary. It is important that the denoising does not change the spectral characteristics like peak volumes as the deduction of the 3D structure of the proteins heavily relies on the latter. Kernel - PCA based denoising techniques [3] have been shown to be very efficient outperforming linear PCA. Kernel PCA actually generalizes linear PCA which hitherto has been used for denoising. PCA denoising follows the idea that retaining only the principal components with highest variance to reconstruct the decomposed signal noise contributions which should correspond to the low variance components can deliberately been omitted hence reducing the noise contribution to the observed signal. Kernel PCA extends this idea to non-linear signal decompositions. The idea is to project observed data non-linearly into a high-dimensional feature space and then to perform linear PCA in feature space. The trick is that the whole formalism can be cast into dot product form hence the latter can be replaced by suitable kernel functions to be evaluated in the lower dimensional input space instead of the high-dimensional feature space. Denoising then amounts to estimating appropriate pre-images in input space of the nonlinearly transformed signals. Kernel PCA based techniques [2] will thus be discussed to try to obtain almost

0-7803-8359-1/04/$20.00 ©2004 IEEE

noise- and artifact - free protein spectra.

II. The Generalized Eigenvalue Decomposition Approach

For convenience we shortly review the generalized eigenvalue decomposition (GEVD) approach using congruent matrix pencils [14], [15], [12]. The generalized eigenvalue decomposition approach to the blind source separation problem, based on the model $\vec{x}(t) = \mathbf{A}\vec{s}(t)$, considers a matrix pencil, $(\mathbf{R}_{\vec{x},1}, \mathbf{R}_{\vec{x},2})$ computed on the sensor signals $\vec{x}(t)$ (henceforth called sensor pencil). The correlation matrices $\mathbf{R}_{\vec{x}} = \langle \vec{x}\vec{x}^T \rangle$ of the pencil have the following property

$$\mathbf{R}_{\vec{x},i} = \mathbf{A}\mathbf{\Lambda}_{\vec{s},i}\mathbf{A}^T \qquad i = 1,2 \tag{1}$$

where $\mathbf{A}$ is the instantaneous but unknown mixing matrix and $\mathbf{\Lambda}_{\vec{s},i}$ are diagonal matrices related to the unknown source signals $\vec{s}(t)$. Then the eigenvector matrix of the generalized eigenvalue decomposition of the sensor pencil provides an estimate of the inverse (or pseudo-inverse) of the mixing matrix. This solution is possible because the sensor pencil, $(\mathbf{R}_{\vec{x},1}, \mathbf{R}_{\vec{x},2})$ and the source pencil $(\mathbf{\Lambda}_{\vec{s},1}, \mathbf{\Lambda}_{\vec{s},2})$ are related as described by eqn (1). In particular if $\mathbf{A}$ is an invertible matrix the two pencils are called congruent pencils. Congruent pencils have identical eigenvalues as can be easily shown writing the characteristic polynomial of the sensor pencil

$$\chi_{\vec{x}}(\lambda) = \det(\mathbf{R}_{\vec{x},1} - \lambda\mathbf{R}_{\vec{x},2}) = 0 \tag{2}$$

If $\mathbf{A}$ represents an invertible matrix then it follows

$$\det(\mathbf{R}_{\vec{x},1} - \lambda\mathbf{R}_{\vec{x},2}) = \det(\mathbf{A})\det(\mathbf{\Lambda}_{\vec{s},1} - \lambda\mathbf{\Lambda}_{\vec{s},2})\det(\mathbf{A}^T) \tag{3}$$

which has the same roots as the related characteristic polynomial of the source pencil

$$\chi_{\vec{s}}(\lambda) = \det(\mathbf{\Lambda}_{\vec{s},1} - \lambda\mathbf{\Lambda}_{\vec{s},2}) \tag{4}$$

The generalized eigenvalue decomposition statement of the sensor pencil

$$\mathbf{R}_{\vec{x},2}\mathbf{E} = \mathbf{R}_{\vec{x},1}\mathbf{E}\mathbf{D} \tag{5}$$

where $\mathbf{E}$ is the eigenvector matrix which will be an unique matrix (with the columns normalized to unit length) if the diagonal matrix $\mathbf{D}$ has distinct eigenvalues λ_i^D. Otherwise the eigenvectors which correspond to the same eigenvalue might be substituted by their linear combinations without affecting the previous equality. Supposing that the diagonal elements λ_i^D of $\mathbf{D}$ are all distinct equation (4) can be written as

$$\mathbf{A}\mathbf{\Lambda}_{\vec{s},2}\mathbf{A}^T\mathbf{E} = \mathbf{A}\mathbf{\Lambda}_{\vec{s},1}\mathbf{A}^T\mathbf{E}\mathbf{D} \tag{6}$$

If $\mathbf{A}$ is an invertible matrix, we can multiply both sides of the eqn.(5) by $\mathbf{A}^{-1}$ and setting $\mathbf{E}_{\vec{s}} = \mathbf{A}^T\mathbf{E}$ leads to the eigenvalue decomposition of the source pencil

$$\mathbf{\Lambda}_{\vec{s},2}\mathbf{E}_{\vec{s}} = \mathbf{\Lambda}_{\vec{s},1}\mathbf{E}_{\vec{s}}\mathbf{D} \tag{7}$$

where $\mathbf{E}_{\vec{s}}$ is the eigenvector matrix. Hence each column of $\mathbf{E}_{\vec{s}}$ is related to a column of $\mathbf{E}$ by the transpose of the mixing matrix. Then the normalized eigenvectors corresponding to a particular eigenvalue are related by $\vec{e}_{\vec{s}} = \alpha\mathbf{A}^T\vec{e}$ where α is a constant that normalizes, to unit length, the eigenvectors. Concerning blind source separation problems the eigenvector matrix $\mathbf{E}$ will provide an approximation to the inverse of the mixing matrix, if the eigenvector matrix $\mathbf{E}_{\vec{s}}$ represents the identity matrix (or a permutation). The latter holds true if, as was assumed, the source pencil is formed with diagonal matrices $\mathbf{\Lambda}_{\vec{s},i}$. Tomé [15] also presented an algebraic formulation of the GEVD problem using the notion of congruent matrix pencils and block matrix operations when the mixing matrix has more rows than columns.

III. Kernel - PCA based denoising

In principal component analysis (PCA) [13] a data vector is represented in an orthogonal basis system such that the projected data have maximal variance. The orthogonal transformation is obtained by diagonalizing the centered covariance matrix of the data set $\{\vec{x}_k \in \mathcal{R}^N \quad (k = 1, ..., l)\}$ defined by

$$\mathbf{C} = \langle (\vec{x}_i - \langle \vec{x}_k \rangle)(\vec{x}_i - \langle \vec{x}_k \rangle)^T \rangle \tag{8}$$

The coordinates of the eigenvector basis are called *principal components* and represent linear features of the data set. The eigenvalue μ_i corresponding to an eigenvector $\vec{c_i}$ of $\mathbf{C}$ equals the variance of the data in the direction of $\vec{\omega}$. Hence the first n eigenvectors$\{\vec{c_i}\}, i = 1, ..., n$ corresponding to the n largest eigenvalues $\mu_1 > \mu_2 > ..... > \mu_n$ cover as much variance as is contained in n orthogonal directions. In denoising applications directions with small (subject to a given criterium) variance are deliberately dropped. However PCA only extracts linear features though with suitable *nonlinear* features more information could be extracted.

It has been shown [3] that Kernel PCA is well suited to extract interesting nonlinear features in the data. Kernel PCA first maps the data $\{\vec{x}_i\}$ into some high-dimensional feature space Ω through a nonlinear function $\Phi(\vec{x}_i)$ and then performs linear PCA on the mapped data in the feature space Ω. Assuming centered data in feature space, i.e. $\sum_{k=1}^{l}\Phi(\vec{x}_k) = 0$ to perform PCA in space Ω means finding the eigenvalues $\lambda > 0$ and eigenvectors $\vec{\omega} \in \Omega$ of the covariance matrix $\bar{\mathbf{R}} = \frac{1}{l}\sum_{j=1}^{l}\Phi(\vec{x}_j)\Phi(\vec{x}_j)^T$ satisfying the eigenequation

$$\bar{\mathbf{R}}\vec{\omega} = \lambda\vec{\omega} \tag{9}$$

Note that all solutions $\vec{\omega}$ with $\lambda \neq 0$ lie in the span of $\Phi(\vec{x}_1), ..., \Phi(\vec{x}_l)$ hence the eigenvectors can be represented via

$$\vec{\omega} = \sum_{i=1}^{l} \alpha_i \Phi(\vec{x}_i) \tag{10}$$

Multiplying eqn.(9) with $\Phi(\vec{x}_k)$ from the left the following modified eigen-equation is obtained

$$\mathbf{K}\vec{\alpha} = l\lambda\vec{\alpha} \tag{11}$$

with $\lambda > 0$ and coefficient eigenvectors $\vec{\alpha} = (\alpha_1, ..., \alpha_l)^T$. With normalized eigenvectors $\vec{\omega} \in \Omega$ obeying $\vec{\omega}_k \cdot \vec{\omega}_k = 1$ the corresponding coefficient eigenvectors are also normalized obeying $\lambda_k(\vec{\alpha}_k \cdot \vec{\alpha}_k) = 1$. The eigenequation now is cast in the form of dot products occuring in feature space through the $l \times l$ matrix $\mathbf{K}$ with elements $K_{ij} = (\Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)) = k(\vec{x}_i, \vec{x}_j)$ which are represented by kernel functions $k(\vec{x}_i, \vec{x}_j)$ to be evaluated in the input space. The principal components of the projected data vectors in feature space are then given by

$$\beta_k = (\vec{\omega}_k \cdot \Phi(\vec{x})) = \sum_{i=1}^{l} \alpha_i^k (\Phi(\vec{x}_i) \cdot \Phi(\vec{x})) = \sum_{i=1}^{l} \alpha_i^k k(\vec{x}_i, \vec{x}) \tag{12}$$

For feature extraction any suitable kernel can be used and knowledge of the nonlinear function $\Phi(\vec{x})$ is not needed. Note that the latter can always be reconstructed from the principal components obtained. The image of a data vector under the map Φ can be reconstructed from its projections β_k via

$$\hat{P}_n \Phi(\vec{x}) = \sum_{k=1}^{n} \beta_k \vec{\omega}_k \tag{13}$$

which defines the projection operator $\hat{P}_n$. If n is deliberately chosen small as is done in de-noising then the squared reconstruction error

$$e_{rec}^2 = \sum_{i=1}^{l} \| \hat{P}_n \Phi(\vec{x}_i) - \Phi(\vec{x}_i) \|^2 \tag{14}$$

is still minimal. To find a corresponding approximate representation of the data in input space it is necessary to estimate a $\vec{z} \in \mathbf{R}^N$ such that

$$\rho(\vec{z}) = \| \hat{P}_n \Phi(\vec{x}) - \Phi(\vec{z}) \|^2 \tag{15}$$

is minimized. In de-noising applications it is hoped that the deliberately neglected dimensions of minor variance contain noise mostly and $\vec{z}$ represents a de-noised version of $\vec{x}$. Eqn. (15) can be minimized via gradient descent techniques using the cost function

$$\rho(\vec{z} = k(\vec{z}, \vec{z}) - 2\sum_{k=1}^{n} \beta_k \sum_{i=1}^{l} \alpha_i^k k(\vec{x}_i, \vec{z}) \tag{16}$$

which is obtained by subtituting into eqn.(15) the corresponding expressions. Note that instead of using an iterative gradient descent technique an analytic solution to the pre-image problem has been given recently in case of invertible kernels [1].

IV. Water Artifact Removal and Denoising of Protein Spectra

FID's $X(t_{1,j}, t_2)$ recorded at fixed evolution times $t_{1,j}$ were sampled over time spans t_2 and Fourier transformed with respect to both time domains to obtain corresponding spectra $X(\omega_1, \omega_2)$ which could be corrected for any phase distortions. Data matrices have been formed with one row representing one single spectrum corresponding to a fixed evolution time $t_{1,j}$. The final matrix, $\mathbf{X}(\omega_2, t_1)$, 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 were sampled of each spectrum. However due to phase cycling only every fourth of the 512 spectra, hence 128 spectra, have been considered at most hence four groups of data matrices of size (128×2048) with identical phase have been used finally.

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 only have been related with the water artefact and have been set to zero deliberately. Then the protein spectrum has been reconstructed with the estimated inverse of the demixing matrix and the corrected matrix of estimated source signals.

Kernel PCA denoising has then been applied to the reconstructed artifact- free protein spectra to remove the additional noise introduced by the statistical separation procedure. The real and imaginary part of the complex nmr spectral data have been divided into several groups and have been projected then in a 1024-dim feature space where 1024 principal components have been determined. Denoising was achieved by projecting the data onto a lower dimensional subspace and estimating an approximate pre-image in input space.

A. Computing the demixing matrix for artifact separation

The matrix pencil $(\mathbf{R}_{\vec{x},1}, \mathbf{R}_{\vec{x},2})$ of zero mean data $\mathbf{X} = [\vec{x}_1\ \vec{x}_2\ \ldots\ \vec{x}_J]^T$ comprises two correlation matrices of the data. The first matrix is computed as follows

$$\mathbf{R}_{\vec{x},1} = \frac{1}{N}\mathbf{X}(\omega_2, t_1)\mathbf{X}^H(\omega_2, t_1) \tag{17}$$

with $N = 2048$ representing the number of samples in the ω_2 domain and $\mathbf{X}^H$ the conjugate transpose of the matrix

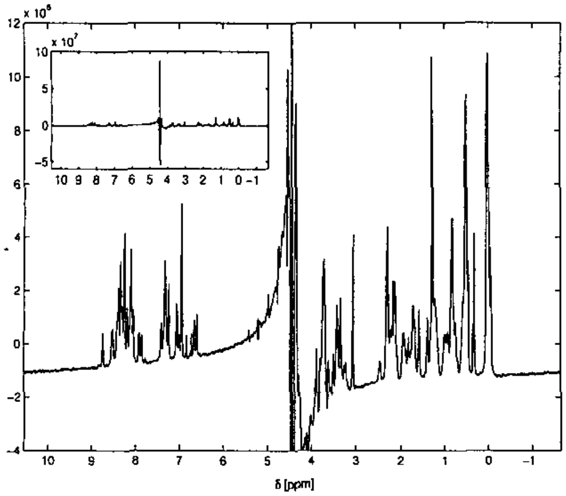

Fig. 1. 1D slice of a 2D NOESY spectrum of the polypeptide p11 in aqueous solution corresponding to the shortest evolution period t_1. The chemical shift ranges from $-1ppm$ to $10ppm$ roughly

$\mathbf{X}$. The second correlation matrix $\mathbf{R}_{\bar{x},2}$ of the pencil has been computed after filtering each single spectrum (each row of $\mathbf{X}(\omega_2, t_1)$) by a function, $h(\omega_2)$ that modifies the spectral shape. Then if H is a matrix with all rows equal to the samples of $h(\omega_2)$ the second matrix can be computed as

$$\mathbf{R}_{\bar{x},2} = \frac{1}{N}[(\mathbf{X}(\omega_2, t_1) \diamond H)(H^T \diamond \mathbf{X}^H(\omega_2, t_1)] \quad (18)$$

where $\diamond$ represents the Hadamard product. A bandpass filter of Gaussian shape centered in the spectrum with a variance of $2 \leq \sigma^2 \leq 4$ has been used in the applications discussed below. Both matrices of the pencil are of dimension 128×128 as we assume as many sources as there are sensor signals.

Substituting this result into the GEVD statement (5) the later can be solved in two steps by reducing it to two standard eigenvalue decomposition (EVD) problems. Starting with

$$\mathbf{R}_{\bar{x},1} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^{\mathbf{T}} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^{\mathbf{T}}\mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^{\mathbf{T}} = \mathbf{W}\mathbf{W} \quad (19)$$

and defining $\mathbf{Z} = \mathbf{W}\mathbf{E}$ yields the transformed equation $\mathbf{C}\mathbf{Z} = \mathbf{Z}\mathbf{D}$ which is of the standard EVD form of a real symmetric matrix $\mathbf{C} = \mathbf{W}^{-1}\mathbf{R}_{\bar{x},2}\mathbf{W}^{-1}$ with $\mathbf{W}^{-1} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^T$. The eigenvalues of the matrix pencil are available from the solution of the EVD of matrix $\mathbf{C}$ while the corresponding eigenvectors are obtained via $\mathbf{E} = \mathbf{W}^{-1}\mathbf{Z}$.

B. Water artifact removal in protein spectra

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 [8]. Fig. (1) shows a 1-dim slice of the 2D NOESY spectrum of p11. Note that the vertical scale has been expanded to show the protein peaks clearly. The insert in Fig. (1) shows the full water resonance at 4.8 ppm demonstrating its dominance compared to the protein resonances. 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 bandpass filter $h(\omega_2)$ centered at the water resonance at $4.8ppm$. The half width of the filter turned out not to be critical and has been chosen to $\sigma = 1$. All 128 spectra $X(t_{1i}, \omega_2), i = 1, ..., 128$ of each group with identical phase of the excitation pulse have been considered hence the demixing matrix had dimension 128×128. Of the 128 estimated ICs 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 as can be seen in Fig. (2). The latter shows the reconstructed spectrum of the protein p11 on an identical scale as the original spectrum shown in Fig. (1). Also almost all baseline artifacts could be removed as well (see Fig. (2)) which means that these distortions also could be separated into different independent components.

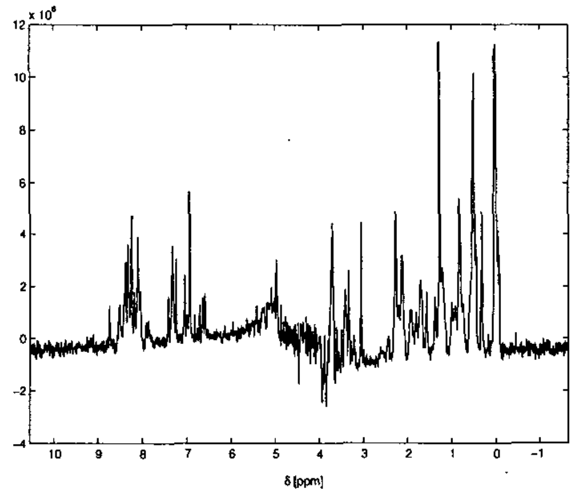

Fig. 2. Reconstructed p11 spectrum corresponding to Fig.1 with the water artifact removed with the matrix pencil algorithm

C. Denoising of the reconstructed artifact-free spectra

As the removal of the water artifact lead to additional noise in the spectra (compare Fig. (1) and Fig. (2)) kernel PCA based denoising was applied. First (almost) noise free samples had to be created in order to determine the principle axes in feature space. For that purpose the first 400 data points of the real and the imaginary part of each of the 512

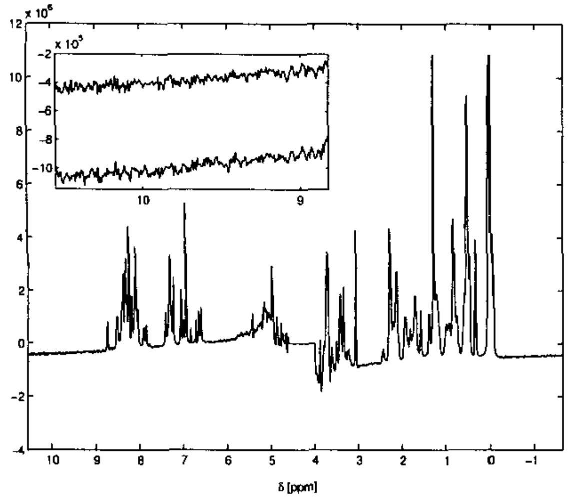

Fig. 3. Result of the kernel PCA denoising of a reconstructed p11 spectrum corresponding to Fig.2. The insert shows the region of the spectrum between 10 and 9 ppm roughly. The upper trace corresponds to the denoised baseline and the lower trace shows the baseline of the original spectrum.

original spectra were used to form a 400×1024 sample matrix $\mathbf{Q}^{(1)}$. Likewise five further sample matrices $\mathbf{Q}^{(m)}$, $m = 2, ..., 6$, were created which now consisted of the data points 401 to 800, 601 to 1000, 1101 to 1500, 1249 to 1648 and 1649 to 2048 respectively. Note that the region (1000 - 1101) of data points comprising the main part of the water resonance was nulled deliberately as it is of no use for the kernel PCA. For each of the sample matrices $\mathbf{Q}^{(m)}$ the corresponding kernel matrix $\mathbf{K}$ was determined by

$$\mathbf{K}_{i,j} = k(\vec{x}_i, \vec{x}_j), \quad i, j = 1, ..., 400, \tag{20}$$

where $\vec{x}_i$ denotes the i-th column of $\mathbf{Q}^{(m)}$. For the kernel function the Gaussian kernel

$$k(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2\sigma^2}\right) \tag{21}$$

where

$$2\sigma^2 = \frac{1}{1024 * 1023} \sum_{i,j=1}^{1024} \| \vec{x}_i - \vec{x}_j \|^2$$

is the width parameter σ, was chosen.

Finally the kernel matrix $\mathbf{K}$ was expressed in terms of its EVD (cf. (11)) which lead to the expansion parameters $\vec{\alpha}$ necessary to determine the principal axes of the corresponding featurespace $\Omega^{(m)}$:

$$\vec{\omega} = \sum_{i=1}^{1024} \alpha_i \Phi(\vec{x}_i).$$

Exactly as the original data the noisy data of the reconstructed spectra were used to form six 400×1024 dimensional pattern matrices $\mathbf{P}^{(m)}$, $m = 1, ..., 6$. Then the principal components β_k of each column of $\mathbf{P}^m$ were calculated in the corresponding feature space $\Omega^{(m)}$. In order to denoise the patterns only projections onto the first $n = 112$ principal axes were considered. This lead to

$$\beta_k = \sum_{i=1}^{1024} \alpha_i^k k(\vec{x}_i, \vec{x}), \quad k = 1, ..., 112,$$

where $\vec{x}$ is a column of $\mathbf{P}^m$.

After reconstructing the image $\hat{P}_n \Phi(\vec{x})$ of the sample vector under the map Φ (cf. (13)) its approximate pre-image was determined by minimizing the cost function

$$\rho(\vec{z}) = -2 \sum_{k=1}^{112} \beta_k \sum_{i=1}^{1024} \alpha_i^k k(\vec{x}_i, \vec{z}).$$

Note that the method described above fails to denoise the region where the water artefact has its peak (datapoints 1001 to 1101) because then the samples formed from the original data differ too much from the noisy data. This is not a major drawback as protein peaks totally hidden under the water artifact cannot be uncovered by the presented blind source separation method. Fig. (3) shows the resulting denoised protein spectrum on an identical vertical scale as Fig. (1) and Fig. (2). The insert compares the noise in a region of the spectrum between 10 and 9 ppm roughly where no protein peaks are found. The upper trace shows the baseline of the denoised reconstructed protein spectrum and the lower trace the corresponding baseline of the original experimental spectrum before the water artefact has been separated out. NOte that the peak intensities are not influenced by the denoising procedure.

V. Conclusions

Multidimensional NMR spectra form an indispensable tool to elucidate the native 3D structure of proteins in their natural aqueous environment. Proton NMR spectra are generally dominated by a huge water resonance contaminating the spectrum of the protein. We have shown that blind source separation algorithms using the time-structure of the signals and relying on second order statistics only can be an useful tool to remove water solvent artifacts in protein spectra. In particular methods based on the GEVD of a matrix pencil proved to be fast and efficient. However these statistical artefact separation methods inevitably introduce additional noise into the spectra not present in the experimental spectra. Hence denoising methods are necessary as a postprocessing of the artifact-free spectra to remove the additional noise. We have shown that kernel PCA based de-noising techniques are well suited to reduce the noise to a level comparable to or below the noise level of the experimental spectra. This is important as the structure determination relies on the integral intensities of the resonances to deduce distance information. Further experiments will have to corroborate these findings on a larger set of protein spectra and to compare the results with other nonlinear de-noising procedures. Finally artifact separation and de-noising need to be automatized to provide a preprocessing tool to NMR spectroscopists starting to analyze these spectra and deduce structural information fast and reliably.

VI. Acknowledgements

This research has been supported by the DFG (GRK 638: Nonlinearity and nonequilibrium in condensed matter) and the BMBF (project: ModKog).

References

[1] J. T. Kwok, I. W. Tsang, The Pre-Image Problem in Kernel Methods, Proceed. Int. Conf. Machine Learning (ICML03) (2003)

[2] S. Mika, B. Schölkopf, A. Smola, K. Müller, M. Scholz, G. Rätsch, Kernel PCA and denoising in feature spaces, Advanc. Neural Inform. Process. Systems **11**, (1998)

[3] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Neural Computation, **10**, 1299 - 1319, (1998)

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

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

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

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

[8] A.Nordhoff and Ch.Tziatzios and J.A.V. Broek and M.Schott and H.-R.Kalbitzer and K.Becker and D.Schubert and R.H.Schirme, Denaturation and reactivation of dimeric human glutathione reductase, Eur. J. Biochem. 273 - 282, (1997)

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

[10] K.Stadlthanner, A.M.Tomé, F.J.Theis, W.Gronwald, H.-R.Kalbitzer, E.W.Lang, Blind Source Separation of Water Artefacts in NMR Spectra Using a Matrix Pencil, Proc. 4^{th} Int. Conf. on Independent Component Analysis and Signal separation (ICA'2003), Nara, Japan, 167 - 172, (2003)

[11] K.Stadlthanner, F.J.Theis,E.W.Lang, A.M.Tomé W.Gronwald, H.-R.Kalbitzer, A matrix pencil approach to the blind source separation of artifacts in 2D NMR spectra, Neural Information Processing - Letters and Reviews, 1(3), 103 - 110, (2003)

[12] L. Parra, P. Sajda, Blind Source Separation via Generalized Eigenvalue Decomposition, J- Machine Learn. Research **4**, 1261 - 1269, (2003)

[13] K. I. Diamantaras, S. Y. Kung, Principal Component Networks, Wiley, New York, 1996

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

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

[16] A. M. Tomé, Separation of a mixture of signals using linear filtering and second order statistics, Proceed. European Symposium on Artificial Neural Networks (ESANN'2002), pp.307 - 312, (2002)