

# Novel Postprocessing Methods to Remove the Water Resonance from Protein NMR Spectra

DISSERTATION ZUR ERLANGUNG DES  
DOKTORGRADES DER NATURWISSENSCHAFTEN  
(DR. RER. NAT.) DER NATURWISSENSCHAFTLICHEN  
FAKULTÄT II - PHYSIK DER UNIVERSITÄT  
REGENSBURG

vorgelegt von

Kurt Stadlthanner

aus Regensburg

2007

Promotionsgesuch eingereicht am: 4. Juli 2007  
Die Arbeit wurde angeleitet von: Prof. Dr. Elmar Lang

Prüfungsausschuss:

Vorsitzender: Prof. Dr. J. Zweck

1. Gutachter: Prof. Dr. E. W. Lang
2. Gutachter: Prof. Dr. A. M. Tomé
3. Gutachter: Prof. Dr. I. Morgenstern

to Natacha



# Summary

NMR spectroscopy is one of the most popular tools used in the spatial structure determination of proteins. It owes much of its popularity to the fact that it is the only method with which proteins can be investigated under quasi-physiological conditions. However, if the behavior of the  $^1H$ -protons of the proteins is studied in an NMR experiment a dominant water signal is observed which exacerbates the automated analysis of the recorded data considerably.

The water resonance appears as the proteins under investigation are usually dissolved in water in order to analyze them under quasi-physiological conditions. In such experiments the concentration of the proteins in water is usually very low such that the recorded  $^1H$ -NMR spectra contain mostly the resonance signal of the water protons while the resonances of the protein protons can hardly be resolved. In the past, experimental methods have been developed with which the water resonance can be reduced such that a quantitative analysis of the protein signals is possible. However, even if these methods are applied the water resonance remains as the largest peak in the spectrum, overlaps neighboring protein peaks and leads to severe baseline distortions.

In this thesis digital signal processing is applied to remove the remaining water signal from the NMR spectra. First, the concept of NMR spectroscopy is reviewed and the main features of the water signal are discussed. Based on this introductory chapter it is shown how Blind Source Separation (BSS) can be used advantageously to remove the water signal from protein spectra. In this approach the obtained NMR spectra are decomposed into protein and water related components. The latter are then neglected and the protein components are reassembled such that pure, water free protein spectra result. For the BSS step a second order algorithm is used which is based on the generalized eigenvalue decomposition (GEVD) of the covariance matrices of the original and of filtered observation signals, respectively.

A drawback of the BSS approach is that it leads to increased noise in the resulting spectra. In order to remove this noise again local and Kernel Prin-

Principal Component Analysis (PCA) denoising are investigated. Both methods map the recorded data into higher dimensional feature spaces which are then divided into signal and noise containing subspaces, respectively, by means of standard PCA. Considering only the signal subspace eventually leads to the desired denoised signals.

In detail, the embedding is carried out in local PCA by means of delayed coordinates. Furthermore, the feature space vectors are first clustered by similarity before the denoising step is carried out. Kernel PCA makes also use of the concept of delayed coordinates but maps the embedded data additionally into an even higher dimensional feature space by means of a nonlinear mapping. In this space standard PCA based denoising is carried out again before the data has to be mapped back to input space. In this procedure the explicit nonlinear mapping is circumvented by making use of the so-called Kernel trick.

The results obtained by both denoising methods are compared quantitatively whereas it turns out that the local PCA approach is superior as it hardly affects the actual signals and better removes the noise.

Both BSS and denoising are carried out simultaneously in the algorithm dAMUSE which is also applied to the NMR data sets. In this algorithm the BSS step is again based on a joint eigenvalue decomposition of covariance matrices and the denoising is achieved by means of delayed coordinates and PCA. Compared with standard BSS combined with local PCA dAMUSE leads to similar results but is more time efficient.

In the context of dAMUSE the algorithm Autoassign is presented with which the estimated water and protein related signals can be detected automatically. Autoassign first estimates the water signal by means of SSA and uses a Genetic Algorithm for the assignment task. For the GA a fixed set of parameters is given with which the global minimum of the used target function can be found reliably and within reasonable computational time.

Finally, also singular spectrum analysis (SSA) is applied to remove the water signal from protein NMR spectra. While the BSS methods lead to better results in frequency domain SSA is applied to the time domain signals (also called free induction decays) recorded throughout the NMR experiment. In SSA these time signals are also embedded by means of delayed coordinates whereupon standard PCA is carried out. In the PCA step the water signal is estimated and then subtracted from the recorded data. This approach has two decisive advantages compared with the BSS based methods: first, protein peaks residing in the immediate vicinity of the water signal remain in the resulting spectra and second, no additional noise appears. Furthermore, SSA is particularly easy to use as only one parameter has to be tuned.

Hence, it is concluded that SSA is the method of choice for removing

the water signal from NMR data. Its robustness is proved by applying it to the data sets of four different proteins, namely HPr, P11, TmCSP, and the RAS-binding domain of the protein RalGEF, respectively. In all these cases SSA removes the water signal almost perfectly while the protein signals in the resulting spectra remain virtually unaltered. A closer investigation of the 2D-spectrum of the protein TmCSP reveals furthermore that SSA is capable of uncovering protein peaks which were formerly hidden by the water resonance. Hence, the SSA approach facilitates not only the automatic analysis of protein spectra by correcting baseline distortions but also unveils new information about the protein under investigation.



# Contents

<b>Summary</b>	<b>v</b>
<b>Introduction</b>	<b>xiii</b>
<b>1 Mathematical Preliminaries</b>	<b>1</b>
1.1 Basic Probability Theory . . . . .	1
1.1.1 Distribution and Density Functions . . . . .	1
1.1.2 Expectations . . . . .	3
1.2 Performance Measures . . . . .	5
1.3 Matrix Pencils and GEVD . . . . .	6
<b>2 NMR Spectroscopy</b>	<b>11</b>
2.1 Basics of NMR . . . . .	11
2.1.1 Magnetic Moment and Nuclear Spin . . . . .	12
2.1.2 Nuclei in External Magnetic Fields . . . . .	12
2.1.3 Bloch Equations . . . . .	14
2.1.4 Chemical Shift . . . . .	17
2.2 Relaxation . . . . .	17
2.2.1 Longitudinal Relaxation . . . . .	18
2.2.2 Transveral Relaxation . . . . .	18
2.2.3 Dipolar Relaxation . . . . .	19
2.2.4 Relaxation by Chemical Shift Anisotropy . . . . .	20
2.2.5 Relaxation by Indirect Spin-Spin Coupling . . . . .	20
2.3 Experimental Methods . . . . .	21
2.3.1 Data Acquisition . . . . .	21
2.3.2 Fourier Transformation . . . . .	23
2.3.3 Spectra Postprocessing . . . . .	24
2.4 3D Structure Determination of Proteins . . . . .	25
2.4.1 Stationary Nuclear Overhauser Effect . . . . .	26
2.4.2 Twodimensional NOE Spectroscopy . . . . .	27
2.5 Conclusions . . . . .	32

<b>3</b>	<b>Water Signal Removal by BSS</b>	<b>33</b>
3.1	The Water Signal . . . . .	33
3.2	Removal of water signal by BSS . . . . .	36
3.2.1	Linear Blind Source Separation . . . . .	36
3.2.2	Linear BSS in the Context of 2D-NOESY Data . . . . .	39
3.3	The Matrix Pencil BSS Algorithm . . . . .	42
3.3.1	Congruent Matrix Pencils in BSS . . . . .	42
3.3.2	The MP-BSS Algorithm . . . . .	45
3.4	Limits of the MP-BSS Algorithm . . . . .	47
3.4.1	Violations of Uncorrelatedness Assumptions . . . . .	47
3.4.2	Violations of the Linear Mixture Model . . . . .	50
3.5	MP-BSS Applied to 2D-NOESY-Data . . . . .	57
3.5.1	MP-BSS Applied to Artificial 2D-NOESY-Data . . . . .	57
3.5.2	MP-BSS Applied to Real World Dataset . . . . .	59
3.6	Conclusions . . . . .	68
<b>4</b>	<b>Denoising of MP-BSS data</b>	<b>69</b>
4.1	Principal Component Analysis . . . . .	69
4.1.1	Concept of PCA Denoising . . . . .	69
4.1.2	PCA Denoising of P11 Spectra . . . . .	73
4.2	Local PCA . . . . .	75
4.2.1	Concept of Local PCA Denoising . . . . .	75
4.2.2	Selection of Parameters . . . . .	78
4.2.2.1	Determination of $U$ . . . . .	79
4.2.2.2	Determination of $D$ , $\Delta t$ , and $K$ . . . . .	81
4.2.3	Local PCA Denoising of P11 Spectra . . . . .	83
4.3	Kernel PCA . . . . .	89
4.3.1	Concept of Kernel PCA . . . . .	90
4.3.1.1	Nonlinear PCA . . . . .	90
4.3.1.2	Centering in Feature Space . . . . .	93
4.3.1.3	Kernel Trick . . . . .	94
4.3.2	Kernel PCA Denoising and Pre-Images . . . . .	96
4.3.2.1	Minimum Feature Space Distance Method . . . . .	97
4.3.2.2	Pre-Images by MDS . . . . .	100
4.3.3	Kernel PCA Denoising of P11 Spectra . . . . .	103
4.3.3.1	Overall Denoising Strategy . . . . .	103
4.3.3.2	Choice of Pre-Image Algorithm . . . . .	105
4.3.3.3	Determination of optimal parameters . . . . .	108
4.4	Conclusions . . . . .	111

<b>5</b>	<b>BSS Comb. Denoising a. Water Assignment</b>	<b>113</b>
5.1	dAMUSE . . . . .	113
5.1.1	Embedding . . . . .	114
5.1.2	Estimation of $\mathbf{A}^{tr}$ and $\mathbf{S}^{tr}$ . . . . .	115
5.1.3	Denoising . . . . .	118
5.1.4	dAMUSE analysis of NMR spectra . . . . .	119
5.2	Automated water assignment . . . . .	120
5.2.1	Autoassign . . . . .	121
5.2.1.1	SSA estimation of the water signal . . . . .	121
5.2.1.2	Assignment of water related sources . . . . .	122
5.2.1.3	GA based optimization . . . . .	123
5.3	dAMUSE Analysis of P11 Spectra . . . . .	125
5.3.1	Water Estimate . . . . .	125
5.3.2	Determination of $D$ , $\Delta t$ , and $\sigma$ . . . . .	125
5.3.3	Estimation of $U$ . . . . .	129
5.4	Conclusions . . . . .	131
<b>6</b>	<b>Water Removal by SSA</b>	<b>135</b>
6.1	SSA Compared to MP-BSS and dAMUSE . . . . .	135
6.2	SSA applied to RalGEF data . . . . .	138
6.3	SSA applied to HPr data . . . . .	141
6.4	SSA applied to TmCSP data . . . . .	142
6.5	Conclusions . . . . .	144
<b>7</b>	<b>Conclusions</b>	<b>151</b>
<b>A</b>	<b>Supplements for MP-BSS</b>	<b>153</b>
A.1	Diagonality of $\mathbf{C}_{\mathfrak{S}}$ . . . . .	153
A.2	Diagonality of the Eigenvector Matrix . . . . .	154
<b>B</b>	<b>Pre-Image Problem</b>	<b>157</b>
<b>C</b>	<b>Genetic Algorithms</b>	<b>159</b>
C.1	Population Representation and Initialisation . . . . .	160
C.2	Fitness Function and Selection Probability . . . . .	161
C.3	Selection . . . . .	162
C.4	Mating . . . . .	162
C.5	Mutation . . . . .	163
C.6	Reinsertion . . . . .	163
C.7	Subpopulations . . . . .	164
C.8	Termination . . . . .	164

**Acknowledgments**

# Introduction

NMR spectroscopy is the state of the art tool used to determine the spatial structure of proteins. Compared with alternative methods like e.g. X-ray analysis it has the decisive advantage that the proteins to be investigated can be dissolved in water which allows their examination under quasi-physiological conditions.

In practice the resonances of the protein protons are typically investigated and used for the spatial structure determination. In these experiments the concentration of the proteins in water is usually very low such that foremostly the signal of the water protons is recorded while any protein signals vanish because of the limited dynamic ranges of the commonly used A/D converters. Hence, experimental methods have been developed in the past which weaken the water signal during the experiment. Although this methods allow a quantitative analysis of the protein signals they still fail to suppress the water signal sufficiently such that it still remains as the largest peak in the recorded spectra. There, it overlaps neighboring protein peaks and often leads to severe baseline distortions which exacerbate an automated analysis of the recorded data.

For this reason, postprocessing methods will be proposed in this thesis with which the water resonance can be removed from the recorded data algorithmically. These methods will be based partly on Blind Source Separation (BSS), a technique in which the observed sensor signals are separated into their underlying sources. From these sources, the water related signals will be identified and nilled out deliberately such that only the protein signals prevail.

A drawback of the BSS methods is that they inevitably lead to increased noise in the resulting spectra. Hence, extensions to Principal Component Analysis (PCA), called local and Kernel PCA, will be applied with which the additional noise can be removed again. Furthermore, the algorithm dAMUSE will presented with which both BSS and denoising can be carried out simultaneously. In this context a novel approach called Autoassing will be proposed with which water and protein related sources can be identified automatically.

Finally, also Singular Spectrum Analysis (SSA) will be used to remove the water signal from NMR spectra. As will be seen this approach has the appealing property that no additional noise appears in the resulting spectra and that no identification of water related sources is necessary.

Altogether the thesis is organized as follows:

*Chapter 1* reviews the mathematical concepts used throughout the thesis.

*Chapter 2* gives an introduction to NMR spectroscopy. Special attention will be paid to the 2D-NOESY NMR protocol as all datasets analyzed in this thesis are of this type.

*Chapter 3* explains in detail how the data recorded in 2D-NOESY experiments can be analyzed by means of Blind Source Separation. A particular BSS algorithm called MP-BSS will be presented and its limits will be investigated using artificial data. MP-BSS will also be applied to real world 2D-NOESY data whereas the noise problem inherent in the BSS approach will become obvious.

*Chapter 4* shows how the additional noise appearing in the MP-BSS spectra can be removed again by local and Kernel PCA based denoising. Furthermore performance measures will be suggested with which the results obtained by both denoising methods can be compared quantitatively. These performance measures will also be used to determine the optimal sets of the parameters appearing in local and Kernel PCA.

*Chapter 5* deals with the algorithm dAMUSE with which BSS and denoising can be carried out simultaneously. Again, performance measures will be proposed with which the optimal set of parameters of dAMUSE can be determined. Moreover, the algorithm Autoassign will be presented with which the water and protein related sources estimated by dAMUSE can be clustered automatically.

*Chapter 6* investigates the suitability of SSA to remove the water signal from NMR spectra. After a comparison with the results obtained by MP-BSS and dAMUSE it will be concluded that SSA is the method of choice for the analysis of NMR data. To show the robustness of the SSA approach it will be applied to several data sets of proteins whereas each time the water signal is removed successfully.

*Chapter 7* finally concludes the thesis.

# Chapter 1

## Mathematical Preliminaries

This chapter offers a short review of some basic mathematical concepts which will be used frequently throughout this thesis. The first part gives a short introduction into some basic probability theory. The second part deals with three common performance measures which are used to quantify similarities between two signals or two matrices, respectively. Finally, so-called matrix pencils and their generalized eigenvalue decomposition (GEVD) will be presented. These GEVDs build the backbone of most of the algorithms presented later on.

### 1.1 Basic Probability Theory

In this section some of the fundamental properties of random variables will be reviewed briefly. For a profounder introduction the interested reader is referred to e.g. [32] or [42].

#### 1.1.1 Distribution and Density Functions

In the field of random variables or vectors the probability with which a certain value is taken is of major interest. In order to quantify this probability the cumulative distribution function is defined:

**Definition 1** (Cumulative distribution function (cdf)). *The cumulative distribution function  $F$  of a continuous random vector  $\mathbf{x} = [x_1, x_2, \dots, x_M]^T$ <sup>1</sup> at point  $\mathbf{x} = \mathbf{x}_0$  is defined as the probability  $P$  that  $\mathbf{x} \leq \mathbf{x}_0$ :*

$$F(\mathbf{x}_0) := P(\mathbf{x} \leq \mathbf{x}_0). \quad (1.1)$$

---

<sup>1</sup>” $\top$ ” denotes the transpose

The range of the cdf is  $[0, 1]$  whereas the minimum value is reached for  $F(\mathbf{x}_0 = -\infty)$  and the maximum for  $F(\mathbf{x}_0 = +\infty)$ . In-between these extremas  $F(\mathbf{x}_0)$  is continuous and nondecreasing.

Directly linked with the cdf is the probability density function which is defined as the derivate of the cdf:

**Definition 2** (Probability density function (pdf)). *The probability density function  $p(\mathbf{x}_0)$  of a continuous random variable  $\mathbf{x}$  is the first derivate of the its cdf:*

$$p(\mathbf{x}_0) = \left. \frac{dF(\mathbf{x}_0)}{d\mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0} = \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \cdots \frac{\partial}{\partial x_M} F(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (1.2)$$

Obviously, the pdf is normalized

$$\int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x} = 1. \quad (1.3)$$

Cdfs and pdfs are also defined for cases of more than just one random vector. In particular, if  $\mathbf{x}$  is an  $M$ -dimensional and  $\mathbf{y}$  is an  $N$  dimensional random vector then both can be combined to form the vector  $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$  for which the above-mentioned formulas can straightforwardly be extended:

**Definition 3** (Joint cdf and pdf). *Let  $\mathbf{x}$  and  $\mathbf{y}$  be two continuous random vectors of dimension  $M$  and  $N$ , respectively. The joint cummulative distribution function of  $\mathbf{x}$  and  $\mathbf{y}$  is defined by*

$$F_{\mathbf{x},\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = P(\mathbf{x} \leq \mathbf{x}_0, \mathbf{y} \leq \mathbf{y}_0) \quad (1.4)$$

Furthermore, the joint probability density function of  $\mathbf{x}$  and  $\mathbf{y}$  is given by

$$p_{\mathbf{x},\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = \left. \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \cdots \frac{\partial}{\partial x_M} \frac{\partial}{\partial y_1} \frac{\partial}{\partial y_2} \cdots \frac{\partial}{\partial y_N} F(\mathbf{x}, \mathbf{y}) \right|_{\mathbf{x}=\mathbf{x}_0, \mathbf{y}=\mathbf{y}_0}. \quad (1.5)$$

Finally, also so-called marginal densities are defined:

**Definition 4** (Marginal densities). *Given the joint pdf  $p_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y})$  of two continuous random vectors  $\mathbf{x}$  and  $\mathbf{y}$  the corresponding marginal densities  $p_{\mathbf{x}}(\mathbf{x})$  and  $p_{\mathbf{y}}(\mathbf{y})$  are defined by*

$$p_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{\infty} p_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \boldsymbol{\eta}) d\boldsymbol{\eta} \quad (1.6)$$

$$p_{\mathbf{y}}(\mathbf{y}) = \int_{-\infty}^{\infty} p_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \mathbf{y}) d\boldsymbol{\xi} \quad (1.7)$$

$$(1.8)$$

### 1.1.2 Expectations

Based on the pdf the expectation of a random vector or a function of it can be defined:

**Definition 5** (Expectation). *Let  $\mathbf{x}$  be a continuous random vector with pdf  $p(\mathbf{x})$  and let  $\mathbf{g}(\mathbf{x})$  denote any quantity derived from  $\mathbf{x}$ . The expectation  $E\{\mathbf{g}(\mathbf{x})\}$  of  $\mathbf{g}(\mathbf{x})$  is defined as*

$$E\{\mathbf{g}(\mathbf{x})\} = \int_{-\infty}^{\infty} \mathbf{g}(\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (1.9)$$

Here,  $\mathbf{g}(\mathbf{x})$  may either be a scalar, a vector or even a matrix. Expectations have some appealing properties which are often utilized in practice:

1. The expectation is a *linear* function, i.e.

$$E\left\{\sum_{m=1}^M a_m \mathbf{x}_m\right\} = \sum_{m=1}^M a_m E\{\mathbf{x}_m\} \quad (1.10)$$

where the  $\mathbf{x}_m$ ,  $m = 1, \dots, M$ , are random vectors while the  $a_m$  are nonrandom scalar factors.

2. This linearity can readily be extended to *linear transforms* of random vectors. Let  $\mathbf{x} = [x_1, x_2, \dots, x_M]^\top$  be an  $M$ -dimensional continuous random vector and let  $\mathbf{A}$  and  $\mathbf{B}$  be two nonrandom matrices of dimension  $K \times M$  and  $M \times L$ , respectively. Then

$$E\{\mathbf{A}\mathbf{x}\} = \mathbf{A}E\{\mathbf{x}\}, \quad E\{\mathbf{x}\mathbf{B}\} = E\{\mathbf{x}\}\mathbf{B} \quad (1.11)$$

3. The expectation is *transformation invariant*, i.e. if  $\mathbf{y} = \mathbf{g}(\mathbf{x})$  is a vector valued function of the continuous random vector  $\mathbf{x}$  then

$$\int_{-\infty}^{\infty} \mathbf{y}p_{\mathbf{y}}(\mathbf{y})d\mathbf{y} = \int_{-\infty}^{\infty} \mathbf{g}(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \quad (1.12)$$

where  $p_{\mathbf{y}}$  and  $p_{\mathbf{x}}$  denote the pdf of  $\mathbf{y}$  and  $\mathbf{x}$ , respectively.

The simplest example of an expectation is the first moment or mean of a random vector which is defined as follows:

**Definition 6** (Mean). *Let  $\mathbf{x} = [x_1, x_2, \dots, x_M]^\top$  be a continuous random vector with pdf  $p_{\mathbf{x}}(\mathbf{x})$ . The mean of  $\mathbf{x}$  is defined by*

$$\mathbf{m}_{\mathbf{x}} = E\{\mathbf{x}\} = \int_{-\infty}^{\infty} \mathbf{x}p_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \quad (1.13)$$

with components

$$m_{x_i} = E\{x_i\} = \int_{-\infty}^{\infty} x_i p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} x_i p_{x_i}(x_i) dx_i, \quad i = 1, 2, \dots, M, \quad (1.14)$$

whereas  $p_{x_i}(x_i)$  denotes the marginal density of the  $i$ -th component  $x_i$  of  $\mathbf{x}$ .

Another important statistical property is the correlation between the elements of a random vector  $\mathbf{x}$ :

**Definition 7** (Correlation). Let  $\mathbf{x} = [x_1, x_2, \dots, x_M]^T$  be a continuous random vector. The  $M \times M$  matrix

$$\mathbf{R}_{\mathbf{x}} = E\{\mathbf{x}\mathbf{x}^H\} \quad (1.15)$$

with elements

$$r_{ij} = E\{x_i \bar{x}_j\} = \int_{-\infty}^{\infty} x_i \bar{x}_j p_{x_i, x_j}(x_i, x_j) dx_i dx_j \quad (1.16)$$

is called the correlation matrix of  $\mathbf{x}$ .

Here  $\mathbf{x}^H$  denotes the conjugate transpose of  $\mathbf{x}$  and  $\bar{x}_j$  is the complex conjugate of  $x_j$ .

The correlation matrix has the following important properties:

1.  $\mathbf{R}_{\mathbf{x}}$  is hermitian:

$$\mathbf{R}_{\mathbf{x}} = \mathbf{R}_{\mathbf{x}}^H. \quad (1.17)$$

2.  $\mathbf{R}_{\mathbf{x}}$  is positive semidefinite i.e.

$$\mathbf{a}^H \mathbf{R}_{\mathbf{x}} \mathbf{a} \geq 0 \quad (1.18)$$

for all  $M$ -dimensional vectors  $\mathbf{a}$ . In practice,  $\mathbf{R}_{\mathbf{x}}$  is usually even positive definite, i.e. for all  $\mathbf{a} \neq 0$  Eq. (1.18) holds as a strict inequality.

3.  $\mathbf{R}_{\mathbf{x}}$  has real nonnegative eigenvalues (even positive eigenvalues if  $\mathbf{R}_{\mathbf{x}}$  is positive definite). The corresponding eigenvectors can always be chosen such that they are mutually orthogonal.

Closely related with the correlation matrix is the covariance matrix:

**Definition 8** (Covariance). Let  $\mathbf{x} = [x_1, x_2, \dots, x_M]^T$  be a continuous random vector. The  $M \times M$  matrix

$$\mathbf{C}_{\mathbf{x}} = E\{(\mathbf{x} - \mathbf{m}_{\mathbf{x}})(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^H\} \quad (1.19)$$

with elements

$$c_{ij} = E\{(x_i - m_i)(\bar{x}_j - \bar{m}_j)\} \quad (1.20)$$

is called the covariance matrix of  $\mathbf{x}$ .

Obviously, the correlation and the covariance matrix are related by

$$\mathbf{R}_x = \mathbf{C}_x + \mathbf{m}_x \mathbf{m}_x^H \quad (1.21)$$

and  $\mathbf{C}_x$  satisfies the same properties as  $\mathbf{R}_x$ .

In practice the expectations of random vectors cannot be computed analytically as usually the relevant probability densities are unknown. However, often a set of  $T$  samples  $\mathbf{x}_1, \dots, \mathbf{x}_T$  of a  $M$ -dimensional random vector  $\mathbf{x} = [x_1, x_2, \dots, x_M]^T$  is available from which the expectations can be estimated. These samples are used to constitute an  $M \times T$  observation matrix  $\mathbf{X}$ , whereas the  $i$ -th row consists of the samples of the random variable  $x_i$ .

The mean of the random vector  $\mathbf{x}$  can then be estimated by the following well-known formula:

$$\mathbf{m}_x \approx \frac{1}{T} \mathbf{X} \mathbf{1}_T \quad (1.22)$$

whereas  $\mathbf{1}_T$  is an  $T$ -dimensional vector consisting of ones:  $\mathbf{1}_T = \underbrace{[1, 1, \dots, 1]^T}_{T\text{-times}}$ .

Similarly, the estimates of the correlation and the covariance matrix are obtained by

$$\mathbf{R}_x \approx \frac{1}{T-1} \mathbf{X} \mathbf{X}^H \quad (1.23)$$

and

$$\mathbf{C}_x \approx \frac{1}{T-1} (\mathbf{X} - \mathbf{m}_x \mathbf{1}_T^T) (\mathbf{X} - \mathbf{m}_x \mathbf{1}_T^T)^H \quad (1.24)$$

respectively.

## 1.2 Performance Measures

Throughout this thesis performance measures will be needed which quantify how well a certain signal or a matrix is estimated by some algorithm. In the first case the correlation coefficient

$$r(s^{est}, s^{org}) = \frac{\sum_{t=1}^T (s^{org}(t) - m_{s^{org}})(s^{est}(t) - m_{s^{est}})}{\sqrt{\sum_{t=1}^T (s^{org}(t) - m_{s^{org}})^2 \sum_{t=1}^T (s^{est}(t) - m_{s^{est}})^2}}. \quad (1.25)$$

can be used to compare the original signal  $s^{org}(t) = [s^{org}(1), s^{org}(2), \dots, s^{org}(T)]$  with its estimate  $s^{est}(t) = [s^{est}(1), s^{est}(2), \dots, s^{est}(T)]$ . The correlation coefficient is a normalized covariance with range  $[-1, 1]$ . In particular,  $r = 1$  means that the two signals are perfectly correlated,  $r = -1$  indicates perfect anticorrelation and for  $r = 0$  no correlation between the original and

the estimated signal exists. Note, that the correlation coefficient only quantifies similarities between the waveforms of the two signals but does not allow for different amplitudes.

Alternatively, the similarity between the above signals can also be measured as a signal to noise ratio ( $SNR$ ) according to

$$SNR(s^{org}, s^{est}) = 20 \log_{10} \frac{\|s^{org}\|}{\|s^{org} - s^{est}\|} [dB]. \quad (1.26)$$

Here, the sampled signals  $s^{org} = [s^{org}(1), s^{org}(2), \dots, s^{org}(T)]$  and  $s^{est} = [s^{est}(1), s^{est}(2), \dots, s^{est}(T)]$  are seen as vectors in an  $T$ -dimensional space and  $\|\cdot\|$  denotes the Euclidean norm. Obviously, large  $SNR$  values indicate a high level of similarity between  $s^{org}$  and  $s^{est}$ .

In order to compare two square matrices  $\mathbf{A}^{est}$  and  $\mathbf{A}^{org}$  the Cross-Talking-Error (CTE) is often used. Its value measures how well two matrices equal each other up to scaling and permutation indeterminacies. For its computation the matrix

$$\mathbf{P} = (\mathbf{A}^{est})^{-1} \mathbf{A}^{org} \quad (1.27)$$

is needed. If  $\mathbf{A}^{org}$  and  $\mathbf{A}^{est}$  are equal up to scaling and permutation indeterminacies then the matrix  $\mathbf{P}$  is a (possibly permuted) diagonal matrix with only one nonzero entry in each row and each column. This structure gets blurred, however, when further differences between  $\mathbf{A}^{org}$  and  $\mathbf{A}^{est}$  exist. In this case the zero elements are replaced by elements whose magnitudes scale with the difference between the original and the estimated matrix. This deviation from the ideal structure of  $\mathbf{P} = (P_{ij})$  is measured by the CTE:

$$CTE = \sum_i \left( \sum_j \frac{|P_{ij}|}{\max_k |P_{ik}|} - 1 \right) + \sum_j \left( \sum_i \frac{|P_{ij}|}{\max_k |P_{ik}|} - 1 \right) \quad (1.28)$$

Obviously, the  $CTE$  vanishes if  $\mathbf{P}$  is a diagonal matrix. Hence, a matrix  $\mathbf{A}^{est}$  is a good estimate of  $\mathbf{A}^{org}$  if the corresponding  $CTE$  is small.

### 1.3 Matrix Pencils and the General Eigenvalue Decomposition

Throughout this thesis the concept of matrix pencils and their general eigenvalue decomposition (GEVD) will be utilized frequently. Hence, a short review of this framework will be presented in the following. First, the term matrix pencil has to be defined precisely:

**Definition 9** ((Hermitian definite) matrix pencil). *Let  $\mathbf{C}$  and  $\mathbf{D}$  be two  $\mathbb{C}^{M \times M}$ -matrices. The set of all matrices of the form*

$$(\mathbf{C}, \mathbf{D}) := \mathbf{C} - \mu \mathbf{D}, \quad \mu \in \mathbb{C} \quad (1.29)$$

*is called matrix pencil. If additionally both  $\mathbf{C}$  and  $\mathbf{D}$  are hermitian and  $\mathbf{D}$  is positive definite then the matrix pencil  $(\mathbf{C}, \mathbf{D})$  is called positive definite.*

The eigenvalues and eigenvectors of such matrix pencils are defined as follows:

**Definition 10** (Eigenvalues and eigenvectors of matrix pencils). *The eigenvalues of a matrix pencil  $(\mathbf{C}, \mathbf{D})$  are the elements of the set*

$$\lambda(\mathbf{C}, \mathbf{D}) := \{\lambda \in \mathbb{C} \mid \det(\mathbf{C} - \lambda \mathbf{D}) = 0\}. \quad (1.30)$$

*If  $\lambda \in \lambda(\mathbf{C}, \mathbf{D})$ ,  $\mathbf{e} = [e_1, e_2, \dots, e_m]^\top \in \mathbb{C}^M$  and furthermore*

$$\mathbf{C}\mathbf{e} = \lambda \mathbf{D}\mathbf{e}, \quad \mathbf{e} \neq 0, \quad (1.31)$$

*then  $\mathbf{e}$  is called an eigenvector of  $(\mathbf{C}, \mathbf{D})$ .*

For positive definite matrix pencils the generalized eigenvalue decomposition Eq. (1.31) always exists as stated by the following theorem [21]:

**Theorem 1** (Generalized eigenvalue decomposition (GEVD) of positive definite matrix pencils). *If  $(\mathbf{C}, \mathbf{D})$  is a positive definite matrix pencil then there exists a matrix  $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_M] \in \mathbb{C}^{M \times M}$  such that*

$$\mathbf{E}^H \mathbf{C} \mathbf{E} = \text{diag}(p_1, \dots, p_M) \quad \text{and} \quad \mathbf{E}^H \mathbf{D} \mathbf{E} = \text{diag}(q_1, \dots, q_M), \quad (1.32)$$

*whereas  $q_i, p_i \in \mathbb{R}$ .*

In other words, Theorem 1 states that the columns of  $\mathbf{E}$  are the eigenvectors of the  $(\mathbf{C}, \mathbf{D})$ , i.e.

$$\mathbf{C}\mathbf{E} = \mathbf{D}\mathbf{E}\mathbf{\Lambda}, \quad i = 1, 2, \dots, M, \quad (1.33)$$

where  $\mathbf{\Lambda}$  is a diagonal matrix containing the eigenvalues  $\lambda_1 = p_1/q_1, \lambda_2 = p_2/q_2, \dots, \lambda_M = p_M/q_M$  of  $(\mathbf{C}, \mathbf{D})$  on its diagonal.

Furthermore, the uniqueness of the GEVD of a positive definite matrix pencil is important:

**Theorem 2** (Uniqueness of the GEVD of positive definite matrix pencils). *Let  $(\mathbf{C}, \mathbf{D})$  be a positive definite matrix pencil with GEVD*

$$\mathbf{C}\mathbf{E} = \mathbf{D}\mathbf{E}\mathbf{A}. \quad (1.34)$$

*and non-degenerated eigenvalues. Then the eigenvalue matrix  $\mathbf{\Lambda}$  and the eigenvector matrix  $\mathbf{E}$  are unique up to scalings and permutations of their columns.*

Apart from positive definiteness also the congruence between two matrix pencils will play an important role later on. This congruence is defined as follows:

**Definition 11** (Congruent matrix pencils). *Two matrix pencils  $(\mathbf{C}_1, \mathbf{D}_1)$  and  $(\mathbf{C}_2, \mathbf{D}_2)$  with  $\mathbf{C}_i, \mathbf{D}_i \in \mathbb{C}^{M \times M}$ ,  $i = 1, 2$ , are called congruent if there exists an invertible matrix  $\mathbf{A} \in \mathbb{C}^{M \times M}$  such that*

$$\mathbf{C}_1 = \mathbf{A}\mathbf{C}_2\mathbf{A}^H \quad \text{and} \quad \mathbf{D}_1 = \mathbf{A}\mathbf{D}_2\mathbf{A}^H. \quad (1.35)$$

For congruent matrix pencils the following two important theorems exist:

**Theorem 3** (Eigenvalues of congruent positive definite matrix pencils[55]). *Congruent positive definite matrix pencils have identical eigenvalues.*

*Proof.* Let  $(\mathbf{C}_1, \mathbf{D}_1)$  and  $(\mathbf{C}_2, \mathbf{D}_2)$  be two congruent matrix pencils fulfilling Eq. (1.35). The eigenvalues of the matrix pencil  $(\mathbf{C}_1, \mathbf{D}_1)$  are the roots of its characteristic polynomial

$$\chi(\lambda) = \det(\mathbf{C}_1 - \lambda\mathbf{D}_1) = 0. \quad (1.36)$$

Making use of the congruence of the two matrix pencils this polynomial can be rewritten as

$$\det(\mathbf{C}_1 - \lambda\mathbf{D}_1) = \det(\mathbf{A}\mathbf{C}_2\mathbf{A}^H - \lambda\mathbf{A}\mathbf{D}_2\mathbf{A}^H) \quad (1.37)$$

$$= \det(\mathbf{A}) \det(\mathbf{C}_2 - \lambda\mathbf{D}_2) \det(\mathbf{A}^H). \quad (1.38)$$

Obviously, the last term only vanishes if  $\lambda$  is also an eigenvalue of  $(\mathbf{C}_2, \mathbf{D}_2)$ .  $\square$

**Theorem 4** (Eigenvectors of congruent positive definite matrix pencils[55]). *Let  $(\mathbf{C}_1, \mathbf{D}_1)$  and  $(\mathbf{C}_2, \mathbf{D}_2)$  be two congruent positive definite matrix pencils*

$$\mathbf{C}_1 = \mathbf{A}\mathbf{C}_2\mathbf{A}^H \quad \text{and} \quad \mathbf{D}_1 = \mathbf{A}\mathbf{D}_2\mathbf{A}^H, \quad \mathbf{A} \text{ invertible} \quad (1.39)$$

*with nondegenerated eigenvalues. Furthermore, let  $\mathbf{E}_1$  and  $\mathbf{E}_2$  denote the eigenvector matrices of  $(\mathbf{C}_1, \mathbf{D}_1)$  and  $(\mathbf{C}_2, \mathbf{D}_2)$ , respectively. Then  $\mathbf{E}_1$  and  $\mathbf{E}_2$  are related by*

$$\mathbf{E}_2 = \mathbf{A}^H\mathbf{E}_1. \quad (1.40)$$

**Data:**  
 ( $\mathbf{C}, \mathbf{D}$ ): positive definite matrix pencil  
**Result:**  
 $\mathbf{E}$ : eigenvector matrix  
 $\mathbf{\Lambda}$ : eigenvalue matrix

```

1 begin
2   EVD of  $\mathbf{D}$ :  $\mathbf{D} = \mathbf{E}_D \mathbf{\Lambda}_D \mathbf{E}_D^H$ 
3    $\mathbf{U} := \mathbf{\Lambda}_D^{-1/2} \mathbf{E}_D^H$ 
4   EVD of  $\mathbf{UCU}^H$ :  $\mathbf{UCU}^H \mathbf{V} = \mathbf{V} \mathbf{\Lambda}$ 
5    $\mathbf{E} = \mathbf{E}_D \mathbf{\Lambda}_D^{-1/2} \mathbf{V}$ 
6 end

```

**Algorithm 1:** GEVD by means of two single EVDs.

*Proof.* Being positive definite  $(C_1, D_1)$  has a GEVD of the form

$$\mathbf{C}_1 \mathbf{E}_1 = \mathbf{D}_1 \mathbf{E}_1 \mathbf{\Lambda}. \quad (1.41)$$

Because of the congruence of the two matrix pencils this equation can be expressed in terms of the matrices  $\mathbf{C}_2$  and  $\mathbf{D}_2$ :

$$\mathbf{A} \mathbf{C}_2 \mathbf{A}^H \mathbf{E}_1 = \mathbf{A} \mathbf{D}_2 \mathbf{A}^H \mathbf{E}_1 \mathbf{\Lambda} \quad (1.42)$$

As  $\mathbf{A}$  is invertible by definition both sides of Eq. (1.42) can be multiplied from the left by  $\mathbf{A}^{-1}$ :

$$\mathbf{C}_2 \mathbf{A}^H \mathbf{E}_1 = \mathbf{D}_2 \mathbf{A}^H \mathbf{E}_1 \mathbf{\Lambda} \quad (1.43)$$

Obviously, this is the GEVD of the matrix pencil  $(\mathbf{C}_2, \mathbf{D}_2)$  with eigenvector matrix

$$\mathbf{E}_2 = \mathbf{A}^H \mathbf{E}_1. \quad (1.44)$$

□

Eventually, the question arises how the GEVD

$$\mathbf{C} \mathbf{E} = \mathbf{D} \mathbf{E} \mathbf{\Lambda} \quad (1.45)$$

of a positive definite matrix pencil  $(\mathbf{C}, \mathbf{D})$  is computed in practice. This can be done by the following two step procedure in which the generalized EVD problem is solved by two standard EVDs (see Alg. 1) [55]. First, the EVD of the matrix  $\mathbf{D}$  is computed

$$\mathbf{D} = \mathbf{E}_D \mathbf{\Lambda}_D \mathbf{E}_D^H = \mathbf{E}_D \mathbf{\Lambda}_D^{1/2} \mathbf{\Lambda}_D^{1/2} \mathbf{E}_D^H \quad (1.46)$$

whereas  $\mathbf{E}_D$  and  $\Lambda_D$  are the eigenvector and eigenvalue matrix of  $\mathbf{D}$ , respectively. Defining

$$\mathbf{U} := \Lambda_D^{-1/2} \mathbf{E}_D^H \quad (1.47)$$

the matrix  $\mathbf{D}$  in Eq. (1.45) can be substituted by

$$\mathbf{D} = \mathbf{U}^{-1} (\mathbf{U}^{-1})^H \quad (1.48)$$

leading to

$$\mathbf{C}\mathbf{E} = \mathbf{U}^{-1} (\mathbf{U}^{-1})^H \mathbf{E} \Lambda \quad (1.49)$$

Next, the matrix

$$\mathbf{V} := (\mathbf{U}^{-1})^H \mathbf{E} \quad (1.50)$$

can be used to further transform Eq. (1.49):

$$\mathbf{C}\mathbf{U}^H \mathbf{V} = \mathbf{U}^{-1} \mathbf{V} \Lambda. \quad (1.51)$$

Multiplying both sides of the above equation by  $\mathbf{U}$  and substituting

$$\mathbf{W} := \mathbf{U}\mathbf{C}\mathbf{U}^{-1} \quad (1.52)$$

leads to the second standard EVD problem, this time of the matrix  $\mathbf{W}$ :

$$\mathbf{W}\mathbf{V} = \mathbf{V}\Lambda. \quad (1.53)$$

Hence, the eigenvalue matrix  $\Lambda$  of the initial GEVD problem is determined after this second standard EVD step. However, also the eigenvector matrix  $\mathbf{E}$  can now be computed by means of Eq. (1.50)

$$\mathbf{E} = \mathbf{U}^H \mathbf{V} = \mathbf{E}_D \Lambda_D^{-1/2} \mathbf{V} \quad (1.54)$$

# Chapter 2

## NMR Spectroscopy

In this chapter the behavior of nuclear spins under the influence of an external magnetic field will be described first. Based on these considerations the concepts of Nuclear Magnetic Resonance (NMR) spectroscopy will be introduced and it will be shown how this technique is applied to distinguish identical nuclear spins in different molecular surroundings. Furthermore, after a short review of the Nuclear Overhauser Effect (NOE), also two dimensional NMR (2D-NMR) spectroscopy will be explained. These 2D-NMR experiments have become very popular in recent years and are nowadays commonly used to determine the spatial structure of proteins. Accordingly, all the datasets analyzed in this thesis stem from 2D-NMR experiments.

### 2.1 Basics of NMR

Generally, processes which occur at the elementary level of single particles and molecules can only be described and understood completely by means of Quantum Mechanics. In the field of NMR, however, the recorded signals originate from a macroscopic sample which contains a large number of molecules. One of the major criteria for the validity of microscopic theories like Quantum Mechanics is, though, that it leads to the same results for macroscopic systems as classical considerations. Thus, the basic principles of NMR can be described using the concepts of classical physics if those are extended by some results stemming from Quantum Mechanics. Hence, in the following a semiclassical elucidation of NMR is preferred over a Quantum Mechanical one as the first is mathematically simpler and vivid.

### 2.1.1 Magnetic Moment and Nuclear Spin

In order to investigate atoms by means of NMR their nuclei have to possess a nuclear spin  $\mathbf{J}$  [28]. Related with this nuclear spin is always a magnetic moment  $\boldsymbol{\mu}_{\mathbf{I}}$  via the following simple relation:

$$\boldsymbol{\mu}_{\mathbf{I}} = \gamma_{\mathbf{I}}\mathbf{J} = \hbar\gamma_{\mathbf{I}}\mathbf{I}. \quad (2.1)$$

where  $\gamma_{\mathbf{I}}$  is a nucleus specific constant called magnetogyric ratio. Usually, the nuclear spin  $\mathbf{J}$  is expressed in terms of Planck's constant  $\hbar = 1.0546 \cdot 10^{-34} \text{m}^2\text{kgs}^{-1}$  and is denoted by the dimensionless symbol  $\mathbf{I}$ . The magnitude of the nuclear spin of a certain molecule is expressed by the spin quantum number  $I$ , which for theoretical reasons can only assume multiple values of  $1/2$ . Quantum mechanical considerations lead to the following relation for the absolute value of the nuclear spin  $\mathbf{J}$ :

$$\langle |\mathbf{J}| \rangle = \hbar \langle |\mathbf{I}| \rangle = \hbar \sqrt{I(I+1)} \quad (2.2)$$

In the thesis at hand all NMR experiments investigate the behavior the hydrogen atom  $^1\text{H}$ . Its nucleus has spin quantum number of  $I = 1/2$  and possesses the largest magnetogyric ratio of  $2.6752 \cdot 10^8 \text{T}^{-1}\text{s}^{-1}$  of all stable isotopes. All further explanations hence refer to nuclei with nuclear spin  $I = 1/2$  and positive gyromagnetic ratio.

### 2.1.2 Nuclei in External Magnetic Fields

The macroscopic samples used in NMR experiments contain a large number of nuclei. For example,  $1 \text{ cm}^3$  of water already contains several  $10^{22}$  of H nuclei. Without an external magnetic field the magnetic dipoles  $\boldsymbol{\mu}_{\mathbf{I}}$  of these nuclei orient themselves at random and compensate each other such that no overall magnetization can be observed. This changes, however, if an external magnetic field  $\mathbf{B}_0$  is applied e.g. in the  $z$ -direction of a Cartesian coordinate system. Similar to a gyro in the gravitation field of the earth the magnetic moments  $\boldsymbol{\mu}_{\mathbf{I}}$  then start to precess around  $\mathbf{B}_0$  with a *Larmor Frequency* of

$$\boldsymbol{\omega}_{\mathbf{I}} = -\gamma_{\mathbf{I}}\mathbf{B}_0. \quad (2.3)$$

Furthermore, Quantum Mechanical considerations lead to the result that the expectation value of the  $z$ -component of the nuclear spin (and hence also the magnetic moment) is quantized as

$$\langle I_z \rangle = \hbar m_{\mathbf{I}} \quad (2.4)$$

where  $m_{\mathbf{I}}$  is the magnetic quantum number. For atoms with nuclear spin  $I = 1/2$   $m_{\mathbf{I}}$  can only take the values  $+1/2$  and  $-1/2$ , respectively.

The energy of a magnetic dipole  $\boldsymbol{\mu}$  in an external magnetic field  $\mathbf{B}_0$  is given by

$$E_m = -\boldsymbol{\mu}\mathbf{B}_0 = -\hbar\gamma_{\mathbf{I}}m_{\mathbf{I}}B_0. \quad (2.5)$$

Accordingly, only the following two energy levels  $E_1$  and  $E_2$  exist for atoms with nuclear spin  $I = 1/2$ :

$$E_1 = -\frac{1}{2}\hbar\gamma_{\mathbf{I}}B_0 \quad (2.6)$$

$$E_2 = \frac{1}{2}\hbar\gamma_{\mathbf{I}}B_0 \quad (2.7)$$

As with any spectroscopy transitions between these two levels can only occur if the system is excited by an electromagnetic radiation whose frequency  $\omega$  fulfills the Bohr condition, i.e. for which  $\Delta E = E_2 - E_1 = \hbar\omega$  holds. This leads to the basic resonance condition of NMR:

$$\omega = \gamma_{\mathbf{I}}B_0. \quad (2.8)$$

In other words, the absolute value of the resonance frequency  $\omega$  used to excite the system must equal the Larmor frequency (Eq. (2.3)) with which the magnetic dipole precesses around  $\mathbf{B}_0$ .

According to the Boltzmann law the energy levels  $E_1$  and  $E_2$  are populated differently. If  $N^+$  denotes the population of  $E_1$  and  $N^-$  the population of  $E_2$  then the following relations holds:

$$\frac{N^-}{N^+} = e^{-\Delta E/kT} \quad (2.9)$$

where  $k = 1.38 \cdot 10^{-23} \text{m}^2\text{kg s}^{-2}\text{K}^{-1}$  is the Boltzmann constant. At room temperature and in the presence of an external magnetic field of 1 T this leads to an absolute population difference of only about  $10^{-6}$ . This difference is sufficient, however, to obtain a detectable macroscopic magnetization  $\mathbf{M}$  parallel to the external magnetic field  $\mathbf{B}_0$ . In thermal equilibrium and at room temperature the magnitude of this magnetization is given by

$$\mathbf{M}_0 = \frac{N\hbar^2\gamma_{\mathbf{I}}^2 I(I+1)}{3kT} \mathbf{B}_0, \quad (2.10)$$

whereas  $N = N^+ + N^-$  is the absolute number of nuclear spins. In NMR experiments this magnetization is perturbed from its equilibrium position and is then detected by additional coils (see Sec. 2.3.1). From the recorded signals inferences can then be made on the molecules under investigation (see Sec. 2.4).

### 2.1.3 Bloch Equations

Shortly after the first NMR signals were detected successfully in 1946 Felix Bloch described classically the behavior of the  $M_x$ ,  $M_y$  and  $M_z$  component of the magnetization  $\mathbf{M}$  by a system of differential equations, the so-called Bloch Equations. In essence, these equations describe mathematically the following experimental findings [28]:

1. If the the magnetization  $\mathbf{M}$  points into another direction than the external magnetic field  $\mathbf{B}_0$  then  $\mathbf{M}$  precesses around the external magnetic field.
2. A sufficiently long time after the perturbation of  $\mathbf{M}$  the magnetization returns to the equilibrium position, i.e.  $M_z$  equals  $M_0$  and the components  $M_x$  and  $M_y$  perpendicular to  $\mathbf{B}_0$  vanish.

Formally, the exponential approach of  $M_z$  to  $M_0$  is described by means of the following equation:

$$\frac{dM_z}{dt} = \frac{M_0 - M_z}{T_1}. \quad (2.11)$$

The time constant  $T_1$  is called longitudinal relaxation time. Similarly, the decay of the transversal components  $M_x$  and  $M_y$  can be expressed mathematically by

$$\frac{dM_x}{dt} = -\frac{M_x}{T_2} \quad (2.12)$$

$$\frac{dM_y}{dt} = -\frac{M_y}{T_2} \quad (2.13)$$

Here, the time constant  $T_2$  is called transversal relaxation time. Combination of the above equations with the differential equation of the precessing magnetization

$$\frac{d\mathbf{M}}{dt} = -\gamma_{\mathbf{I}} (\mathbf{B} \times \mathbf{M}) \quad (2.14)$$

leads to the following *Bloch equations*

$$\begin{aligned} \frac{dM_x}{dt} &= -\gamma_{\mathbf{I}} (\mathbf{B}_0 \times \mathbf{M})_x - \frac{M_x}{T_2} \\ \frac{dM_y}{dt} &= -\gamma_{\mathbf{I}} (\mathbf{B}_0 \times \mathbf{M})_y - \frac{M_y}{T_2} \\ \frac{dM_z}{dt} &= -\gamma_{\mathbf{I}} (\mathbf{B}_0 \times \mathbf{M})_z - \frac{M_0 - M_z}{T_1}. \end{aligned} \quad (2.15)$$

In typical NMR experiments, the magnetization  $\mathbf{M}_0$  is perturbed from its equilibrium position by means of a linearly polarized, high frequency magnetic field which points into the  $x$ -direction

$$\mathbf{B}_\perp = 2B_1 \cos(\omega_{hf}t) \mathbf{e}_x \quad (2.16)$$

( $\mathbf{e}_x$  unit norm vector in  $x$ -direction). The frequency  $\omega_{hf}$  of this perpendicular magnetic field has to be chosen such that it is close to the Larmor frequency of the nuclei under investigation.

$\mathbf{B}_\perp$  can be decomposed into two circularly polarized magnetic fields  $\mathbf{B}^+$  and  $\mathbf{B}^-$  with frequencies of  $+\omega_{hf}$  and  $-\omega_{hf}$ , respectively. While the component rotating in the direction opposite to the larmor precession of the nuclear spins only exerts a negligible influence on the magnetization  $\mathbf{M}$  the other component is “seen” by the nuclear spins and leads to the perturbation of  $\mathbf{M}$  from its equilibrium position. For  $^1\text{H}$  nuclei, which rotate clockwise around the static magnetic field  $\mathbf{B}_0$  because of their positive magnetogyric ratio, this means that the magnetization  $\mathbf{M}$  is influenced by the component  $\mathbf{B}^-$ . Formally, this means that  $\mathbf{M}$  is affected by the overall magnetic field

$$\mathbf{B} = \mathbf{B}^- + \mathbf{B}_0 = B_1 \cos(\omega_{hf}t) \mathbf{e}_x - B_1 \sin(\omega_{hf}t) \mathbf{e}_y + B_0 \mathbf{e}_z \quad (2.17)$$

whereas  $\mathbf{e}_x$ ,  $\mathbf{e}_y$  and  $\mathbf{e}_z$  denote the unit norm vectors in  $x$ -,  $y$ - and  $z$ -direction, respectively. Inserting this magnetic field into the Bloch equations Eq. (2.15) leads to

$$\begin{aligned} \frac{dM_x}{dt} &= \gamma_{\mathbf{I}} (M_y B_0 + M_z B_1 \sin(\omega_{hf}t)) - \frac{M_x}{T_2} \\ \frac{dM_y}{dt} &= \gamma_{\mathbf{I}} (-M_x B_0 + M_z B_1 \cos(\omega_{hf}t)) - \frac{M_y}{T_2} \\ \frac{dM_z}{dt} &= \gamma_{\mathbf{I}} (-M_x B_1 \sin(\omega_{hf}t) - M_y B_1 \cos(\omega_{hf}t)) + \frac{M_0 - M_z}{T_1}. \end{aligned} \quad (2.18)$$

For nuclei with a positive magnetogyric ratio  $\gamma_{\mathbf{I}} > 0$ , which precess according to Eq. (2.14) clockwise around  $\mathbf{B}_0$  one often uses an alternative coordinate system  $x', y', z' = z$  which rotates with  $-\omega_{hf}$  around the  $z$ -axis. This coordinate transform allows a simplification of the above equations and also offers a more vivid illustration of the ongoing processes. For instance, the magnetic field  $\mathbf{B}^-$ , which rotates with  $-\omega_{hf}$  in the laboratory system is static in the rotating coordinate system and points into the  $x'$ -direction:

$$\mathbf{B}'^- = B_1 \mathbf{e}_{x'}. \quad (2.19)$$

Furthermore, the magnetization  $\mathbf{M}'$ , which revolves in the laboratory system with  $\omega_{\mathbf{I}}$  ( $\omega_{\mathbf{I}} < 0$  as  $\gamma_{\mathbf{I}} > 0$ ) around  $\mathbf{B}_0$ , precesses in the rotation

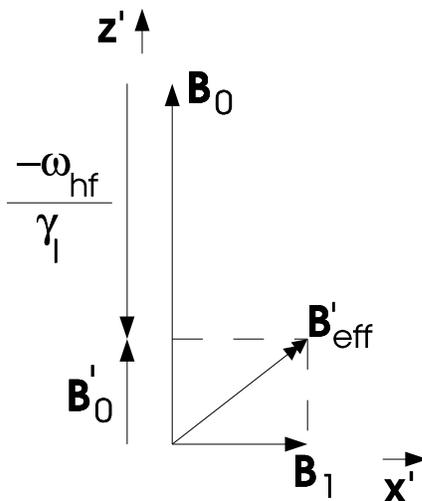


Figure 2.1: Illustration of the reduced magnetic field  $\mathbf{B}'_0$  and the effective field  $\mathbf{B}'_{eff}$  (see text).

coordinate system with

$$\Omega = \omega_{\mathbf{I}} + \omega_{hf}. \quad (2.20)$$

Because of Eq. (2.3) this reduced precession frequency also leads to a reduced static magnetic field  $\mathbf{B}'_0$  in the rotating coordinate system:

$$\mathbf{B}'_0 = -\frac{\Omega}{\gamma_{\mathbf{I}}} \mathbf{e}_{z'} \quad (2.21)$$

Usually,  $\mathbf{B}'_0$  is called reduced magnetic field. In the rotating coordinate system the magnetization  $\mathbf{M}'$  is hence subject to the effective magnetic field

$$\mathbf{B}'_{eff} = \mathbf{B}'_0 + \mathbf{B}'_{\perp} \quad (2.22)$$

around which it precesses with the effective frequency

$$\omega_{eff} = -\gamma_{\mathbf{I}} \mathbf{B}'_{eff}. \quad (2.23)$$

(see Fig. 2.1).

If the frequency  $\omega_{hf}$  of the perpendicular magnetic field  $\mathbf{B}_{\perp}$  equals the Larmor frequency  $\omega$  of the nuclei, then  $\Omega$  vanishes such that

$$\mathbf{B}'_{eff} = \mathbf{B}'_{\perp} = B_1 \mathbf{e}_{x'}. \quad (2.24)$$

Thus, the magnetization  $\mathbf{M}'$  is turned only around the  $x'$ -axis in the rotating coordinate system. If the HF-field  $\mathbf{B}_{\perp}$  is switched off after

$$\tau = \frac{\pi}{2\omega_{eff}} \quad (2.25)$$

then the magnetization points in the positive  $y'$  direction and one speaks of an  $\pi/2$  or  $90^\circ$  pulse. Furthermore, if the HF-field is switched off after  $2\tau$  the magnetization points in the negative  $z'$  direction the corresponding pulse is called a  $\pi$  or  $180^\circ$  pulse.

Finally, the Bloch equations (Eq. (2.15)) are also simplified in the rotating coordinate system as can be seen if Eq. (2.22) is inserted into them:

$$\begin{aligned}\frac{dM_{x'}}{dt} &= -M_{y'}\Omega - \frac{M_{x'}}{T_2} \\ \frac{dM_{y'}}{dt} &= \gamma_{\mathbf{I}}B_1M_{z'} + M_{x'}\Omega + \frac{M_{y'}}{T_2} \\ \frac{dM_{z'}}{dt} &= -\gamma_{\mathbf{I}}B_1M_{y'} + \frac{M_0 - M_{z'}}{T_1}\end{aligned}\quad (2.26)$$

The describe a damped precession of the magnetization around the effective magnetic field  $\mathbf{B}'_{eff}$ .

### 2.1.4 Chemical Shift

The considerations made so far only dealt with isolated nuclei in free space. In fact, though, the nuclei are incorporated into molecules where they are surrounded by an electronic shell. This shell slightly shields the nuclei from the external magnetic field  $\mathbf{B}_0$  such that they experience only a weaker effective field  $\mathbf{B}'$  [28]

$$\mathbf{B}' = \mathbf{B}_0(1 - \sigma) \quad (2.27)$$

whereas  $\sigma$  is called shielding constant. This effective field must also be inserted into the resonance equation of NMR:

$$\omega_{\mathbf{I}} = -\gamma_{\mathbf{I}}B' = -\gamma_{\mathbf{I}}B_0(1 - \sigma). \quad (2.28)$$

As the resulting shift of the frequency depends on the chemical surroundings of the nuclei it is usually termed *chemical shift*.

## 2.2 Relaxation

After the magnetization has been perturbed from its equilibrium position by irradiation of the perpendicular magnetic field  $\mathbf{B}_\perp$  it gradually returns back to its original position [28]. This processes can be divided into two parts called longitudinal and vertical relaxation. Both relaxation mechanisms are caused by mechanisms at the molecular level and are based on interactions of the nuclear spins with their surroundings. The observed relaxation rates are proportional to the square of magnitude of these interactions.

### 2.2.1 Longitudinal Relaxation

Longitudinal relaxation, also called spin-lattice-relaxation, describes the relaxation of the  $M_{z'}$ -component of the magnetization  $\mathbf{M}$  to its equilibrium value  $M_{z'}^0$ . Seen from the microscopic level the irradiated magnetic field  $\mathbf{B}_\perp$  transfers numerous nuclear spins from the energy level  $E_1$  to the higher level  $E_2$ , i.e. compared with the Boltzmann distribution Eq. (2.9)  $N^-$  became overpopulated at the expense of  $N^+$ . Once  $\mathbf{B}_\perp$  is switched off again the surplus spins populating  $E_2$  can only return to  $E_1$  if they are able to transmit the energy quantum  $\hbar\omega_{\mathbf{I}} = E_2 - E_1$  to their surroundings. This means that the interactions with the surroundings must be time dependent and must contain a frequency component at the Larmor frequency  $\omega_{\mathbf{I}}$ .

In fluids the time dependence of the interactions originates from the Brownian molecular movement. In order to quantify the latter the rotational correlation time  $\tau_{rot}$  is introduced which describes the average time span it takes a molecule to rotate by an angle of 1 rad. The intensity of the noise spectrum at the frequency  $\omega_{\mathbf{I}}$  then determines the efficiency of the relaxation process and hence the longitudinal relaxation time  $T_1$ .

### 2.2.2 Transversal Relaxation

In contrast to longitudinal relaxation transversal relaxation, also called spin-spin-relaxation, describes the decay of the  $M_{x'}$  and  $M_{y'}$  component after  $\mathbf{B}_\perp$  has been switched off [28]. To illustrate this mechanism, assume that the magnetization has been perturbed from its equilibrium position by a  $90^\circ$  pulse such that it points in the positive  $y'$  direction. As the external magnetic field  $\mathbf{B}_0$  is never perfectly homogeneous nuclear spins at different locations in the sample will precess at slightly different frequencies. This leads to a fanning out of the magnetization until, after a sufficiently long time, the fan is spread homogeneously over the  $x'y'$ -plane such that the macroscopic magnetization vanishes. If after some time  $\tau$  a further  $180^\circ$  pulse is applied the faster rotating spins will be located behind those which rotate at lower frequencies, i.e. the dephasing will be inverted such that after the time  $2\tau$  a macroscopic magnetization in negative  $y'$ -direction is observed. If there were no further relaxation mechanisms the magnetization pointing in the negative  $y'$  direction would be of the same magnitude as the original magnetization obtained after the first  $90^\circ$  pulse. Hence, this mechanism is called reversible transversal relaxation (see also Fig. 2.2).

However, also irreversible transverse relaxation is observed in real life experiments. A major source for such irreversible effects are spin-lattice relaxations as during the flipping of the spins between the individual energy

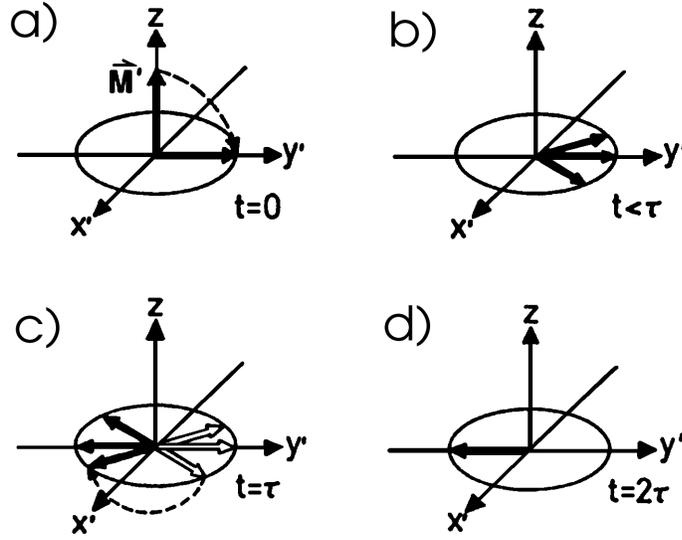


Figure 2.2: Reversible transversal relaxation [28]: a) Magnetization is turned into  $xy$ -plane by  $90^\circ$  pulse. b) Fanning of the magnetization caused by inhomogeneities in the magnetic field. c) In order to invert the fanning a  $180^\circ$  pulse is applied. d) After the time  $2\tau$  all spins are in-phase again if no irreversible relaxation mechanisms appear.

levels the phase coherence with their neighboring spins is destroyed. This leads to an exponential decay of the transversal magnetization with decay constant  $T_2$ . Altogether, an effective relaxation time  $T_2^*$  is observed in real life experiments which is the result of the reversible as well as for the irreversible relaxation processes [34]:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{2,inhomo}}. \quad (2.29)$$

Here,  $T_{2,inhomo}$  reflects the relaxation caused by inhomogeneities in the static magnetic field  $\mathbf{B}_0$ .

### 2.2.3 Dipolar Relaxation

Dipolar relaxation plays an important role in the case of nuclei like e.g.  $^1\text{H}$  which have a large magnetic moment  $\mu_{\text{I}}$ . To illustrate this relaxation mechanism consider two adjacent nuclear spins in a molecule. Each of these two nuclear spins experiences, apart from the external magnetic field  $\mathbf{B}_0$ , also a local magnetic field  $B_{loc}$  stemming from the other nuclear spin, respectively.

The size of this local magnetic field is given by

$$B_{loc} = \frac{\mu_0 \mu_z (3 \cos^2 \theta - 1)}{4\pi r^3} \quad (2.30)$$

where  $\mu_0$  is the magnetic permeability of vacuum,  $r$  is the distance between the two dipoles,  $\mu_z$  the component of the magnetic moment  $\boldsymbol{\mu}_I$  parallel to the external magnetic field  $\mathbf{B}_0$ , and  $\theta$  the angle between  $\mathbf{B}_0$  and the vector  $\mathbf{r}$  which connects the two dipoles.

In liquids this local magnetic field is time dependent as the angle  $\theta$  between  $\mathbf{r}$  and  $\mathbf{B}_0$  varies because of the Brownian motion. Furthermore, water is often used as a solvent in liquid NMR such that the concentration of  $^1\text{H}$  nuclei in the sample is very high. Hence, a nucleus in a molecule does not only experience the local fields of the nuclei belonging to the same molecule but also the local magnetic fields of nuclei situated in adjacent molecules. As the distance between the individual molecules in the sample varies also because of the Brownian motion the distance  $r$  in Eq. (2.30) becomes time dependent. Hence, the nuclei are subject to a fluctuating magnetic field and the larger the frequency component of these fluctuations at frequency  $\omega_I$  is, the faster the spins will relax.

### 2.2.4 Relaxation by Chemical Shift Anisotropy

Nuclei experience a further magnetic field which is caused by the electron cloud surrounding them. This field is almost always anisotropic such that its magnitude at the location of the nuclei changes when the molecule is turned around by Brownian motion [28]. For protons, however, which only have a single proton this additional field is rather weak. Accordingly, relaxation by chemical shift anisotropy is negligible compared to the relaxation processes induced by dipole-dipole interactions.

### 2.2.5 Relaxation by Indirect Spin-Spin Coupling

Spin-spin coupling may occur between two atoms of a molecule which are connected via an electron pair [28]. In the following this mechanism will be described considering as example the hydrofluoric acid molecule HF. If at a certain time one of the two binding electrons  $e_1$  is close to proton  $A$  then the probability for a parallel orientation of  $A$  and  $e_1$  is higher than that for antiparallel orientation. This holds as electrons possess a negative magnetic moment while protons have a positive one. Likewise, the second electron  $e_2$  will orient itself antiparallel to the nucleus  $X$  of the fluorine atom. Because of the Pauli exclusion principle, however, the spins of the two electrons  $e_1$

and  $e_2$  must be oriented antiparallel with respect to each other such that an indirect coupling between the nuclear spins  $A$  and  $X$  results. As this interaction is independent of the orientation of the molecule it is sufficient to quantify it by the scalar  $J_{AX}$  which is called spin-spin coupling constant. Thus, no time dependent interactions between  $A$  and  $X$  can appear by means of Brownian molecular movement.

Time dependency of  $J_{AX}$  may occur, however, if one of the two nuclei participates at chemical exchange processes, i.e. if for instance a proton of the molecule under investigation is exchanged by a proton of the solvent (e.g. water). This process is usually called scalar relaxation of the first type whereas the corresponding correlation time equals the reciprocal exchange rate.

Furthermore, the local magnetic field which is caused by nuclear spin  $X$  at the location of nuclear spin  $A$  can be modulated if nuclear spin  $X$  relaxes. This is called scalar relaxation of the second type and the corresponding correlation time equals the relaxation time of nuclear spin  $X$ . Type two scalar relaxation is particularly important for nuclear spins  $A$  that are coupled to a nuclear spin  $X$  with a spin quantum number  $I \geq 1$ .

## 2.3 Experimental Methods

### 2.3.1 Data Acquisition

In simple  $^1\text{H}$ -NMR experiments the magnetization  $\mathbf{M}$  is perturbed from its equilibrium position by  $90^\circ$  pulse before the time course of the transversal magnetization is recorded. This detection is achieved by means of an additional coil pointing in the  $x$ -direction in which the rotating magnetization induces a voltage  $U_x$ . If all protons in the sample have identical chemical shifts then the coil detects an exponentially decreasing cosine function with decay constant  $T_2^*$  and Larmor frequency  $\omega_{\mathbf{I}}$

$$U_x(t) \propto M_0 \exp(-t/T_2^*) \cos(\omega_{\mathbf{I}}t) \quad (2.31)$$

whereas  $M_0$  denotes, as before, the magnitude of the equilibrium magnetization.

A phase shifter is then used to generate a second signal  $U_y$  which is phase shifted with respect to  $U_x$  by  $\pi/2$ :

$$U_y(t) \propto M_0 \exp(-t/T_2^*) \cos(\omega_{\mathbf{I}}t + \pi/2) = M_0 \exp(-t/T_2^*) \sin(\omega_{\mathbf{I}}t). \quad (2.32)$$

This signal corresponds to the voltage which would be measured if an additional detection coil pointing in the  $y$ -direction was used. Typically,

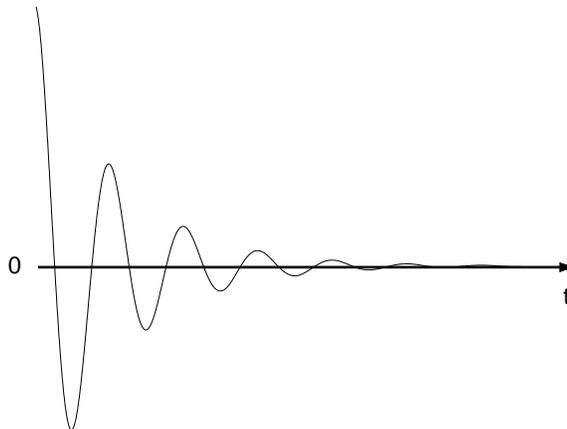


Figure 2.3: Real part of an ideal FID.

Eq. (2.31) and Eq. (2.32) are combined to a complex valued signal

$$U(t) = M_0 \exp(-t/T_2^*) \exp(i\omega_{\mathbf{I}}t) \quad (2.33)$$

such that  $U_x$  corresponds to the real part and  $U_y$  to the imaginary part of  $U(t)$ . This signal is usually termed free induction decay (FID), see Fig. 2.3.

Such pure FIDs are not obtained if larger molecules are investigated which contain several nonequivalent protons, i.e. protons in different chemical environments with varying chemical shifts and  $T_2^*$  times. In this case, the recorded signal  $U(t)$  is a superposition of the FIDs originating from the individual nonequivalent protons

$$U(t) \propto \sum_{k=1}^p M_{0,k} \exp(-t/T_{2,k}^*) \exp(i\omega_{\mathbf{I},k}t) \quad (2.34)$$

whereas  $p$  denotes the number of nonequivalent protons,  $T_{2,k}^*$  their individual relaxation times and  $\omega_{\mathbf{I},k}$  their specific larmor frequencies.

For the further analysis these signals have to be digitized first. However, the proton resonance frequencies obtained by modern spectrometers are in the range of 400 to 800 MHz and thus exceed the conversion capabilities of common 32 bit analogue digital converters (AD converters). For this reason a mixer is used to subtract the so-called receiver reference frequency  $\omega_{ref}$  from the recorded signal [30]. If the proton resonance frequencies lie in the range  $\omega$  to  $\omega + \Delta\omega$  then the reference frequency  $\omega_{ref}$  is usually set to  $\omega + \Delta\omega/2$ , i.e. to the middle of proton frequency range. Thus, protons which precess with a frequency lower than  $\omega_{ref}$  seem to rotate with a negative angular rate, otherwise their frequency is positive.

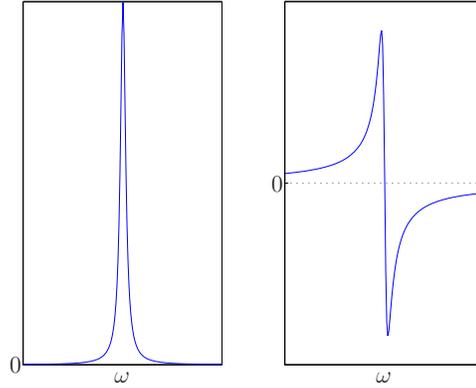


Figure 2.4: Ideal Lorentzian absorption (left) and dispersion peak (right).

After subtraction of the reference frequency  $\omega_{ref}$  the signals can be digitized by means of an AD converter which saves the amplitudes of the frequency reduced signals (denoted by  $\tilde{U}$  in the sequel) in time intervals of size  $t_{samp}$ . The largest frequency that can be determined uniquely in this way is called the *Nyquist* frequency and is determined by

$$f_{Nyquist} = \frac{1}{2t_{samp}} \quad (2.35)$$

Hence, the frequency range  $[-f_{Nyquist}, f_{Nyquist}]$  can be measured by means of the presented detection method.

### 2.3.2 Fourier Transformation

As mentioned in the last section the recorded signal  $U(t)$  and hence also the frequency reduced signal  $\tilde{U}(t)$  are the combinations of the FIDs originating from the individual proton groups contained in the sample. In order to determine the resonance frequencies and concentrations of these proton groups the signal  $\tilde{U}(t)$  is Fourier transformed [28]:

$$F(\omega) = \int_{-\infty}^{\infty} \tilde{U}(t) \exp^{-i\omega t} dt. \quad (2.36)$$

The real part of the spectrum  $F(\omega)$  then contains Lorentzian shaped absorption peaks

$$L_{abs,k}(\omega) = C_k \frac{\Delta\omega_{1/2,k}^2}{\Delta\omega_{1/2,k}^2 + \Delta\omega_k^2}. \quad (2.37)$$

at the frequencies  $\omega_{\mathbf{I},k}$ ,  $k = 1, \dots, p$ , of the  $p$  different proton groups contained in the sample.

In contrast, the imaginary part of  $F(\omega)$  consists of dispersion peaks [34]

$$L_{disp,k}(\omega) = C_k \frac{\Delta\omega_{1/2,k}^2 \Delta\omega_k}{\left(\Delta\omega_{1/2,k}^2 + \Delta\omega_k^2\right)^2}. \quad (2.38)$$

(see Fig. 2.4).

Here,  $\Delta\omega_k = \omega_{\mathbf{I},k} - \omega_{ref}$  while  $\Delta\omega_{1/2,k}$  denotes the full width at half maximum of the peaks. The latter is via

$$\Delta\omega_{1/2,k} = \frac{1}{T_{2,k}^*} \quad (2.39)$$

related with the transversal relaxation time  $T_{2,k}^*$ . Finally,  $C_k$  denotes the magnitude of the  $k$ -th peak which is proportional to the concentration of the  $k$ -th individual proton group in the sample.

Finally, a normalization of the frequency axis of the spectra is necessary in order to be able to compare data recorded by different NMR spectrometer. The reason is that because of Eq. (2.3) the frequencies detected by the spectrometer depend on the magnitude of the used external magnetic field  $\mathbf{B}_0$ . A scaling  $\delta$  of the frequency axis which is independent of  $\mathbf{B}_0$  can be obtained, though, if the following normalization is used [30]

$$\delta = \frac{f_{sig} - f_{ref}}{f_{ref}} \cdot 10^6. \quad (2.40)$$

Here,  $f_{sig}$  is the frequency of the measured signal and  $f_{ref}$  is the resonance frequency a standard substance possesses when exposed to the used external magnetic field  $\mathbf{B}_0$ . The normalized frequency  $\delta$  is called *chemical shift* and is expressed in units of parts per million (ppm).

### 2.3.3 Spectra Postprocessing

Until now it has been assumed implicitly that  $M_x(t)$  and  $M_y(t)$  carry out dampened cosine and sine oscillations with phase  $\Phi = 0$  or, in other words, that  $M_x(t = 0)$  is maximal while  $M_y(t = 0)$  vanishes. If these assumptions hold the real and imaginary part of the spectra obtained by means of Eq. (2.37) and Eq. (2.38) contain pure absorption and dispersion peaks, respectively. For technical reasons, however, it is impossible to turn on the detection coil exactly at time  $t = 0$  such that nonvanishing phase shifts  $\Phi \neq 0$  are observed. Accordingly, the computed spectra contain mixtures

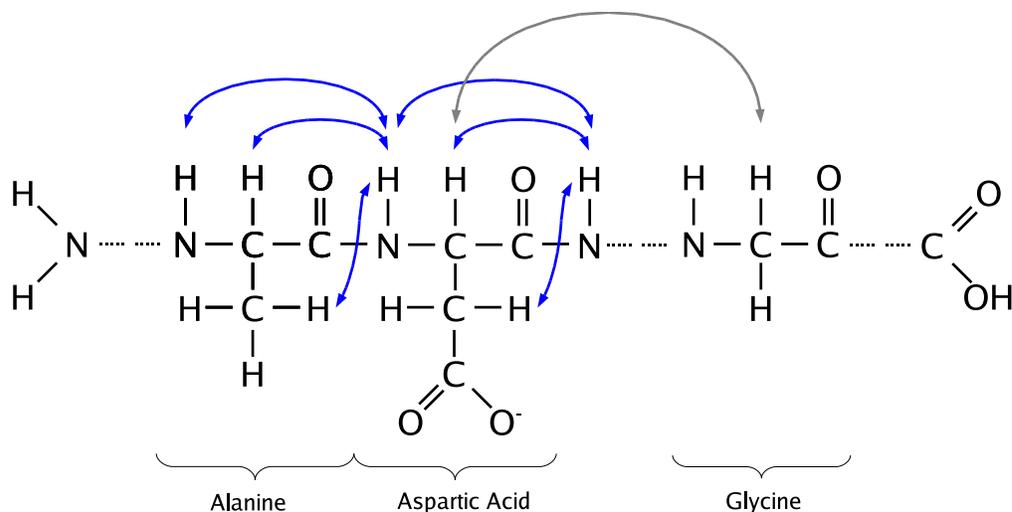


Figure 2.5: Fictitious protein consisting of several amino acids. Blue arrows indicate distances between neighboring amino acids which are commonly measured in 2D-NOESY experiments. Proteins are often twisted such that some of their protons can be spatially close even if they are far apart along the protein sequence (e.g. the protons of Aspartic acid and Glycine which are connected by the grey arrow). Such proteins can also be detected and the distance between them can be determined by 2D-NOESY.

of absorption and dispersion peaks. To amend this flaw the recorded time signal is multiplied by an exponential function  $e^{i\phi_{corr}}$ , where  $\phi_{corr} = -\Phi$ , in a postprocessing step called phase correction [30].

Furthermore, the recorded FIDs are often multiplied by a window function before they are Fourier transformed. If an exponential function with negative decay constant is used as a window function then the signal to noise ratio is increased at the expense of an increasing line width. If importance is rather attached to small line widths then a Gaussian function is often used, which, however, leads to an decreasing signal to noise ratio.

## 2.4 3D Structure Determination of Proteins

In recent years enormous efforts have been made in order to decode the human genome. Although these investigations have now lead to the identification of all coding parts (called genes) of the human DNA the enormous challenge to elucidate the function of these genes still remains. One way to tackle this task is to investigate the function of the proteins which are synthesized according to the blueprint encoded in the corresponding genes.

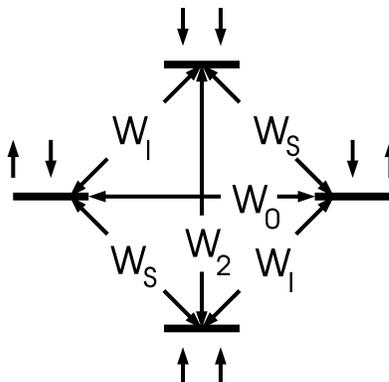


Figure 2.6: Energy levels of two dipolar coupled nuclei. The external magnetic field  $\mathbf{B}_0$  is assumed to point in the positive  $z$ -direction. See text for details.

For this purpose, the spatial structure of the proteins is often determined as it is directly linked with their functionality. These structure determinations are often carried out by means of NMR as it allows the investigation of proteins under quasi-physiological conditions. An NMR experiment which is typically used for this purpose is 2D Nuclear Overhauser Effect Spectroscopy (2D-NOESY) which is capable of detecting spatially close molecules even if they are far apart along the protein sequence (see Fig. 2.5).

### 2.4.1 Stationary Nuclear Overhauser Effect

NMR spectroscopy based on the Nuclear Overhauser Effect (NOE) between dipolar coupled nuclei is one of the most commonly used technique to determine the spatial structure of proteins [41]. Generally speaking the NOE describes the transfer of spin polarization from one spin population to another. The magnitude of the NOE highly depends on the distance between the dipolar coupled spins such that it can be used to determine interatomic distances within molecules.

To illustrate the NOE consider as example a molecule in which two nuclei  $I$  and  $S$  with nuclear spin  $1/2$  are dipolar coupled. Fig. 2.6 depicts the energy levels of such a system if the static magnetic field  $\mathbf{B}_0$  points in the positive  $z$ -direction. The scalars  $W$  in Fig. 2.6 denote the transition probabilities between the individual energy levels. In thermal equilibrium the populations of these energy levels are subject to the Boltzmann probability function.

In the following assume that an ensemble of such molecules is available and let  $\mathbf{M}_0^I$  and  $\mathbf{M}_0^S$  denote the magnetizations resulting from the equilibrium

distributions of the nuclear spins  $I$  and  $S$ , respectively. The equilibrium distribution of the spins  $I$  can now be perturbed selectively by irradiating a perpendicular magnetic HF field  $\mathbf{B}_\perp$  whose frequency equals the Larmor frequency of  $I$ . Because of the dipolar coupling this perturbation also affects the spins  $S$  such that eventually both  $z$ -magnetizations  $M_z^I$  and  $M_z^S$ , which originate from the spins  $I$  and  $S$ , respectively, differ from their equilibrium values  $M_0^I$  and  $M_0^S$ .

Quantitatively, the reaction of  $M_z^S$  to a change of  $M_z^I$  is described by the *Solomon equation* which are obtained by adding a coupling term to the Bloch equation:

$$\frac{dM_z^S}{dt} = -\varrho(M_z^S - M_0^S) - \sigma(M_z^I - M_0^I). \quad (2.41)$$

Here,  $\varrho = (2W_s + W_2 + W_0)$  and  $\sigma = (W_2 - W_0)$  denotes the cross relaxation parameter.

In stationary NMR experiments the HF field with frequency  $\omega^I$  is irradiated until the system reaches a new equilibrium state. This means that the nuclear spins  $I$  are saturated such that  $M_z^I$  vanishes and that  $dM_z^S/dt$  can be set to zero. Hence, the Solomon equation (2.41) now reads

$$\frac{M_z^S - M_0^S}{M_0^S} = \frac{\sigma \gamma_I}{\varrho \gamma_S} \quad (2.42)$$

whereas  $M_z^I/M_z^S = \gamma_I/\gamma_S$  has been used. This is the quantitative equation of the NOE. In homonuclear problems  $\gamma_I = \gamma_S$  holds such that only the cross relaxation parameter  $\sigma$  determines if the NOE enhances or weakens the magnetization  $M_z^S$ . Finally note, that in the cases where  $W_2 = W_0$  the cross relaxation  $\sigma$  and hence also the NOE vanish.

### 2.4.2 Twodimensional NOE Spectroscopy

In order to determine the spacial structure of proteins it is often sufficient to examine which of their protons are dipolar coupled. For this purpose the stationary NOE experiment described above would have to be repeated for all nonequivalent protons. These repetitions become too time consuming for practical applications, however, if large proteins with numerous nonequivalent protons are to be examined.

A solution to this problem is two dimensional Nuclear Overhauser Spectroscopy (2D-NOESY) in which all types of nuclei are investigated simultaneously. In this procedure several FIDs are recorded successively whereas for each FID the experimental protocol is slightly changed. Fig. 2.7 shows the basic pulse sequence applied in a 2D-NOESY experiment. The time  $t_1$

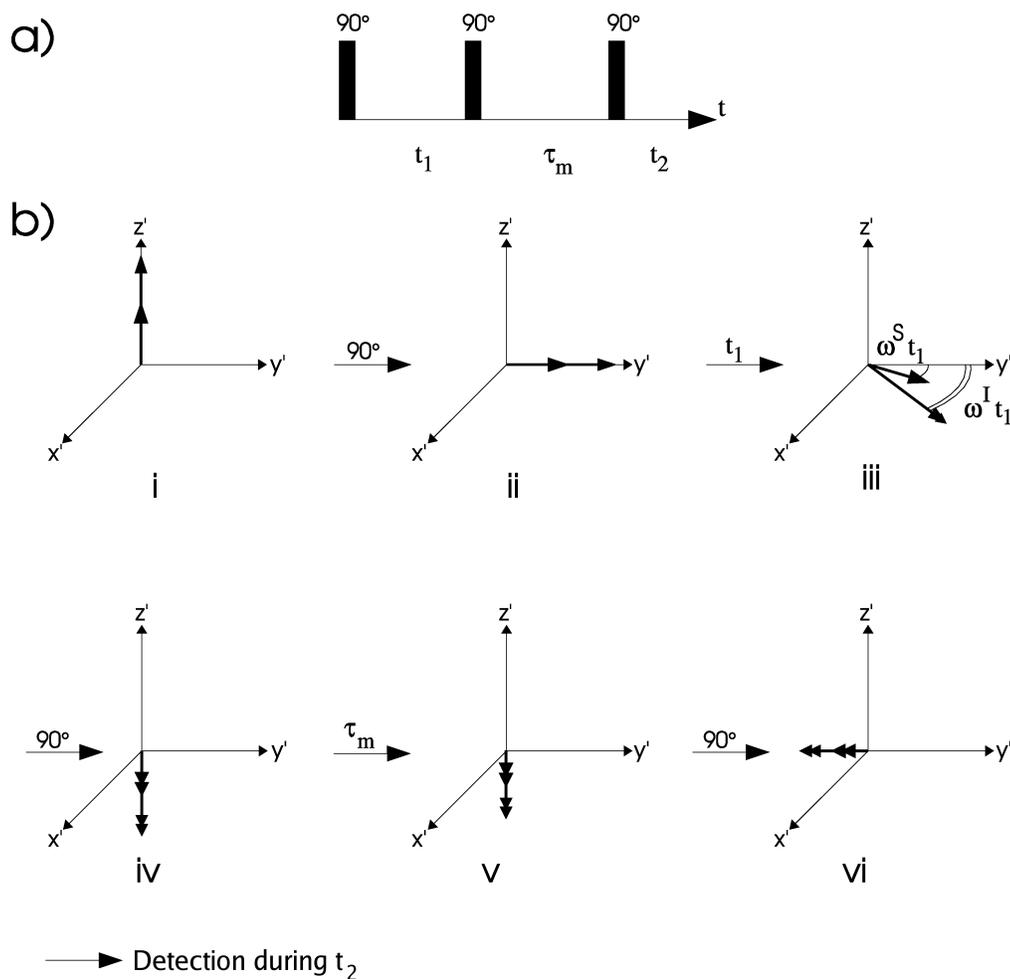


Figure 2.7: a) 2D-NOESY sequence consisting of three  $90^\circ$  pulses. The time  $t_1$  is incremented throughout the experiment while  $\tau_m$  and  $t_2$  are kept constant. b) Impacts of the 2D-NOESY sequence on the magnetizations of two dipolar coupled spins  $I$  and  $S$ . i) Equilibrium situation at the beginning of the experiment.  $\mathbf{M}^I$  and  $\mathbf{M}^S$  (single headed arrows) point in the positive  $z'$ -direction. ii)  $\mathbf{M}^I$  and  $\mathbf{M}^S$  are turned into the  $x'y'$ -plane by the first  $90^\circ$  pulse. iii) During  $t_1$   $\mathbf{M}^I$  and  $\mathbf{M}^S$  cover different angles  $\omega^S t_1$  and  $\omega^I t_1$ , respectively. iv) The second  $90^\circ$  pulse turns the  $y'$ -components  $\tilde{\mathbf{M}}^I$  and  $\tilde{\mathbf{M}}^S$  (depicted by double headed arrows) into the negative  $z'$ -direction. v) During  $\tau_m$  the magnetizations are altered because of the NOE. vi) The last  $90^\circ$  pulse turns the magnetisations into the negative  $y'$ -direction. Their precessions in the  $x'y'$ -plane are jointly detected during  $t_2$ .

separating the first and the second 90° pulse is called evolution time and is incremented from FID to FID. In contrast, the time between the second and the third 90° pulse, called mixing time  $\tau_m$ , and the time  $t_2$  during which the FID is eventually recorded are kept fixed throughout the experiment.

In order to illustrate the fundamental idea of 2D-NOESY consider an ensemble of molecules which contain two dipolar coupled nuclei  $I$  and  $S$ , respectively. Let  $\omega^I$  and  $\omega^S$  denote the Larmor frequencies of these two nuclei and assume

$$\omega^I > \omega^S \quad (2.43)$$

without loss of generality. The magnetisations originating from these two types of spins will be denoted by  $\mathbf{M}^I$  and  $\mathbf{M}^S$  with equilibrium magnitudes  $M_0^I$  and  $M_0^S$ , respectively, in the following.

Before the onset of the experiment the system is in thermal equilibrium such that  $\mathbf{M}^I$  and  $\mathbf{M}^S$  point in the positive  $z$ -direction (see Fig. 2.7b i). These magnetisations are turned into the positive  $y'$  direction by the first 90° pulse (see Fig. 2.7b ii). Once arrived there they start to precess with their Larmor frequencies  $\omega^I$  and  $\omega^S$  in the  $xy$ -plane and cover angles of  $\omega^I t_1$  and  $\omega^S t_1$ , respectively, during evolution time  $t_1$  (see Fig. 2.7b iii).

The second 90° pulse turns the remaining  $y'$ -components of  $\mathbf{M}^I$  and  $\mathbf{M}^S$  into the negative  $z'$ -direction while the  $x'$ -components remain unchanged. The latter are made vanish, however, by a series of additional pulses (omitted in Fig. 2.7) such that eventually two magnetisations  $\tilde{\mathbf{M}}^I$  and  $\tilde{\mathbf{M}}^S$  are obtained which point in the negative  $z'$  direction and have magnitudes of

$$\tilde{M}^I = M_0^I \cos(\omega^I t_1) \quad (2.44)$$

$$\tilde{M}^S = M_0^S \cos(\omega^S t_1) \quad (2.45)$$

respectively (see Fig. 2.7b iv). Finally,  $\tilde{\mathbf{M}}^I$  and  $\tilde{\mathbf{M}}^S$  are turned into the negative  $y'$  direction by the last 90° pulse whereupon they start to precess in the  $xy$ -plane. During  $t_2$  these precessions induce an FID-signal  $U(t_1, t_2)$  in the detection coil of which the Fourier transform

$$U(t_1, \omega_2) = \int_{-\infty}^{\infty} U(t_1, t_2) e^{-i\omega_2 t_2} dt_2 \quad (2.46)$$

is computed. The real part

$$u_{\omega_2}(t_1) = \text{Re}\{U(t_1, \omega_2)\} \quad (2.47)$$

of this Fourier transform contains (after appropriate phase correction) two absorption peaks at frequencies  $\omega^I$  and  $\omega^S$  whose amplitudes are proportional to  $\tilde{M}^I$  and  $\tilde{M}^S$  if  $\tau_m$  is set to zero.

This procedure is reiterated several times whereas the evolution time  $t_1$  is incremented from repetition to repetition. The incrementations of  $t_1$  lead to modulations of the absorption peaks with frequencies of  $\omega^I$  and  $\omega^S$ , respectively (cf. Eq. (2.44) and Eq. (2.45)).

In order to detect dipolar coupled spins by 2D-NOESY nonvanishing mixing times  $\tau_m$  have to be used. To better demonstrate the impacts of a nonzero  $\tau_m$  assume that  $t_1$  is chosen such that

$$\omega^I t_1 = \frac{\pi}{2} \quad (2.48)$$

holds. In this case  $\mathbf{M}^I$  points in the positive  $x'$ -direction and  $\tilde{M}^I$  vanishes after  $t_1$  (cf. Eq.(2.44)) while  $\tilde{M}^S$  is nonzero (cf. Eq. (2.45)). Principally this situation is similar to that occurring during a stationary NOE experiment, though, two differences appear: on the one hand the nuclear spins  $I$  are saturated after the second  $90^\circ$  pulse as in a stationary NOE experiment, however, there is no external field during  $\tau_m$  which maintains this saturation. On the other hand the nuclear spins  $S$  are already perturbed from their equilibrium distribution. This means that both  $\tilde{M}^I$  as well as  $\tilde{M}^S$  are altered by the Nuclear Overhauser Effect during  $\tau_m$  (in particular,  $\tilde{M}^I$  will be nonzero after  $\tau_m$ ). The more the spins  $S$  have been perturbed from their equilibrium distribution, i.e. the larger the difference

$$\delta^S = \tilde{M}^S - M_0^S \quad (2.49)$$

the more  $\tilde{M}^I$  will change during  $\tau_m$  (cf. Fig. 2.7 v). Thus, as  $\delta^S$  depends on  $\omega^S$  according to Eq. (2.45) also  $\tilde{M}^I$  becomes  $\omega^S$ -dependent. The same considerations hold for the modifications of  $\tilde{M}^S$ , i.e. also  $\tilde{M}^S$  depends on  $\omega^I$  because of the NOE.

In the real part  $u_{\omega_2}(t_1)$  of the spectra  $U(t_1, \omega_2)$  the amplitudes of the peaks at  $\omega^I$  and  $\omega^S$  thus oscillate with *both* frequencies  $\omega^I$  and  $\omega^S$  for increasing  $t_1$ -times. These  $t_1$ -dependent modulations are further analyzed by means of a second Fourier transform of  $u_{\omega_2}(t_1)$  along  $t_1$ . For this purpose, additional experimental methods are needed in order to be able to discriminate between positive and negative frequencies (see. App.). The resulting spectra contains four peaks: two diagonal peaks  $U(\omega_1 = \omega^I, \omega_2 = \omega^I)$  and  $U(\omega_1 = \omega^S, \omega_2 = \omega^S)$  which originate from the incrementation of  $t_1$  throughout the experiment (cf. Eq. (2.44), Eq.(2.45)) and two cross peaks  $U(\omega_1 = \omega^I, \omega_2 = \omega^S)$  and  $U(\omega_1 = \omega^S, \omega_2 = \omega^I)$  which stem from the NOE between the dipolar coupled spins  $S$  and  $I$  (see Fig. 2.8).

For large molecules with  $N_{sur}$  nonequivalent protons the following 2D-NOESY plot is obtained: first,  $N_{sur}$  diagonal peaks  $U(\omega_1 = \omega^k, \omega_2 = \omega^k)$ ,

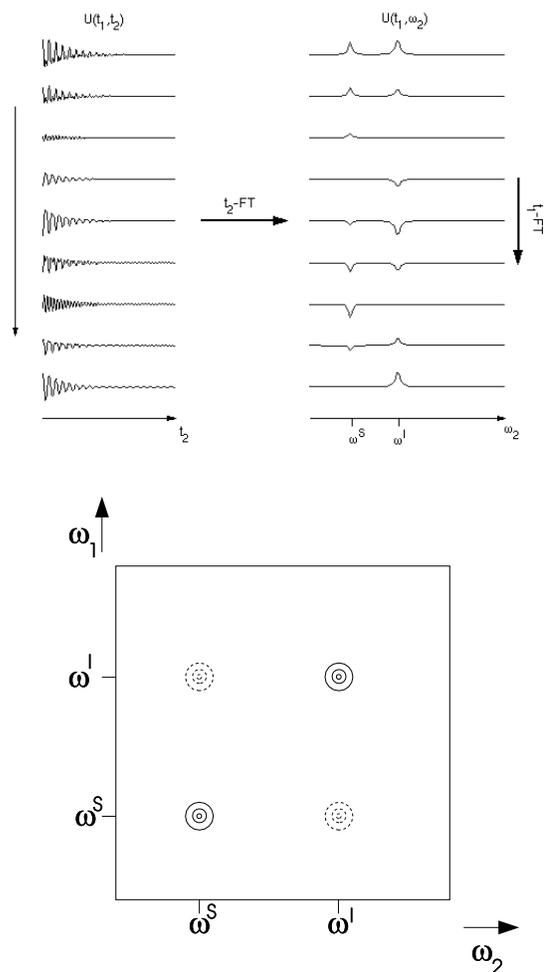


Figure 2.8: Generation of a 2D-NOESY plot considering as example two dipolar coupled spins  $I$  and  $S$ . Top: For each evolution time  $t_1$  the FIDs  $U(t_1, t_2)$  are recorded during the time  $t_2$ . After the first Fourier transform (FT) the spectra  $U(t_1, \omega_2)$  are obtained which contain peaks at the Larmor frequencies  $\omega^I$  and  $\omega^S$  of the spins  $I$  and  $S$ , respectively. The modulations of these peaks originate from the incrementation of the evolution time  $t_1$  throughout the experiment as well as from the NOE between the spins (see text). Bottom (Contour plot): If the spins  $I$  and  $S$  are not dipolar coupled only the diagonal peaks at  $(\omega^I, \omega^I)$  and  $(\omega^S, \omega^S)$  can be observed after the second Fourier transform. The dashed crosspeaks at  $(\omega^I, \omega^S)$  and  $(\omega^S, \omega^I)$  appear additionally if  $I$  and  $S$  interact via the NOE.

$k = 1, \dots, N_{sur}$ , appear. Furthermore, if the  $k$ -th nucleus is dipolar coupled with  $N_k$  other spins, then  $N_k$  cross peaks  $U(\omega_1 = \omega^k, \omega_2 = \omega^j)$ ,  $1 \leq j \leq N_k$  can be found additionally whereas  $\omega^j$  denotes the Larmor frequency of the dipolar coupled spin. Generally, cross peaks can only be observed in the 2D plot if the distance between the interacting nuclei is smaller than 5 nm. Furthermore, the volume of the cross peaks decreases with the sixth power of the distance between the nuclei, a property which allows a good quantitative determination of internuclear distances.

## 2.5 Conclusions

This chapter provided a short review of the physical concepts of nuclear spins in external magnetic fields. It was explained how these properties can be exploited advantageously in NMR spectroscopy to determine the spatial structure of proteins. Special attention was paid to 2D-NOESY, not only because it is the standard NMR procedure to detect interatomic distances, but also because spectra stemming from this type of experiment will be used throughout the remainder of this thesis. It may be noted that a variety of other two-dimensional NMR experiments have been developed and are frequently used in the structure determination of proteins. For a concise overview of these methods the reader is referred to [61], [62], a more detailed explanation can be found in e.g. [28].

# Chapter 3

## Water Signal Removal by BSS

A general problem in the structure determination of proteins is that the latter often change their conformation when they are removed from their natural surroundings. For this reason the proteins investigated by means of NMR are often dissolved in water in order to expose them to at least quasi-physiological conditions during the experiment. It is, however, difficult to produce large amounts of proteins such that their concentration in water is usually very low. Thus, in NMR experiments in which proton resonances are determined (so-called  $^1\text{H}$ -NMR experiments) the signal originating from the water protons is several magnitudes higher than that stemming from the protein protons. Given the limited resolution of the AD converter this means that the protein signals cannot be detected sufficiently well to analyze them quantitatively. Currently, this problem is partly solved by saturating the resonances of the water protons before the FID is recorded. As can be seen in Fig. 3.1, however, the remaining water signal is still dominating the recorded spectra, overlapping protein peaks and additionally leading to severe baseline distortions.

In the following, it will be shown how this remaining water signal can be removed from the spectra by means of blind source separation (BSS). The used BSS algorithm is based on second order statistics and is particularly fast and easy to apply. However, as will be seen, this procedure also leads to additional noise in the spectra.

### 3.1 The Water Signal

In order to reduce the water signal in  $^1\text{H}$ -NOESY experiments the water protons are saturated by a continuous wave electromagnetic field oscillating at the Larmor frequency of the water protons. This field is irradiated during

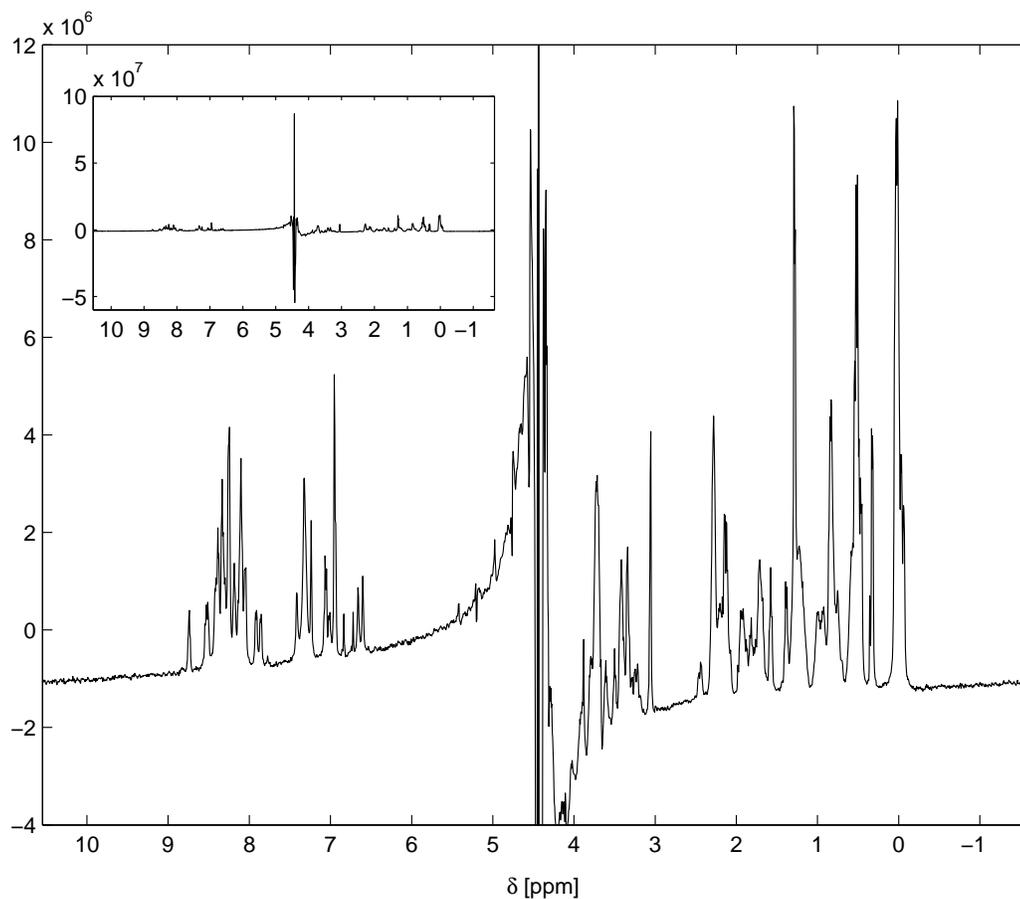


Figure 3.1: The  $^1\text{H}$ -NMR spectrum of the protein P11 as recorded during a NOESY experiment ( $t_1 = 0$ ). Insert: the whole NMR spectrum. Despite saturation of the water protons the spectrum is dominated by a strong water signal at 4.765 ppm which is about 10 times larger than the peaks originating from the protein protons left and right of it. Main figure: blow-up of the spectrum depicted in the insert. The water signal leads to significant baseline distortions and overlaps some of the protein peaks.

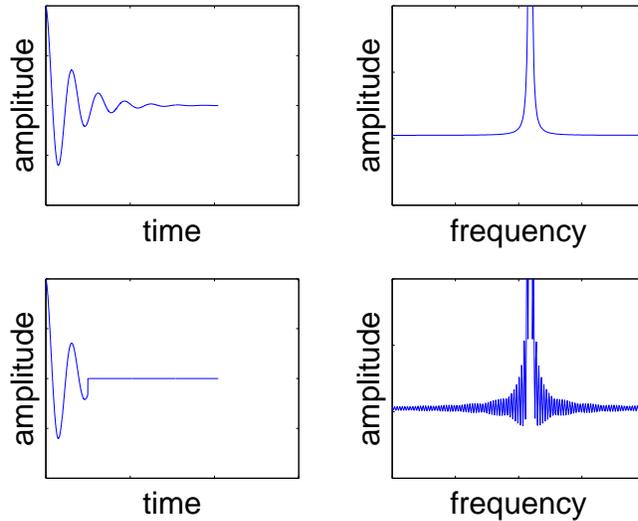


Figure 3.2: Baseline distortions caused by too small  $t_2$ -times. Top:  $t_2$  is long enough for the FID to fade away. Accordingly, a smooth baseline is obtained after Fourier transform. Bottom: too short  $t_2$ -times lead to truncated FIDs. In such cases baseline oscillations are observed in the spectra.

the evolution time  $t_1$  as well as during the mixing time  $t_m$  and leads apart from a suppression of the water signal also to a phase shift between the magnetizations originating from the water and the protein protons. During phase correction (cf. Sec. 2.3.3) only the phases of the protein signals are adjusted, however, such that a mixture of an absorption and a dispersion peak appears at the resonance frequency of the water protons (cf. Fig. 3.1).

Another problem is the comparatively large  $T_{2,H_2O}^*$  time of the water protons as the maximal time span a complete 2D-NOESY experiment may take is confined by the life time of the protein under investigation. This limit often requires that the  $t_2$  time during which the individual FIDs are recorded has to be reduced so drastically that at its end the  $xy$ -magnetization originating from the water protons has not vanished yet [30].

Mathematically, this means that the signal

$$U_{H_2O}(t) \propto \exp(-t/T_{2,H_2O}^*) \exp(i\omega_{H_2O}t) \quad (3.1)$$

induced by the magnetization of the water protons is multiplied by the rectangular function

$$h(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq t_2 \\ 0 & \text{else} \end{cases} \quad (3.2)$$

of width  $t_2$ , i.e. the function

$$U^{H_2O}(t)h(t) \quad (3.3)$$

is recorded. Fourier transforming this function and making use of the convolution theorem leads to spectra of the form

$$\int_{-\infty}^{\infty} U^{H_2O}(t)h(t)e^{-i\omega t} dt = \hat{U}^{H_2O}(\omega) \otimes \hat{h}(\omega) \quad (3.4)$$

whereas  $\hat{U}_{H_2O}(\omega)$  and

$$\hat{h}(\omega) = \frac{\sin \omega}{\omega} \quad (3.5)$$

are the Fourier transforms of  $U^{H_2O}(t)$  and  $h(t)$ , respectively, and  $\otimes$  denotes a convolution. In the obtained spectrum the convolution with  $\hat{h}(\omega)$  leads to additional oscillations left and right of the water peak which distort the baseline and exacerbate a quantitative analysis of the protein peaks (see Fig. 3.2).

These oscillations can be suppressed by multiplying the FID by a rapidly decaying window function whereas, however, the line width or the signal to noise ratio are affected. Such a procedure has been used during the computation of the spectrum shown in Fig. 3.1 such that no baseline oscillations appear there.

## 3.2 Removal of water signal by BSS

Although the experimental water suppression techniques significantly reduce the water signal in NMR experiments the remaining peak still dominates the resulting spectra and leads to severe baseline distortions. Hence, sophisticated postprocessing procedures are needed which are capable of removing the water remnant without influencing the shape or intensity of the protein peaks. A promising approach to this problem is linear Blind Source Separation (BSS) in which the signals recorded during a 2D-NOESY experiment are decomposed into protein and water related components. The latter are then deliberately set to zero such that only protein related signals remain.

### 3.2.1 Linear Blind Source Separation

In linear BSS it is assumed that the observed signals are instantaneous linear mixtures of some unknown underlying sources signals. The goal of a BSS analysis is to recover these sources as well as the weights with which

they contributed to the observations. Such a recovery is already considered successful, though, if the individual sources are determined up to a scaling factor. Furthermore, the order in which the sources are recovered may be arbitrary.

The linear BSS problem can be formalized by means of the following equation

$$\mathbf{X} = \mathbf{A}\mathbf{S} \quad (3.6)$$

whereas the rows of the  $M \times T$  *mixture matrix*  $\mathbf{X}$  contain the observations, the underlying sources constitute the rows of the  $N \times T$ -*source matrix*  $\mathbf{S}$  and the weights with which the  $n$ -th source contributes to the  $m$ -th observation is stored in the element  $a_{mn}$  of the  $M \times N$ -*mixing matrix*  $\mathbf{A}$ .

Based on this notation the objective of a BSS analysis can be restated as that of finding a factorization of the observation matrix  $\mathbf{X}$  into two matrices  $\mathbf{A}^{est}$  and  $\mathbf{S}^{est}$  which equal  $\mathbf{A}$  and  $\mathbf{S}$ , respectively, up to the following two indeterminacies:

1.  $\mathbf{A}^{est}$  and  $\mathbf{S}^{est}$  may differ from  $\mathbf{A}$  and  $\mathbf{S}$ , respectively, by an  $N \times N$  diagonal matrix  $\mathbf{D}$  with diagonal elements  $d_{nn} \neq 0$ :

$$\begin{aligned} \mathbf{A} &= \mathbf{A}^{est} \mathbf{D} \\ \mathbf{S} &= \mathbf{D}^{-1} \mathbf{S}^{est}. \end{aligned} \quad (3.7)$$

This corresponds to the scaling indeterminacy allowed in the BSS analysis. Note, that in the case of complex  $\mathbf{X}$  also the elements of  $\mathbf{D}$  are complex valued, i.e. the  $d_{nn}$  are of the form

$$d_{nn} = r_n \exp(i\phi_n) \quad (3.8)$$

whereas  $i = \sqrt{-1}$ ,  $r_n \in \mathbb{R}$ , and  $\phi_n \in [0, 2\pi]$ . Implicitly, this means that the sources are not only determined up to a scaling but also up to a phase indeterminacy.

2.  $\mathbf{A}^{est}$  and  $\mathbf{S}^{est}$  may differ from  $\mathbf{A}$  and  $\mathbf{S}$ , respectively, by an  $N \times N$  permutation matrix  $\mathbf{P}$ :

$$\begin{aligned} \mathbf{A} &= \mathbf{A}^{est} \mathbf{P} \\ \mathbf{S} &= \mathbf{P} \mathbf{S}^{est}, \end{aligned} \quad (3.9)$$

These equations reflect the arbitrariness with which the sources may be recovered.

Linear BSS problems are commonly divided into three categories depending on the number of available observations and underlying sources:

- In the quadratic case the number of sources equals the number of observations ( $M = N$ ) and the matrix  $\mathbf{A}$  is assumed to have full rank (i.e.  $\text{rank}(\mathbf{A}) = M$ ). This is the standard BSS problem for which a multitude of different algorithms have been developed.
- In the undercomplete case the number of observations exceeds the number sources ( $M > N$ ) and the rank of  $\mathbf{A}$  is at least  $N$ . This problem can often be transferred to the square case by dimension reduction techniques like principal component analysis (see Sec. 4.1).
- In the overcomplete case less observations than sources are available ( $M < N$ ,  $\text{rank}(\mathbf{A}) = M$ ). This BSS problem is particularly difficult to solve and mathematical proofs of uniqueness are often delicate.

Throughout the remainder of this thesis the quadratic case will always be presumed. Obviously, even this BSS problem is highly underdetermined such that additional constraints are needed in order to recover the source and the mixing matrix up to the above mentioned indeterminacies.

In BSS algorithms based on second order statistics such constraints appear in the shape of uncorrelatedness assumptions on the sources. These are, however, insufficient to solve the BSS problem uniquely such that further constraints like e.g. nonstationarity [11] or autocorrelatedness assumptions [7] [40] [57] [64] on the sources or nonnegativity constraints on both  $\mathbf{A}$  and  $\mathbf{S}$  are needed additionally [44].

Also independent component analysis (ICA) has often been used successfully to recover the source and the mixing matrix [6] [9] [31] [29] [48]. In this procedure the columns of the mixture matrix  $\mathbf{X}$  are considered as the realizations of a random vector  $\mathbf{x} = [x_1, \dots, x_m]^\top$  and the goal is to find an  $M \times M$ -matrix  $\mathbf{W}$  such that the components of the random vector  $\mathbf{y} = [y_1, \dots, y_m]^\top$  determined by

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{3.10}$$

are as statistically independent as possible. Concerning BSS it can be proved [12] [13] that the matrix  $\mathbf{W}$  discovered by ICA equals the inverse of the mixing matrix  $\mathbf{A}$  (up to the indeterminacies described above) if the underlying sources are statistically independent and if at most one of them has a Gaussian distribution. These sources are then readily recovered by multiplying the mixture matrix  $\mathbf{X}$  by  $\mathbf{W}$ , i.e.

$$\mathbf{S}^{est} = \mathbf{W}\mathbf{X}. \tag{3.11}$$

and

$$\mathbf{A}^{est} = \mathbf{W}^{-1}. \tag{3.12}$$

There exist, though, also other approaches in which the underlying sources are allowed to be correlated. For instance, sparse nonnegative matrix factorization can be used in such settings given that both  $\mathbf{A}$  and  $\mathbf{S}$  are nonnegative and that  $\mathbf{S}$  contains several nil entries (i.e. is sparsely represented) [52]. The nonnegativity assumption can also be dropped, however, the columns of  $\mathbf{S}$  must then fulfill rather specific sparseness constraints [17]. Finally, also algorithms have been developed in which general, nonsparse sources are made sparse by representing them over a signal dictionary (e.g. wavelet packets, stationary wavelets etc.) [63]. Here, apart from the observations the correct signal dictionary must be known *a priori* in order to recover the sources and the mixing matrix.

### 3.2.2 Linear BSS in the Context of 2D-NOESY Data

In order to illustrate how the data obtained from  $^1\text{H}$ -NOESY fit into the linear mixture model (Eq. (3.6)) assume that  $M$  individual 1D-FIDs were recorded during the experiment, or, in other words, that  $t_1$  was incremented  $M$  times. For simplicity, let  $M$  also equal the number of unequivalent protons in the sample, i.e. the sum of the number of unequivalent protein protons plus the number of unequivalent solvent protons. According to Eq. (2.34) the  $M$  1D-FIDs  $U(m, t_2)$ ,  $m = 1, 2, \dots, M$ , are of the form

$$U(m, t_2) = \sum_{n=1}^M U_{0,n}(m) \underbrace{\exp(-t/T_{2,n}^*) \exp(i\omega_{\mathbf{I},n}t_2)}_{\text{waveform of the } n\text{-th proton signal}}. \quad (3.13)$$

where  $U_{0,n}(m)$  is the maximal amplitude of the  $n$ -th proton signal ( $1 \leq n \leq M$ ) and  $\omega_{\mathbf{I},n}$  and  $T_{2,n}^*$  are the corresponding Larmor frequency and spin-lattice relaxation time, respectively.

As explained in Sec. 2.4.2 only the amplitudes  $U_{0,n}(m)$  of the individual proton signals vary for different evolution times while their waveforms remain unchanged. This property can be exploited to write the  $M$  equations (3.13) in the matrix notation of Eq. (3.6) whereas the  $m$ -th measured FID  $U(m, t_2)$  is used to constitute the  $m$ -th row of  $\mathbf{X}$ , the waveform of the  $n$ -th proton signal forms the  $n$ -th row of the matrix  $\mathbf{S}$ , and the maximal amplitude  $U_{0,n}(m)$  of the  $n$ -th proton signal given the  $m$ -th  $t_1$ -time are the elements of the matrix  $\mathbf{A}$ , i.e.

$$a_{mn} = U_{0,n}(m). \quad (3.14)$$

This procedure leads to matrices  $\mathbf{X}$  and  $\mathbf{S}$  of size  $M \times T$  if the individual FIDs were sampled at discrete time points  $t_2 = 0, 1, \dots, T - 1$ .

Likewise, the spectra

$$U(m, \omega_2) = \int_{-\infty}^{\infty} U(m, t_2) \exp(-i\omega_2 t_2) dt_2 \quad (3.15)$$

obtained after the first Fourier transform also fit into the linear mixture model. This can be seen directly because of the linearity of the Fourier transform  $\mathcal{F}$ :

$$\mathcal{F}(\mathbf{X}) = \mathcal{F}(\mathbf{A}\mathbf{S}) = \mathbf{A}\mathcal{F}(\mathbf{S}). \quad (3.16)$$

Thus, the rows of the  $M \times T$ -matrix  $\mathcal{F}(\mathbf{X})$  consist of the  $t_2$ -Fourier transformed FIDs  $U(m, \omega_2)$ , the  $n$ -th row of  $\mathcal{F}(\mathbf{S})$  contains a spectrum with a single peak at the Larmor frequency of the  $n$ -th proton while the matrix  $\mathbf{A}$  has the same meaning as in time domain. For simplicity of notation the  $t_2$ -Fourier transformed variables will be labelled with a hat (e.g.  $\hat{\mathbf{X}} := \mathcal{F}(\mathbf{X})$ ) in the following. In practice, the BSS algorithms are applied to frequency domain data as the obtained sources are easier to interpret.

Once the estimates of the source and the mixing matrix have been determined by an appropriate BSS algorithm the water signal can readily be removed. For this purpose the contributions  $\hat{y}_m$  of the individual sources to the first observation  $U(m = 1, \omega_2)$  are computed:

$$\hat{y}_m = a_{1m}^{est} \hat{s}_m^{est}, \quad m = 1, \dots, M, \quad (3.17)$$

where  $\hat{s}_m^{est}$  is the  $m$ -th estimated source as stored in the  $m$ -th row of  $\hat{\mathbf{S}}^{est}$  and the  $a_{1m}^{est}$ 's are the elements of the first row of  $\mathbf{A}^{est}$ . The source signal representing the water protons can then be identified as its corresponding  $\hat{y}$  is the only one among all the  $\hat{y}_m$  which fulfills all of the following criteria:

1. It has a high frequency component at the resonance frequency of the water protons.
2. The amplitude of its peak is significantly larger than that of the other  $\hat{y}_m$ .
3. It contains a dephased peak (all other  $\hat{y}_m$  should contain pure absorption peaks after phase correction, cf. Sec. 2.3.3).

If the  $k$ -th source is identified to contain the water signal the matrix  $\hat{\mathbf{S}}^{est}$  is transferred to the matrix  $\hat{\mathbf{S}}^{protein}$  by setting its  $k$ -th row to zero. Finally, the water free spectra  $\hat{\mathbf{X}}^{protein}$  are obtained by mixing the remaining protein spectra in  $\hat{\mathbf{S}}^{protein}$  by the mixing matrix  $\mathbf{A}^{est}$ :

$$\hat{\mathbf{X}}^{protein} = \mathbf{A}^{est} \hat{\mathbf{S}}^{protein} \quad (3.18)$$

(see Fig. 3.3)

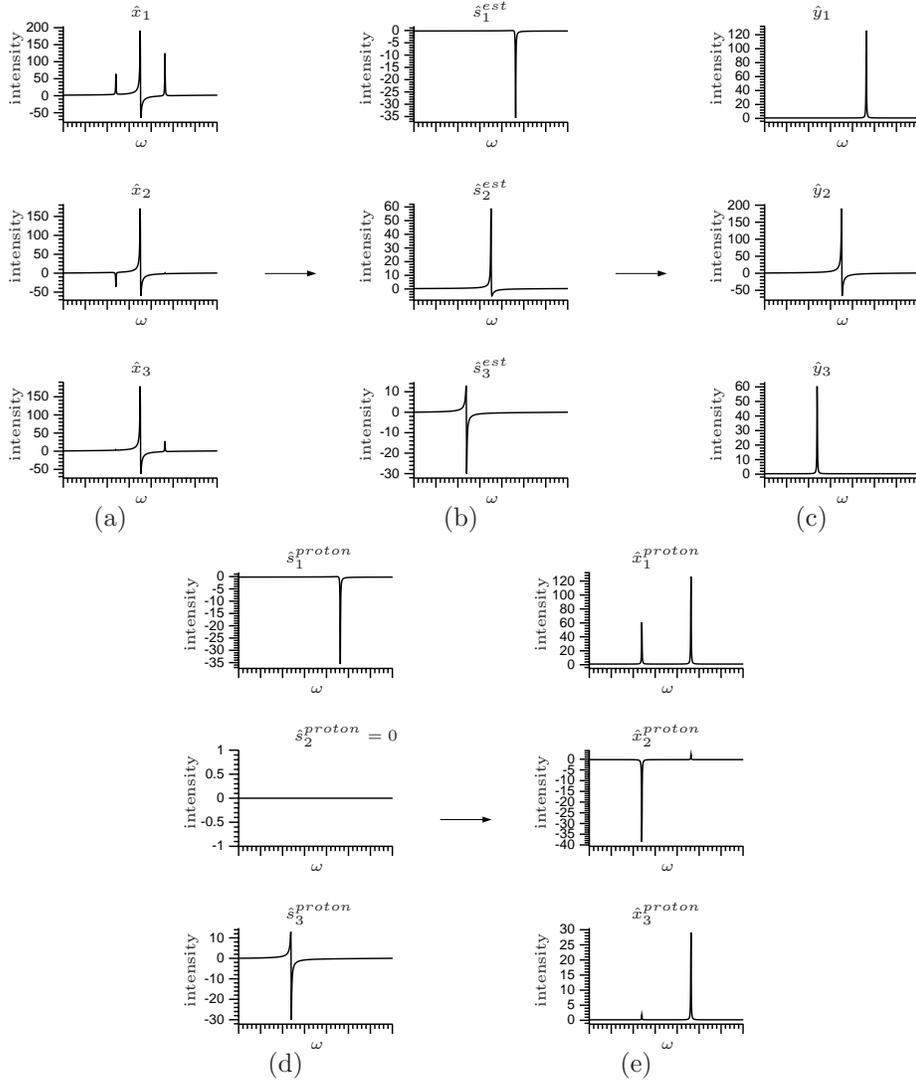


Figure 3.3: BSS based water removal. a) Artificial 2D-NOESY data  $\hat{\mathbf{X}}$  with dominant water peak in the middle and two protein peaks left and right of it. b) Estimated sources  $\hat{\mathbf{S}}^{est}$  obtained by BSS. Because of the scaling indeterminacy (Eq. (3.7)) the peaks have arbitrary amplitude and consist of absorption and dispersion components. c) The contributions  $\hat{y}_m$  of the sources to the first observation. Obviously,  $y_2$  contains the water signal as its peak is located at the resonance frequency of the water protons, has the largest amplitude, and shows some dispersion. d) The matrix  $\hat{\mathbf{S}}^{proton}$  is obtained from the matrix  $\hat{\mathbf{S}}^{est}$  by setting the row containing the water signal to zero. e) The water free spectra  $\hat{\mathbf{X}}^{proton} = \mathbf{A}^{est}\hat{\mathbf{S}}^{proton}$ .

### 3.3 The Matrix Pencil BSS Algorithm

So far it has only been explained in general how BSS can be applied to remove the water signal from 2D-NOESY datasets. In contrast, a concrete BSS algorithm will be presented in this section which is capable of solving the outlined BSS problem. Essentially, this algorithm uses the concept of congruent matrix pencils and joint diagonalization of covariance matrices in order to estimate the underlying sources and the mixing matrix. As I have already shown in my diploma thesis [53] this algorithm outperforms other BSS approaches when applied to NMR datasets.

#### 3.3.1 Congruent Matrix Pencils in BSS

The theoretical backbone of the matrix pencil BSS (MP-BSS) algorithm is the congruence between an observation and a source matrix pencil. In the following, it will be explained which matrices constitute these two matrix pencils and how their congruence can be proved.

For the observation matrix pencil filtered versions of the observed signals are needed. In time domain this filtering is achieved by means of a finite impulse response (FIR) filter which is applied to all observations, i.e. to all rows  $x_m(t) = [x_m(0), x_m(2), \dots, x_m(T-1)]$ ,  $m = 1, 2, \dots, M$ , of the  $M \times T$  observation matrix  $\mathbf{X}$ :

$$x_{F,m}(t) = \sum_{l=0}^{L-1} h(l)x_m(t-l), \quad m = 1, \dots, M. \quad (3.19)$$

Here,  $x_{F,m}$  contains the  $m$ -th filtered observation,  $L$  denotes the order of the filter and the  $h(l) \in \mathbb{C}$ ,  $l = 0, 1, \dots, L-1$ , are the filter coefficients.

In the following, it will be more convenient to express the above filtering process in matrix notation. For this purpose the so-called convolution matrix  $\mathbf{H}$  of the FIR-filter is used. This  $T \times T$ -matrix is of Toeplitz form with the filter coefficients  $h[0], h[1], \dots, h[L-1]$  on its diagonal:

$$\mathbf{H} = \begin{bmatrix} h[0] & 0 & 0 & 0 & \dots & 0 \\ \vdots & h[0] & 0 & 0 & \dots & 0 \\ h[L-1] & \vdots & h[0] & 0 & \dots & 0 \\ 0 & h[L-1] & \vdots & & & \vdots \\ \vdots & 0 & h[L-1] & \ddots & & 0 \\ \vdots & \vdots & 0 & & & h[0] \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & & & h[L-1] \end{bmatrix}. \quad (3.20)$$

With this matrix the filtering process in Eq. (3.19) can be written as

$$\mathbf{X}_F = \mathbf{X}\mathbf{H}^\top \quad (3.21)$$

whereas the rows of the matrix  $\mathbf{X}_F$  contain the  $M$  filtered signals  $x_{F,m}(t)$ .

For both the original as well as the filtered observations the covariance matrices<sup>1</sup>

$$\begin{aligned} \mathbf{C}_\mathbf{X} &= \frac{1}{T-1}(\mathbf{X} - \mathbf{m}_\mathbf{X}\mathbf{1}_T^\top)(\mathbf{X} - \mathbf{m}_\mathbf{X}\mathbf{1}_T^\top)^H \\ \mathbf{C}_{\mathbf{X}_F} &= \frac{1}{T-1}(\mathbf{X}_F - \mathbf{m}_{\mathbf{X}_F}\mathbf{1}_T^\top)(\mathbf{X}_F - \mathbf{m}_{\mathbf{X}_F}\mathbf{1}_T^\top)^H \end{aligned} \quad (3.22)$$

are computed and used to constitute the observation matrix pencil  $(\mathbf{C}_\mathbf{X}, \mathbf{C}_{\mathbf{X}_F})$ .

Likewise, the source matrix pencil is formed. As before, consider filtered versions

$$\mathbf{S}_F = \mathbf{S}\mathbf{H}^\top \quad (3.23)$$

of the original sources which have passed the same filter as the observations. The covariance matrices

$$\begin{aligned} \mathbf{C}_\mathbf{S} &= \frac{1}{T-1}(\mathbf{S} - \mathbf{m}_\mathbf{S}\mathbf{1}_T^\top)(\mathbf{S} - \mathbf{m}_\mathbf{S}\mathbf{1}_T^\top)^H \\ \mathbf{C}_{\mathbf{S}_F} &= \frac{1}{T-1}(\mathbf{S}_F - \mathbf{m}_{\mathbf{S}_F}\mathbf{1}_T^\top)(\mathbf{S}_F - \mathbf{m}_{\mathbf{S}_F}\mathbf{1}_T^\top)^H \end{aligned} \quad (3.24)$$

for both the original and the filtered sources are determined again and are used to form the source matrix pencil  $(\mathbf{C}_\mathbf{S}, \mathbf{C}_{\mathbf{S}_F})$ .

Making use of the linear mixture model Eq. (3.6) and the relation  $\mathbf{m}_\mathbf{X} = \mathbf{A}\mathbf{m}_\mathbf{S}$  between the mean of the rows of  $\mathbf{X}$  and  $\mathbf{S}$ , respectively, it can be shown

---

<sup>1</sup>The following derivations are made for the estimates of the covariance matrices as this leads to the formulas eventually used in the algorithm. Equivalent results are obtained with the exact definitions of covariance matrices.

that the source and the observation matrix pencil are congruent:

$$\begin{aligned}
\mathbf{C}_{\mathbf{X}} &= \frac{1}{T-1} (\mathbf{X} - \mathbf{m}_{\mathbf{X}} \mathbf{1}_T^\top) (\mathbf{X} - \mathbf{m}_{\mathbf{X}} \mathbf{1}_T^\top)^H \\
&= \frac{1}{T-1} (\mathbf{A} \mathbf{S} \mathbf{S}^H \mathbf{A}^H - \mathbf{A} \mathbf{S} \mathbf{1}_T \mathbf{m}_{\mathbf{S}}^H \mathbf{A}^H - \mathbf{A} \mathbf{m}_{\mathbf{S}} \mathbf{1}_T^\top \mathbf{S}^H \mathbf{A}^H \\
&\quad + \mathbf{A} \mathbf{m}_{\mathbf{S}} \mathbf{1}_T^\top \mathbf{1}_T \mathbf{m}_{\mathbf{S}}^H \mathbf{A}^H) \\
&= \frac{1}{T-1} \mathbf{A} (\mathbf{S} - \mathbf{m}_{\mathbf{S}} \mathbf{1}_T^\top) (\mathbf{S} - \mathbf{m}_{\mathbf{S}} \mathbf{1}_T^\top)^H \mathbf{A}^H = \mathbf{A} \mathbf{C}_{\mathbf{S}} \mathbf{A}^H \\
\mathbf{C}_{\mathbf{X}_F} &= \frac{1}{T-1} (\mathbf{X}_F - \mathbf{m}_{\mathbf{X}_F} \mathbf{1}_T^\top) (\mathbf{X}_F - \mathbf{m}_{\mathbf{X}_F} \mathbf{1}_T^\top)^H \\
&= \frac{1}{T-1} (\mathbf{A} \mathbf{S} \mathbf{H}^\top \mathbf{H} \mathbf{S}^H \mathbf{A}^H - \mathbf{A} \mathbf{S} \mathbf{H}^\top \mathbf{1} \mathbf{m}_{\mathbf{S}_F}^H \mathbf{A}^H - \mathbf{A} \mathbf{m}_{\mathbf{S}_F} \mathbf{1}_T^\top \mathbf{H} \mathbf{S}^H \mathbf{A}^H \\
&\quad + \mathbf{A} \mathbf{m}_{\mathbf{S}_F} \mathbf{1}_T^\top \mathbf{1}_T \mathbf{m}_{\mathbf{S}_F}^H \mathbf{A}^H) \\
&= \frac{1}{T-1} \mathbf{A} (\mathbf{S}_F - \mathbf{m}_{\mathbf{S}_F} \mathbf{1}_T^\top) (\mathbf{S}_F - \mathbf{m}_{\mathbf{S}_F} \mathbf{1}_T^\top)^H \mathbf{A}^H = \mathbf{A} \mathbf{C}_{\mathbf{S}_F} \mathbf{A}^H. \quad (3.25)
\end{aligned}$$

Similar results are obtained if the observed data was already transformed to frequency domain (e.g. if the rows of the observation matrix contain  $t_2$ -Fourier transformed spectra of a 2D-NOESY-experiment). In this case the filtering of the source and the observation matrix is achieved by elementwise multiplication with the frequency response of the filter:

$$\hat{\mathbf{X}}_F = \hat{\mathbf{X}} \diamond \hat{\mathbf{H}} \quad \text{and} \quad \hat{\mathbf{S}}_F = \hat{\mathbf{S}} \diamond \hat{\mathbf{H}}. \quad (3.26)$$

Here, “ $\diamond$ ” denotes the Hadamard product and each row of the  $M \times T$ -matrix  $\hat{\mathbf{H}}$  contains the frequency response of the applied filter. After computing covariance matrices in analogy to Eq. (3.22) and Eq. (3.24) the observation matrix pencil ( $\mathbf{C}_{\hat{\mathbf{X}}}, \mathbf{C}_{\hat{\mathbf{X}}_F}$ ) and the source matrix pencil ( $\mathbf{C}_{\hat{\mathbf{S}}}, \mathbf{C}_{\hat{\mathbf{S}}_F}$ ) can be formed. Carrying out identical transformations as for the matrix  $\mathbf{C}_{\mathbf{X}}$  in Eq. (3.25) it can readily be proved that

$$\begin{aligned}
\mathbf{C}_{\hat{\mathbf{X}}} &= \frac{1}{T-1} (\hat{\mathbf{X}} - \mathbf{m}_{\hat{\mathbf{X}}} \mathbf{1}_T^\top) (\hat{\mathbf{X}} - \mathbf{m}_{\hat{\mathbf{X}}} \mathbf{1}_T^\top)^H \\
&= \frac{1}{T-1} (\hat{\mathbf{S}} - \mathbf{m}_{\hat{\mathbf{S}}} \mathbf{1}_T^\top) (\hat{\mathbf{S}} - \mathbf{m}_{\hat{\mathbf{S}}} \mathbf{1}_T^\top)^H = \mathbf{A} \mathbf{C}_{\hat{\mathbf{S}}} \mathbf{A}^H \quad (3.27)
\end{aligned}$$

holds. Furthermore, by noting that  $\mathbf{X} \diamond \mathbf{H} = (\mathbf{A} \mathbf{S}) \diamond \mathbf{H} = \mathbf{A} (\mathbf{S} \diamond \mathbf{H})$  as  $\mathbf{H}$  has identical entries in all of its rows the congruence of  $\mathbf{C}_{\hat{\mathbf{X}}_F}$  and  $\mathbf{C}_{\hat{\mathbf{S}}_F}$  can also

be shown:

$$\begin{aligned}
\mathbf{C}_{\mathbf{X}_F} &= \frac{1}{T-1} (\hat{\mathbf{X}}_F - \mathbf{m}_{\hat{\mathbf{X}}_F} \mathbf{1}_T^\top) (\hat{\mathbf{X}}_F - \mathbf{m}_{\hat{\mathbf{X}}_F} \mathbf{1}_T^\top)^H \\
&= \frac{1}{T-1} (\mathbf{A}(\hat{\mathbf{S}} \diamond \hat{\mathbf{H}})(\hat{\mathbf{H}}^H \diamond \hat{\mathbf{S}}^H) \mathbf{A}^H - \mathbf{A}(\hat{\mathbf{S}} \diamond \hat{\mathbf{H}}) \mathbf{1}_T \mathbf{m}_{\hat{\mathbf{S}}_F}^H \mathbf{A}^H \\
&\quad - \mathbf{A} \mathbf{m}_{\hat{\mathbf{S}}_F} \mathbf{1}_T^\top (\hat{\mathbf{H}}^H \diamond \hat{\mathbf{S}}^H) \mathbf{A}^H + \mathbf{A} \mathbf{m}_{\hat{\mathbf{S}}_F} \mathbf{1}_T^\top \mathbf{1}_T \mathbf{m}_{\hat{\mathbf{S}}_F}^H \mathbf{A}^H) \\
&= \frac{1}{T-1} (\hat{\mathbf{S}}_F - \mathbf{m}_{\hat{\mathbf{S}}_F} \mathbf{1}_T^\top) (\hat{\mathbf{S}}_F - \mathbf{m}_{\hat{\mathbf{S}}_F} \mathbf{1}_T^\top)^H = \mathbf{A} \mathbf{C}_{\mathbf{S}_F} \mathbf{A}^H \quad (3.28)
\end{aligned}$$

Thus, also the matrix pencils  $(\mathbf{C}_{\hat{\mathbf{X}}}, \mathbf{C}_{\hat{\mathbf{X}}_F})$  and  $(\mathbf{C}_{\hat{\mathbf{S}}}, \mathbf{C}_{\hat{\mathbf{S}}_F})$  are congruent.

Eventually, it should be pointed out that in both time and frequency domain the matrix pencils of the sources and the observations are positive definite as they consist of covariance matrices only.

### 3.3.2 The MP-BSS Algorithm

As already mentioned in Sec. 3.2.1 additional constraints on the sources and/or the mixing matrix are needed to solve the linear BSS problem uniquely. In the case of the MP-BSS algorithm presented in this section the following two assumptions are made:

1. The underlying sources to be recovered are uncorrelated, i.e. their covariance matrix is diagonal:

$$\mathbf{C}_{\mathbf{S}} = \frac{1}{T-1} (\hat{\mathbf{S}} - \mathbf{m}_{\hat{\mathbf{S}}} \mathbf{1}_T^\top) (\hat{\mathbf{S}} - \mathbf{m}_{\hat{\mathbf{S}}} \mathbf{1}_T^\top)^H = \Lambda_{\mathbf{S}}, \quad (3.29)$$

where  $\Lambda_{\mathbf{S}}$  is a diagonal matrix <sup>2</sup>

2. The eigenvalues of the observation matrix pencil are non-degenerated.

As shown in the App. A.2 the first assumption means that the eigenvector matrix  $\mathbf{E}_{\mathbf{S}}$  in the GEVD

$$\mathbf{C}_{\mathbf{S}} \mathbf{E}_{\mathbf{S}} = \mathbf{C}_{\mathbf{S}_F} \mathbf{E}_{\mathbf{S}} \Lambda \quad (3.30)$$

of the source pencil  $(\mathbf{C}_{\mathbf{S}}, \mathbf{C}_{\mathbf{S}_F})$  is diagonal.

The second condition ensures (cf. theorem 2) that the eigenvector matrix of the observation matrix pencil  $(\mathbf{C}_{\mathbf{X}}, \mathbf{C}_{\mathbf{X}_F})$  is unique up to scaling and permutation indeterminacies.

---

<sup>2</sup>As shown in App. A.1 diagonality of  $\mathbf{C}_{\mathbf{S}}$  also infers diagonality of the correlation matrix  $\mathbf{C}_{\hat{\mathbf{S}}}$  computed in frequency domain. Hence, all considerations made in the sequel also hold for frequency domain data.

<b>Data:</b> $\mathbf{X}$ <b>Input:</b> $\mathbf{H}$ <b>Result:</b> $\mathbf{A}^{est}, \mathbf{S}^{est}$ <b>begin</b> $\mathbf{X}_F = \mathbf{X}\mathbf{H}^H;$ Compute $\mathbf{C}_\mathbf{X};$ Compute $\mathbf{C}_{\mathbf{X}_F};$ GEVD: $\mathbf{C}_\mathbf{X}\mathbf{E}_\mathbf{X} = \mathbf{C}_{\mathbf{X}_F}\mathbf{E}_\mathbf{X}\mathbf{\Lambda};$ $\mathbf{A}^{est} = (\mathbf{E}_\mathbf{X}^{-1})^H;$ $\mathbf{S}^{est} = (\mathbf{A}^{est})^{-1}\mathbf{X};$ <b>1 end</b>	<b>Data:</b> $\hat{\mathbf{X}}$ <b>Input:</b> $\hat{\mathbf{H}}$ <b>Result:</b> $\mathbf{A}^{est}, \hat{\mathbf{S}}^{est}$ <b>begin</b> $\hat{\mathbf{X}}_F = \hat{\mathbf{X}} \diamond \hat{\mathbf{H}}^H;$ Compute $\mathbf{C}_{\hat{\mathbf{X}}};$ Compute $\mathbf{C}_{\hat{\mathbf{X}}_F};$ GEVD: $\mathbf{C}_{\hat{\mathbf{X}}}\mathbf{E}_{\hat{\mathbf{X}}} = \mathbf{C}_{\hat{\mathbf{X}}_F}\mathbf{E}_{\hat{\mathbf{X}}}\mathbf{\Lambda};$ $\mathbf{A}^{est} = (\mathbf{E}_{\hat{\mathbf{X}}}^{-1})^H;$ $\hat{\mathbf{S}}^{est} = (\mathbf{A}^{est})^{-1}\hat{\mathbf{X}};$ <b>end</b>
---	---

**Algorithm 2:** The MP-BSS algorithm for time domain (left) and frequency domain data (right).

Now recall from the last section that the observation and the source matrix pencil are congruent and positive definite. Hence, their eigenvector matrices are related by

$$\mathbf{E}_\mathbf{S} = \mathbf{A}^H \mathbf{E}_\mathbf{X} \quad (3.31)$$

(cf. Sec. 1.3, Theo. 4) where  $\mathbf{E}_\mathbf{X}$  denotes the eigenvector matrix of the observation pencil  $(\mathbf{C}_\mathbf{X}, \mathbf{C}_{\mathbf{X}_F})$ . Because of the diagonality of  $\mathbf{E}_\mathbf{S}$  this means that once the GEVD of the observation pencil is computed the mixing matrix can be estimated immediately as

$$\mathbf{A}^{est} = (\mathbf{E}_\mathbf{X}^{-1})^H. \quad (3.32)$$

Note, that the scaling and permutation indeterminacies with which the matrix  $\mathbf{E}_\mathbf{X}$  is determined in the GEVD of the observation pencil equal the indeterminacies inherent in the BSS model.

Once the matrix  $\mathbf{A}^{est}$  is estimated the sources can be determined directly as

$$\mathbf{S}^{est} = (\mathbf{A}^{est})^{-1}\mathbf{X}. \quad (3.33)$$

Hence, the MP-BSS algorithm can be summarized as follows (cf. Alg. 2): first the filtered versions of the observations have to be generated. For both the original and the filtered observations the corresponding covariance matrices are computed and used to constitute the observation pencil  $(\mathbf{C}_\mathbf{X}, \mathbf{C}_{\mathbf{X}_F})$ . For this matrix pencil, the eigenvector matrix  $\mathbf{E}_\mathbf{X}$  is determined whereupon the mixing and the source matrix can be estimated according to Eq. (3.32) and Eq. (3.33), respectively.

	$k = 1$	$k = 2$
$s_{0,k}$ [ $l.u.$ ]	5	2
$T_{2,k}$ [ $t.u.$ ]	500	100
$\omega_k$ [ $1/t.u.$ ]	0	variable
$\Phi_k$	$\pi/4$	0

Table 3.1: Parameters of the sources Eq. (3.34) used in the simulation ( $t.u.$ ,  $l.u.$  arbitrary time and length units, respectively).

## 3.4 Limits of the MP-BSS Algorithm

### 3.4.1 Violations of Uncorrelatedness Assumptions

In order to solve the BSS problem uniquely the MP-BSS algorithm requires that the observation matrix pencil has non-degenerated eigenvalues and that the sources are uncorrelated. While the first condition can usually be met in practice by choosing a suitable filter the second poses a problem in the context of NMR if the recorded spectra contain overlapping peaks. The corresponding proton signals are then correlated and cannot be separated satisfactorily by MP-BSS. This means, that in contrast to the delineation in Sec. 3.2.2 estimated sources emerge that contain more than just one FID or peak, respectively. Accordingly, protein signals with a frequency very close to that of the water protons appear in the same estimated source as the water signal and hence get removed erroneously during the annulation step (step d) in Fig. 3.3) of the BSS procedure.

In order to investigate the robustness of the MP-BSS algorithm with respect to violations of the uncorrelatedness assumption of the sources the following experiment was carried out. First, two source signals of the form

$$s_k(t) = s_{0,k} \exp(-t/T_{2,k}) \exp(i(\omega_k t + \Phi_k)), \quad k = 1, 2 \quad (3.34)$$

were generated. The first source should resemble a typical water FID, i.e. it should have frequency zero, be dephased and should have a large amplitude and  $T_2$  time constant. Accordingly,  $\Phi_1 = \pi/4$ ,  $T_{2,1} = 500 t.u.$  (time unit) and  $s_{0,k} = 5 l.u.$  (length unit) were used for its generation while  $\omega_1$  was set to zero (cf. Tab. 3.1). In contrast, the second source imitated a FID originating from a single group of equivalent protein protons. As such signals are comparatively weak and in phase  $s_{0,2}$  was set to  $0.4s_{0,1}$  and  $\Phi_2 = 0$  was used. Furthermore, the value of  $T_{2,2}$  was set to  $100 t.u.$  in order to allow for the comparatively short relaxation times of protein protons. The frequency of this protein FID was varied throughout the experiment from -0.65 to 0.65

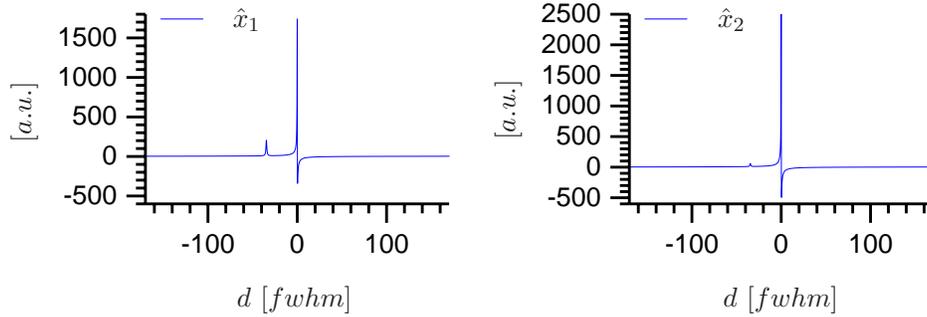


Figure 3.4: The rows  $\hat{x}_1$  and  $\hat{x}_2$  of the frequency domain mixture matrix  $\hat{\mathbf{X}}$ . The  $x$ -axis measures the distance  $d$  between the protein and the water peak in units of the full width at half maximum ( $fwhm$ ) of the protein peak. The dominant water peak resides at  $d = 0 fwhm$ . The significantly smaller protein peak (here shown at  $d = -34.5 fwhm$ ) was shifted from  $d = -34.5 fwhm$  to  $d = 34.5 fwhm$  during the simulations.

[1/ $t.u.$ ] in steps of 0.05, or, in other words, the distance between the protein and the water peak ranged between  $-34.5 fwhm$  and  $34.5 fwhm$  where  $fwhm$  denotes the full width at half maximum of the protein peak.

The two source signals were used to constitute the rows of the source matrix  $\mathbf{S}$  and were mixed by the mixing matrix

$$\mathbf{A} = \begin{bmatrix} 1.00 & 1.00 \\ 1.45 & 0.25 \end{bmatrix} \quad (3.35)$$

The resulting mixtures  $\mathbf{X} = \mathbf{A}\mathbf{S}$  (see Fig. 3.4) were fed into the time domain MP-BSS algorithm. Furthermore, the mixtures were also Fourier-transformed and then provided to the frequency domain version of the MP-BSS algorithm for comparison. For both approaches filtered signals were needed in order to compute the second covariance matrix  $\mathbf{C}_{\hat{\mathbf{X}}_F}$  of the matrix pencil. This filtering was carried out by means of a filter with Gaussian shaped frequency response centered at the resonance frequency of the water protons.

To quantify how well the MP-BSS algorithm could recover the water and the protein signal the minimal correlation coefficient

$$\rho = \min(r(s_1^{org}, s_1^{est}), r(s_2^{org}, s_2^{est})) \quad (3.36)$$

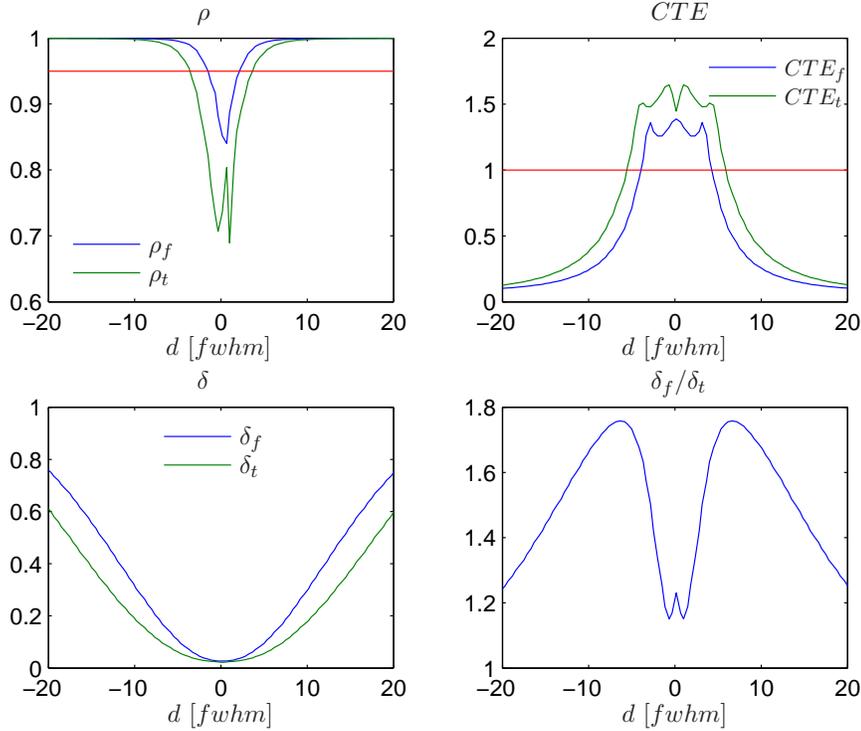


Figure 3.5: Robustness of MP-BSS. Top: the minimal correlation coefficient  $\rho$  and the  $CTE$  for both time (subscript  $t$ ) and frequency domain (subscript  $f$ ) MP-BSS. Usually, a BSS is considered successful if  $\rho > 0.95$  and  $CTE < 1$  (red lines). Especially for small distances  $d$  between the protein and the water peak better results are obtained by the frequency domain MP-BSS algorithm. Bottom: dependency of the eigenvalue distance measure  $\delta$  on the peak distance  $d$ . The left plot shows the course of  $\delta$  for the time ( $\delta_t$ ) and the frequency ( $\delta_f$ ) observation matrix pencil separately. The closer the frequencies of the water and the protein peak the smaller the difference between the eigenvalues. Right: the ratio  $\delta_f/\delta_t$  quantitatively compares the difference of the eigenvalues of the time and the frequency domain matrix pencil.

between the original and the estimated sources  $s_k^{est}$ ,  $k = 1, 2$ , was computed. Additionally, the *CTE* between the estimated and the recovered mixing matrix was determined. The results are depicted in Fig. 3.5. As can be seen both the sources as well as the mixing matrix are better estimated by the frequency than by the time domain MP-BSS algorithm.

This phenomenon can be explained by investigating the eigenvalues of the observation matrix pencils  $(\mathbf{C}_{\mathbf{X}}, \mathbf{C}_{\mathbf{X}_F})$  and  $(\mathbf{C}_{\hat{\mathbf{X}}}, \mathbf{C}_{\hat{\mathbf{X}}_F})$ , respectively. As explained in Sec. 3.3.2 these eigenvalues have to be non-degenerated in order to ensure unique results of the MP-BSS algorithm. However, the more correlated the underlying sources are the closer the eigenvalues lie together no matter which filter is used in the MP-BSS algorithm. This relationship is depicted in Fig. 3.5 (bottom) where the distance measure

$$\delta = \frac{|\lambda_1 - \lambda_2|}{|\lambda_1| + |\lambda_2|} \quad (3.37)$$

was used to quantify the difference between the eigenvalues  $\lambda_1$  and  $\lambda_2$  of the observation matrix pencils. As can be seen the difference between the eigenvalues of the time domain matrix pencil decreases faster than that of the frequency domain pencil for increasing correlation of the sources. Hence, the robust results of the frequency domain MP-BSS approach.

In particular, if frequency domain MP-BSS was applied the sources could be well-separated even if the protein and the water peak overlapped significantly (see Fig. 3.6). Only if the distance between the two peaks lay in the range  $-1.2 \text{ fwhm} < d < 1.5 \text{ fwhm}$  the sources could not be recovered satisfactorily, i.e. their correlation coefficients with the corresponding original sources were smaller than 0.95.

The CTE between the estimated and the original mixing matrix was more sensitive to increasing correlations between the sources. Still, acceptable recoveries (i.e.  $CTE < 1$ ) were observed for peak distances smaller than  $-3.7 \text{ fwhm}$  and larger than  $4 \text{ fwhm}$ . Finally note, that perfect recoveries of the mixing matrix are not really needed for the removal of the water signal as long as the sources are estimated sufficiently well.

### 3.4.2 Violations of the Linear Mixture Model

Apart from violations of uncorrelatedness assumptions on which MP-BSS is based it must also be expected that data recorded in real life experiments does not strictly fit into the linear mixture model presented in Sec. 3.2.2. In particular, difficulties are likely to arise if corresponding peaks in spectra recorded for different  $t_1$  times distinguish themselves not only by their amplitude but also by their shape. In order to investigate how sensitively

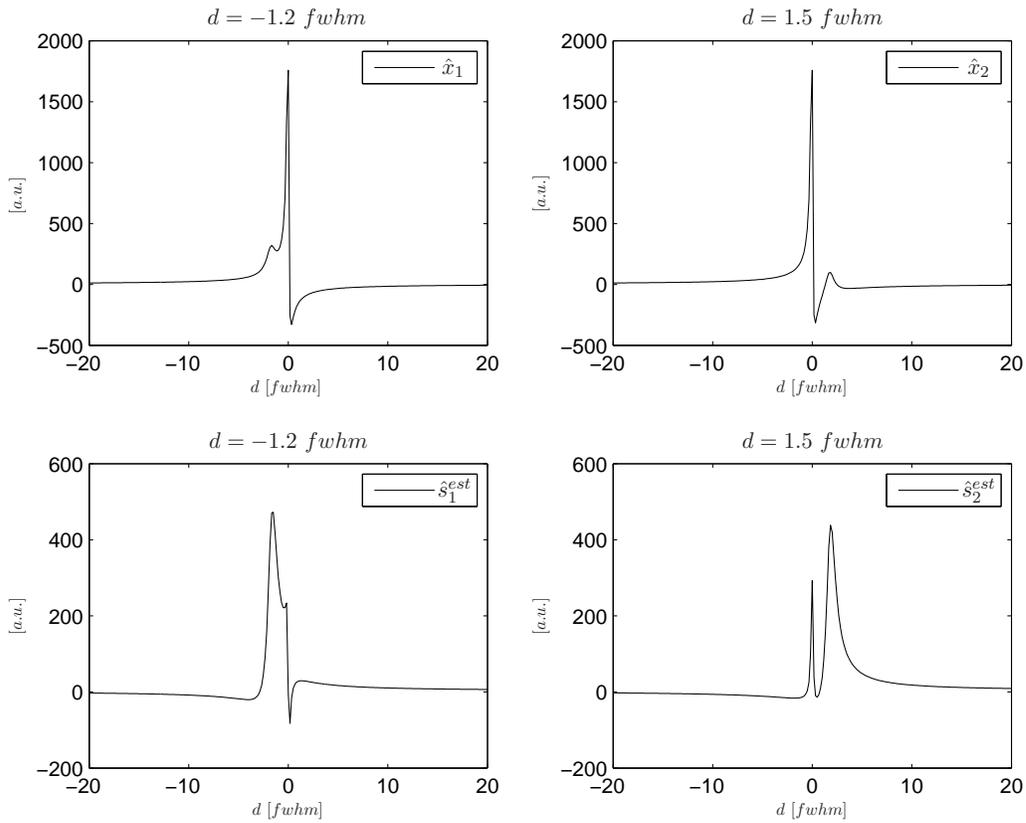


Figure 3.6: Top: for  $d = -1.2 \text{ fwhm}$  and  $d = 1.5 \text{ fwhm}$  the frequency domain MP-BSS can still recovery the sources satisfactorily ( $\rho \approx 0.95$ ). The first rows of the corresponding mixture matrices are depicted. Obviously, the water and the protein peaks overlap significantly. Bottom: the recovered protein peaks for  $d = -1.2 \text{ fwhm}$  and  $d = 1.5 \text{ fwhm}$ , respectively. Only small remnants of the water peak prevail and the baseline is almost flat.

	$s$	$\mathfrak{s}$
$\beta_1$	1	1
$\beta_2$	1	1
$\beta_3$	1	0.56
$\omega_1$ [1/t.u.]	2.5	2.5
$\omega_2$ [1/t.u.]	0	0
$\omega_3$ [1/t.u.]	0	0
$\phi_1$	0	0
$\phi_2$	$\pi/4$	$\pi/4$
$\phi_3$	$\pi/4$	$\pi/4$
$T_{2,1}$ [t.u.]	30	30
$T_{2,2}$ [t.u.]	50	50
$T_{2,3}$ [t.u.]	50	90

Table 3.2: The parameters used in Eq. (3.38) to generate the sources  $s$  and  $\mathfrak{s}$ . [t.u.] stands for arbitrary time units.

MP-BSS reacts to such deviations of the linear mixture model the following simulation was carried out.

First, three sources  $s_k$ ,  $k = 1, 2, 3$ , of the form

$$s_k(t) = \beta_k \exp(-t/T_{2,k}) \exp(i(\omega_k t + \phi_k)) \quad (3.38)$$

were generated whereas the parameters as listed in Tab. 3.2 were used. The sources  $s_2$  and  $s_3$  were identical and imitated a water FID in an NMR data set, i.e. their frequency  $\omega$  was zero and their phases  $\phi$  were nonvanishing. In contrast, the first signal  $s_1$  resembled a protein FID with zero phase, nonzero frequency and a  $T_2$ -time which was smaller than that of the water signals.

Additionally, the sources  $\mathfrak{s}$  were formed according to Eq. (3.38) (see Tab. 3.2 for parameters). The signals  $\mathfrak{s}_1$  and  $\mathfrak{s}_2$  equalled  $s_1$  and  $s_2$ , respectively, and mimicked again typical protein and water FIDs. The third source  $\mathfrak{s}_3$  was also similar to a water signal, however, its  $T_2$  was larger than that of  $\mathfrak{s}_2$ . Furthermore,  $\mathfrak{s}_3$  was scaled down by  $\beta_3 = 0.56$  such that the Fourier transforms  $\hat{\mathfrak{s}}_3$  and  $\hat{\mathfrak{s}}_2$  of  $\mathfrak{s}_2$  and  $\mathfrak{s}_3$ , respectively, had identical amplitudes. As can be seen in the bottom right Fig. 3.7  $\hat{\mathfrak{s}}_2$  and  $\hat{\mathfrak{s}}_3$  differ only marginally from each other.

Fourier transforms  $\hat{s}_k$  and  $\hat{\mathfrak{s}}_k$  of all sources  $s_k$  and  $\mathfrak{s}_k$ ,  $k = 1, 2, 3$ , were computed and used to constitute the rows of the source matrices  $\hat{\mathbf{S}}$  and  $\hat{\mathbf{\mathfrak{S}}}$ , respectively (see Fig. 3.7). These matrices were multiplied by the mixing

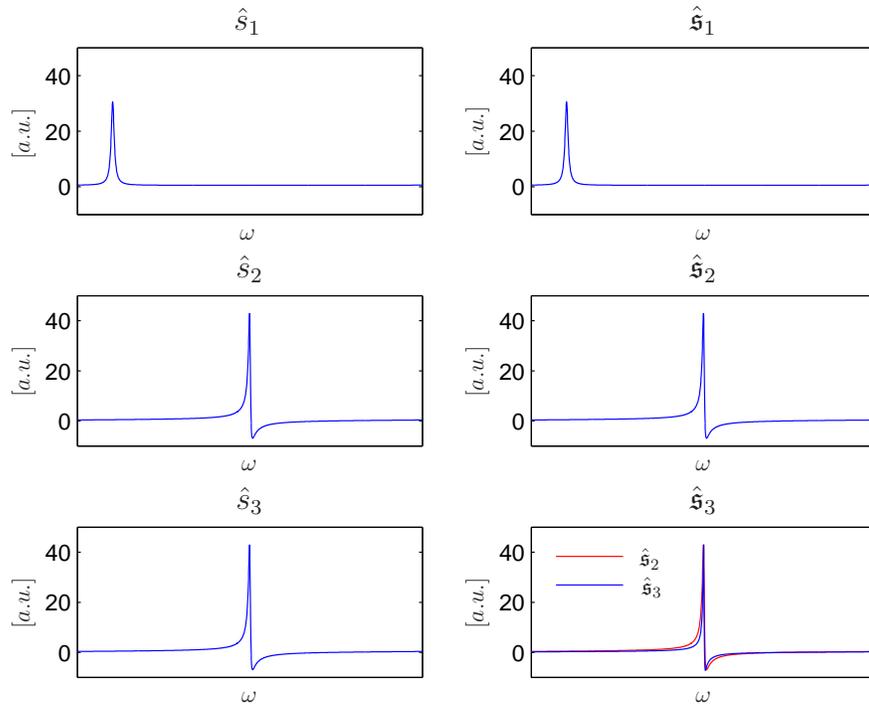


Figure 3.7: Fourier transformed sources. Left: the sources  $\hat{s}$ . Right: the sources  $\hat{\mathfrak{s}}$ . Note that source  $\hat{\mathfrak{s}}_2$  and  $\hat{\mathfrak{s}}_3$  differed only marginally from each other.

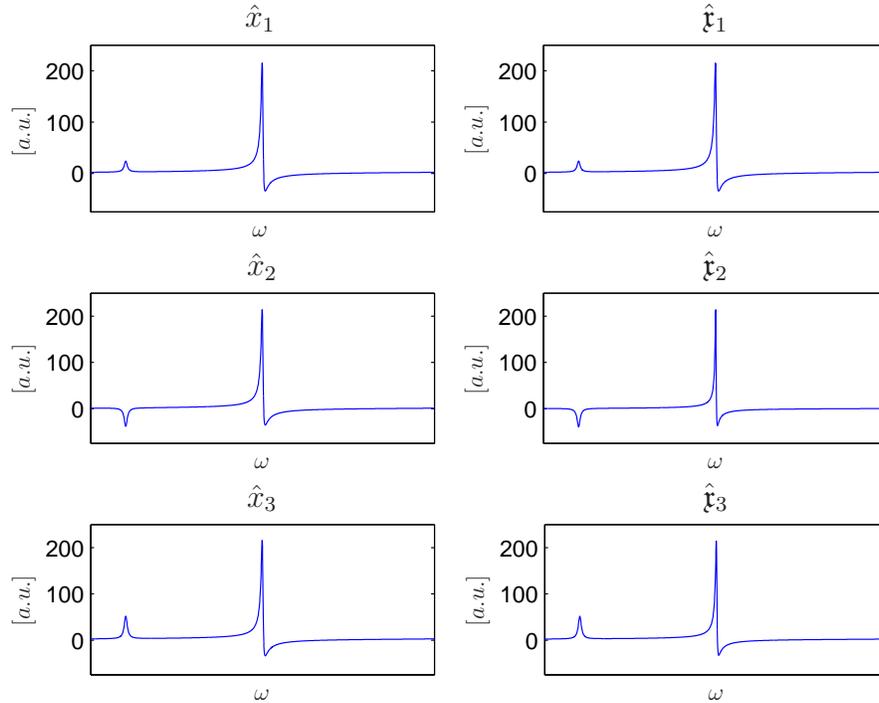


Figure 3.8: The mixtures  $\hat{\mathbf{X}}$  (left) and  $\hat{\mathbf{x}}$  (right). While in all observations  $\hat{x}_k$  the same water peak appeared its shape changed slightly in the mixtures  $\hat{\mathbf{x}}_k$ .

matrix

$$\mathbf{A} = \begin{bmatrix} 0.7143 & 5 & 0 \\ -1.3362 & 0 & 5 \\ 1.6236 & 2.5 & 2.5 \end{bmatrix} \quad (3.39)$$

such that observation matrices

$$\hat{\mathbf{X}} = \mathbf{A}\hat{\mathbf{S}} \quad \text{and} \quad \hat{\mathbf{x}} = \mathbf{A}\hat{\mathbf{G}} \quad (3.40)$$

were obtained (see Fig. 3.8).

The particular structure of the last two columns of  $\mathbf{A}$  lead to an identical water signal in all observations  $\hat{x}_k$ . In contrast, the shape of the water signal differed slightly in the  $\hat{\mathbf{x}}_k$ 's as a consequence of the variations between  $\hat{\mathbf{s}}_2$  and  $\hat{\mathbf{s}}_3$ . In detail,  $\hat{\mathbf{x}}_1$  contained a scalar multiple of the water signal  $\hat{\mathbf{s}}_2$ ,  $\hat{\mathbf{x}}_2$  a scalar multiple of  $\hat{\mathbf{s}}_3$  and the water signal found  $\hat{\mathbf{x}}_3$  was a mixture of both  $\hat{\mathbf{s}}_2$  and  $\hat{\mathbf{s}}_3$ . In other words, the water signals in the individual observations  $\hat{\mathbf{x}}_k$  did not differ in their amplitudes but marginally in their shapes.

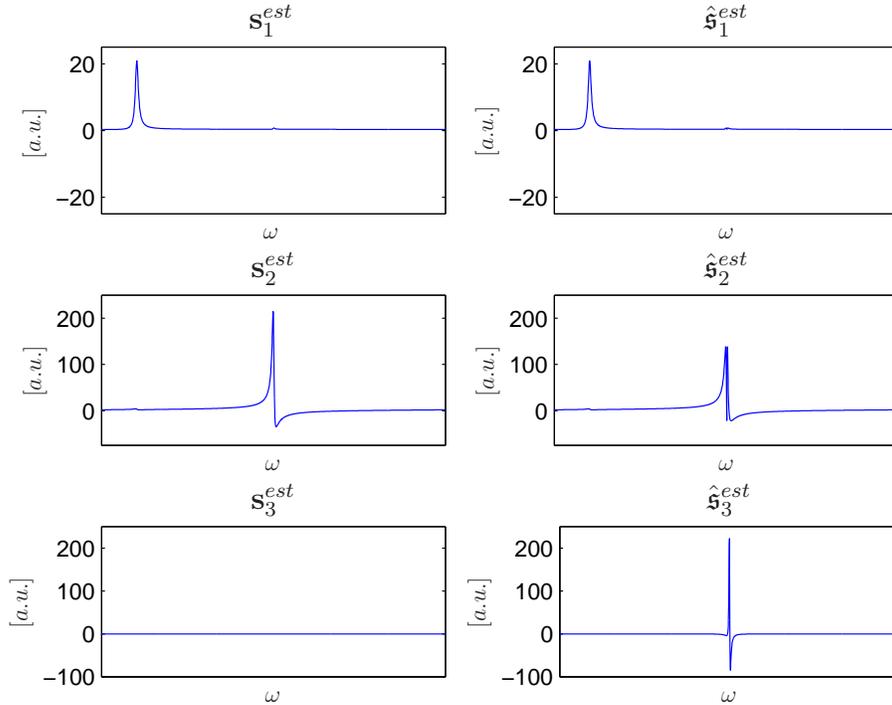


Figure 3.9: Estimated sources. Left: estimated sources obtained when  $\hat{\mathbf{X}}$  was analyzed. The water signal appears in a single source. Right: estimated sources determined when MP-BSS was applied to  $\hat{\tilde{\mathbf{X}}}$ . The water signal is split up and appears in  $\hat{\mathbf{s}}_2^{est}$  and  $\hat{\mathbf{s}}_3^{est}$  simultaneously.

Both observations  $\hat{\mathbf{X}}$  and  $\hat{\tilde{\mathbf{X}}}$  were fed into the MP-BSS algorithm whereas a Gaussian shaped filter of width  $\sigma = 100$  [1/t.u.] was used to filter the observations. The resulting estimated sources  $\hat{\mathbf{s}}_k^{est}$  and  $\hat{\tilde{\mathbf{s}}}_k^{est}$ ,  $k = 1, 2, 3$ , are depicted in Fig. 3.9. As expected, MP-BSS lead to a pure water ( $\hat{\mathbf{s}}_2^{est}$ ) and a clear protein signal ( $\hat{\mathbf{s}}_1^{est}$ ) when it was applied to  $\hat{\mathbf{X}}$ . The third estimated source  $\hat{\mathbf{s}}_3^{est}$  was zero which is not surprising as  $\hat{\mathbf{X}}$  was the result of mixing two *different* sources only.

In contrast, three nonzero estimated sources emerged when MP-BSS was applied to  $\hat{\tilde{\mathbf{X}}}$ . As before, the protein signal only appeared in the first estimated source  $\hat{\tilde{\mathbf{s}}}_1^{est}$  and was well separated from the water signal. The latter, however, got split up and appeared in both estimated sources  $\hat{\tilde{\mathbf{s}}}_2^{est}$  and  $\hat{\tilde{\mathbf{s}}}_3^{est}$  simultaneously.

Note that the present BSS problem  $\hat{\mathbf{X}} = \mathbf{A}\hat{\mathbf{S}}$  is undercomplete, i.e. more observations than sources are available. This case is rather rare and usually

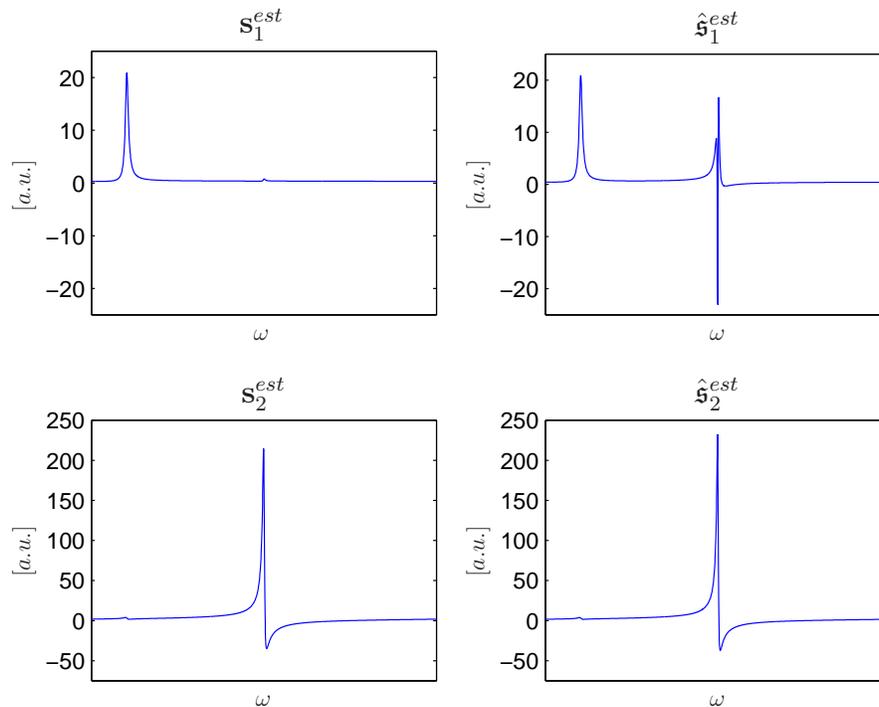


Figure 3.10: Estimated sources for the  $2 \times 2$  BSS problem. Left: applying MP-BSS to the first two rows of  $\hat{\mathbf{X}}$  still leads to a perfect separation of the water and the protein signal. In contrast, remnants of the water signal remain in source  $\hat{s}_1^{est}$  (right) when the first two rows of  $\hat{\mathbf{X}}$  are provided to MP-BSS.

the complete case, in which the number of sources equals the number of observations is assumed. Hence, the above simulation was repeated whereas only the first two rows of  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{X}}$ , respectively, were provided to the MP-BSS algorithm. As can be seen in Fig. 3.10 MP-BSS still separated the water and the protein signal perfectly when  $\hat{\mathbf{X}}$  was analyzed. However, such a clear separation of the water and the protein signal could not be observed if  $\hat{\mathbf{X}}$  was decomposed by MP-BSS. In that case remnants of the water signal remained in the protein related source  $\hat{s}_1^{est}$  which were, however, of much smaller amplitude than the original water signal.

To conclude, these simulations show that MP-BSS is indeed sensitive to violations of the linear mixture model. In particular, it cannot be expected that the water signal will only appear in a single source when MP-BSS is applied to real world NMR data. Even if these simulations indicate that a perfect separation of the water signal from protein signals is unlikely in

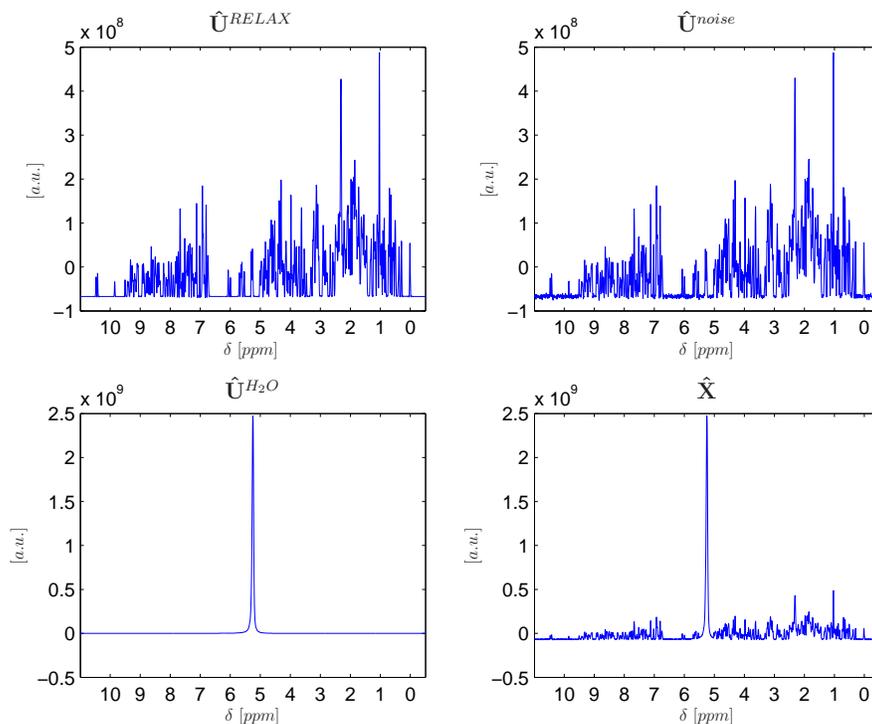


Figure 3.11: Toy data used to probe MP-BSS. Top left: the pure spectrum  $\hat{\mathbf{U}}^{RELAX}$  of TmCSP as determined by RELAX. Top right: the TmCSP spectrum after adding Gaussian noise ( $\hat{\mathbf{U}}^{noise}$ ). Bottom left: the recorded water spectrum ( $\hat{\mathbf{U}}^{H_2O}$ ). Bottom right: the noisy TmCSP spectrum plus the water spectrum ( $\hat{\mathbf{X}}$ ). All figures show the spectra corresponding to the smallest  $t_1$  time of the NOESY experiment.

the square BSS case good results can still be presumed if a large number of observations is available.

## 3.5 MP-BSS Applied to 2D-NOESY-Data

### 3.5.1 MP-BSS Applied to Artificial 2D-NOESY-Data

In order to probe its suitability to remove the water signal from NMR spectra MP-BSS was first applied to an artificially created 2D NOESY data set of the cold shock protein TmCSP of the bacterium *Thermotoga maritima*. This dataset was obtained with the aid of RELAX [22], an algorithm which is capable of generating 2D NOESY data based on the known spatial structure

of a protein. Altogether, 128 FIDs  $U^{RELAX}(t_1, t_2)$  were created whereas each consisted of 2048 data points. The  $t_2$ -Fourier transformed spectrum  $\hat{U}^{RELAX}(t_{1,min}, \omega_2)$  corresponding to the smallest  $t_1$ -time is shown in the top left of Fig. 3.11.

These spectra were used to constitute the rows of the  $128 \times 2048$ -matrix  $\hat{U}^{RELAX}$ . As real life NMR data are always contaminated by noise the matrix  $\mathbf{N}$  whose elements were drawn from the normal distribution was added to  $\hat{U}^{RELAX}$ :

$$\hat{U}^{noise} = \hat{U}^{RELAX} + \lambda \mathbf{N}. \quad (3.41)$$

Here, the scalar  $\lambda$  was chosen such that the SNR between the matrices  $\hat{U}^{noise}$  and  $\hat{U}^{RELAX}$  was 25 dB (see top right of Fig. 3.11). Such noise levels can also be expected in real life NMR datasets.

Additionally, 128 FIDs  $U^{H_2O}(t_1, t_2)$  (each consisting again of 2048 data points) of pure water were recorded in a 2D NOESY experiment. These water signals were  $t_2$ -Fourier transformed and collected in the  $128 \times 2048$  matrix  $\hat{U}^{H_2O}(t_1, \omega_2)$ . Adding this matrix to  $\hat{U}^{noise}$  lead to the actual  $128 \times 2048$  observation matrix

$$\hat{\mathbf{X}} = \hat{U}^{noise} + \hat{U}^{H_2O} \quad (3.42)$$

(see bottom of Fig. 3.11).

From the spectra in  $\hat{\mathbf{X}}$  the water signal was to be removed again by means of MP-BSS. Therefore, the matrix pencil  $(\mathbf{C}_{\hat{\mathbf{X}}}, \mathbf{C}_{\hat{\mathbf{X}}_F})$  had to be formed whereas for the first matrix  $\mathbf{C}_{\hat{\mathbf{X}}}$  the covariance matrix of the observations  $\hat{\mathbf{X}}$  was used as usual. For the second matrix  $\mathbf{C}_{\hat{\mathbf{X}}_F}$  the covariance matrix of the theoretical protein spectra with added noise (i.e. the covariance matrix of  $\hat{U}^{noise}$ ) was computed.

Analysis of  $\hat{\mathbf{X}}$  by MP-BSS lead to 128 estimated sources which constituted the rows of the  $128 \times 2048$  matrix  $\mathbf{S}^{est}$ . Furthermore, a  $128 \times 128$  mixing matrix  $\mathbf{A}^{est}$  was estimated. Following the procedure described in Sec. 3.2.2 the contributions of the sources in  $\mathbf{S}^{est}$  to the first observation (i.e. to the first row of  $\hat{\mathbf{X}}$ ) were examined. In this process the three sources shown in Fig. 3.12 were assigned to the water signal and were set to zero in  $\mathbf{S}^{est}$  such that the matrix  $\hat{\mathbf{S}}^{proton}$  was obtained.

Finally, the water free spectra  $\hat{\mathbf{X}}^{protein}$  were retrieved by left multiplying  $\hat{\mathbf{S}}^{proton}$  by  $\mathbf{A}^{est}$ :

$$\hat{\mathbf{X}}^{protein} = \mathbf{A}^{est} \hat{\mathbf{S}}^{proton} \quad (3.43)$$

As can be seen in Figs. 3.13 and 3.14 MP-BSS removed the water signal almost perfectly and only affected the peaks in the center (at about 5.2 ppm) of the spectrum. Given the simulation results of Sec. 3.4 distortions of the central peaks had to be expected and are the only drawback of the MP-BSS

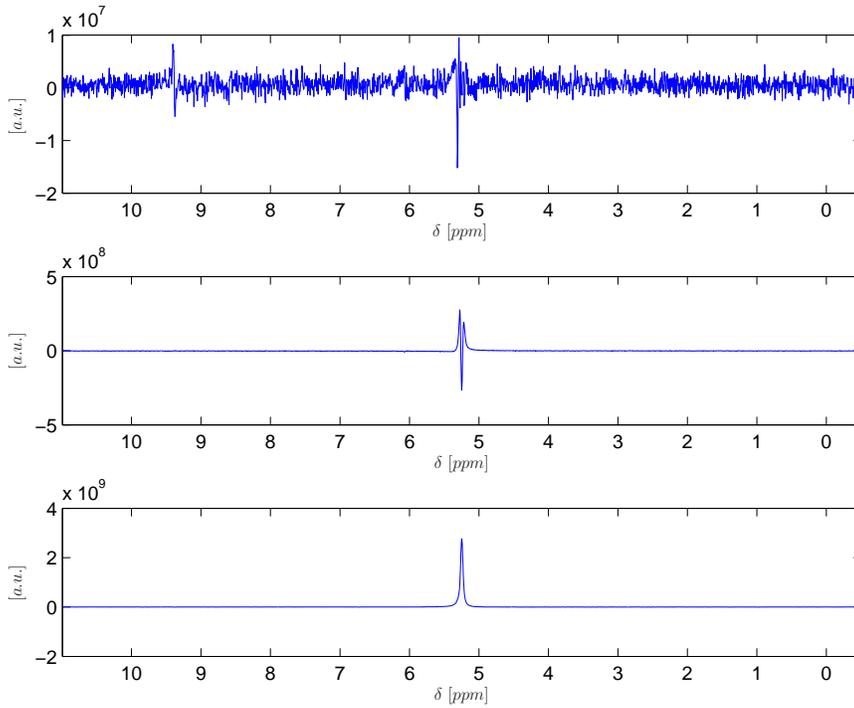


Figure 3.12: The contributions of the three sources related with the water signal to the first observation. The corresponding rows of  $\mathbf{S}^{est}$  were set to zero in order to remove the water signal from the spectra.

procedure. Note, that the baseline around the water signal is almost flat in the recovered spectrum. These good results are also reflected by the high SNR of 20 dB between the original spectra  $\hat{\mathbf{U}}^{RELAX}(t_1, \omega_2)$  and the recovered protein spectra  $\hat{\mathbf{X}}^{protein}$ .

### 3.5.2 MP-BSS Applied to Real World Dataset

Apart from the artificial TmCSP spectra discussed in the last section MP-BSS was also applied to remove the water signal from a real world 2D NOESY dataset which was recorded during an investigation of the protein P11. The latter corresponds to the helix 11 of the human glutathione reductase and consists of 24 amino acids.

In the experiment 512 single FIDs  $U(t_2, t_2)$ , each consisting of 2048 data points, were recorded and Fourier transformed with respect to both time domains. The resulting 2D spectrum  $\hat{U}(\omega_1, \omega_2)$  was phase corrected before it

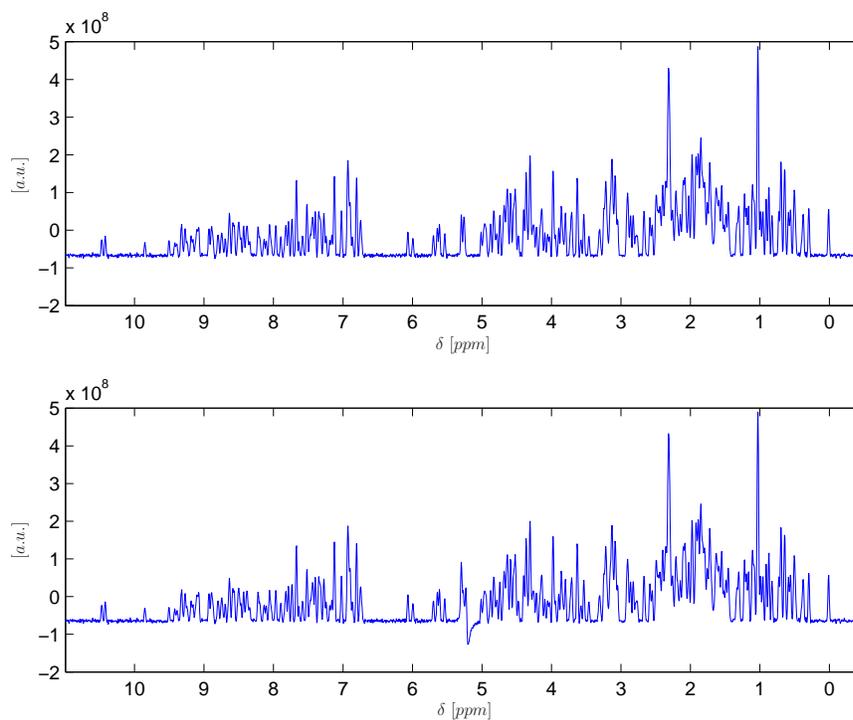


Figure 3.13: Comparison between the original and the recovered TmCSP spectrum. Top: the noisy RELAX spectrum  $\hat{\mathbf{U}}^{noise}$  for the shortest evolution time  $t_1$ . Bottom: the spectrum as recovered by MP-BSS (first row of  $\hat{\mathbf{X}}^{protein}$ ). The water signal got almost perfectly removed. Only the peaks in the middle of the spectrum were affected by the MP-BSS water removal procedure (see also Fig. 3.14).

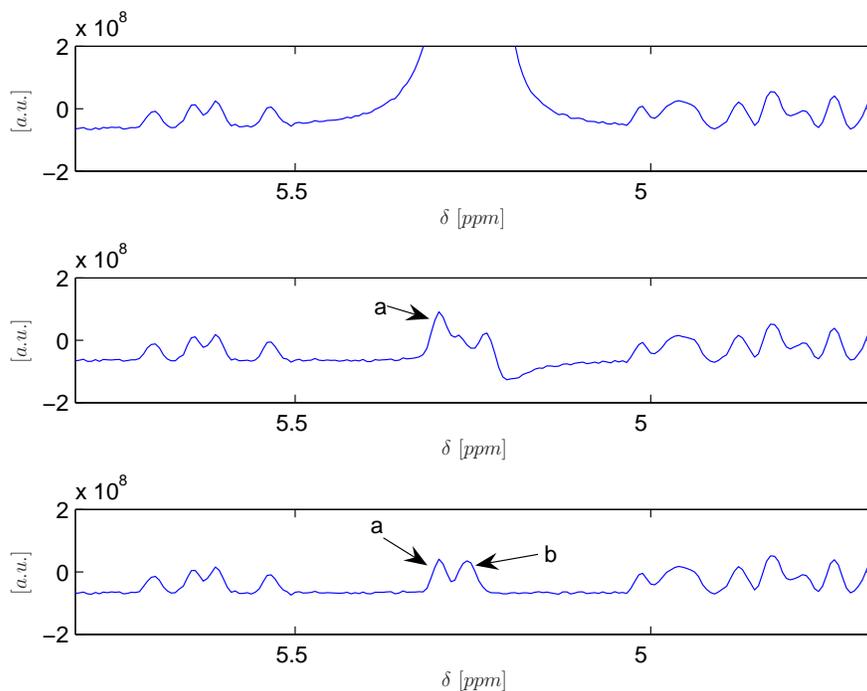


Figure 3.14: Close-up of the section of the TmCSP spectrum which contains the water signal. Top: the noisy TmCSP spectrum plus water (first row of  $\hat{\mathbf{X}}$ ). Middle: the spectrum  $\hat{\mathbf{X}}^{protein}$  as recovered by MP-BSS. Bottom: the original noisy spectrum  $\hat{\mathbf{U}}^{noise}$ . While peak *a* is still clearly visible in the recovered spectrum peak *b* vanishes. Additionally some minor remnants of the water signal appear in the recovered spectrum.

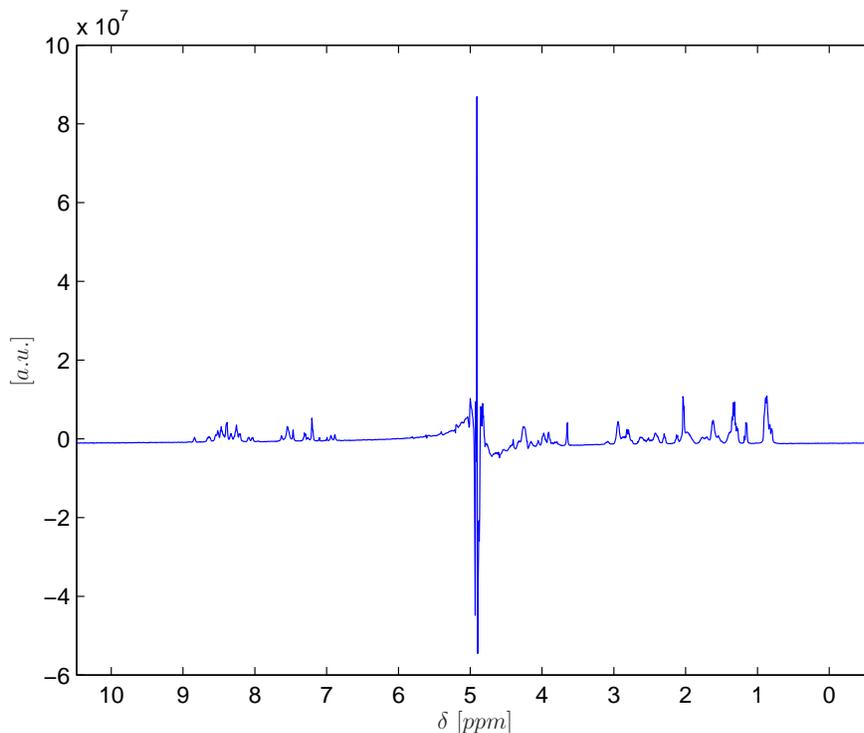


Figure 3.15: The spectrum of P11 for the smallest  $t_1$ -time. Clearly, the dephased water signal at about 5 ppm dominates the spectrum and leads to baseline distortions.

was transformed back to the  $(t_1, \omega_2)$  domain by an inverse  $t_1$ -Fourier transform. From the resulting spectra  $\hat{U}(t_1, \omega_2)$  only every fourth could be used to constitute the  $128 \times 2048$  observation matrix  $\hat{\mathbf{X}}$  because of a specific phase cycling scheme applied during the recording of the FIDs. Fig. 3.15 depicts the first spectrum of  $\hat{\mathbf{X}}(t_1, \omega_2)$  which was recorded for the shortest evolution time  $t_1$ . Obviously, the spectrum is corrupted by a predominant water peak in the middle which is dephased because of a presaturation technique used to suppress the water signal during the recording of the individual FIDs. As can be seen, this water signal leads to considerable baseline distortions in the spectrum.

In order to remove this water signal by means of MP-BSS the matrix pencil  $(\mathbf{C}_{\hat{\mathbf{X}}}, \mathbf{C}_{\hat{\mathbf{X}}_F})$  had to be formed. As in the case of the artificial TmCSP spectra the covariance matrix  $\mathbf{C}_{\hat{\mathbf{X}}}$  of the  $128 \times 2048$  observation matrix  $\hat{\mathbf{X}}$  was used to constitute the first matrix of the pencil.

As no water free spectra were available, the observations had to be filtered

explicitly before the second matrix of the pencil could be formed. For this purpose, a Gaussian shaped filter of width  $\sigma = 1$  was used which was centered at the resonance frequency of the water signal (i.e. in the middle of the spectrum):

$$\hat{\mathbf{X}}_F = \hat{\mathbf{X}} \diamond \hat{\mathbf{H}} \quad (3.44)$$

Here,  $\hat{\mathbf{X}}_F$  contains the filtered observations and  $\hat{\mathbf{H}}$  denotes the  $128 \times 2048$  matrix having the frequency response of the Gaussian filter in each of its rows. For the filtered observation matrix  $\hat{\mathbf{X}}_F$  the covariance matrix  $\mathbf{C}_{\hat{\mathbf{X}}_F}$  was computed and used as the second matrix in the pencil.

The GEVD of the matrix pencil  $(\mathbf{C}_{\hat{\mathbf{X}}}, \mathbf{C}_{\hat{\mathbf{X}}_F})$  lead to an estimated  $128 \times 128$ -mixing matrix  $\hat{\mathbf{A}}^{est}$  whose inverse was multiplied with the observation matrix  $\hat{\mathbf{X}}$

$$\hat{\mathbf{S}}^{est} = (\mathbf{A}^{est})^{-1} \hat{\mathbf{X}} \quad (3.45)$$

such that the  $128 \times 2048$  estimated source matrix  $\hat{\mathbf{S}}^{est}$  was obtained. The rows of the latter matrix contained 128 estimates sources  $\hat{s}_m^{est}$ ,  $m = 1, 2, \dots, 128$ , whose contributions

$$\hat{y}_m^{est} = a_{1m}^{est} \hat{s}_m^{est}, \quad m = 1, 2, \dots, 128 \quad (3.46)$$

to the first observation (i.e. the first row of  $\hat{\mathbf{X}}$ ) were analyzed in order to identify water related signals (cf. Sec. 3.2.2). In this procedure 70 sources were assigned to the water signal and set to zero in the matrix  $\hat{\mathbf{S}}^{est}$  such that the matrix  $\hat{\mathbf{S}}^{protein}$  was obtained. The water free spectra  $\hat{\mathbf{X}}^{protein}$  were eventually determined by mixing the remaining sources in  $\hat{\mathbf{S}}^{protein}$  by the estimated mixing matrix  $\mathbf{A}^{est}$ :

$$\hat{\mathbf{X}}^{protein} = \mathbf{A}^{est} \hat{\mathbf{S}}^{protein}. \quad (3.47)$$

Fig. 3.16 shows the water free spectra for the smallest evolution period  $t_1$ . As can be seen the water signal got almost perfectly removed while the protein signals remained in the spectrum. Further, as in the case of the artificial TmCSP spectra, also the baseline distortions caused by the water signal were reduced considerably.

Compared with the simulation carried out with the artificial TmCSP spectrum two differences appeared. On the one hand, the number of sources assigned to the water signal was much larger in the case of P11 than when the artificial TmCSP spectra were decomposed by MP-BSS. As already explained in Sec. 3.4.2 such a splitting up of the water signal occurs if corresponding peaks in spectra recorded for different  $t_1$ -times differ in their shape. Such differences appeared in the P11 spectra as can be seen in Fig. 3.17. In its top two close-ups of the spectrum recorded for the shortest evolution time  $t_1$

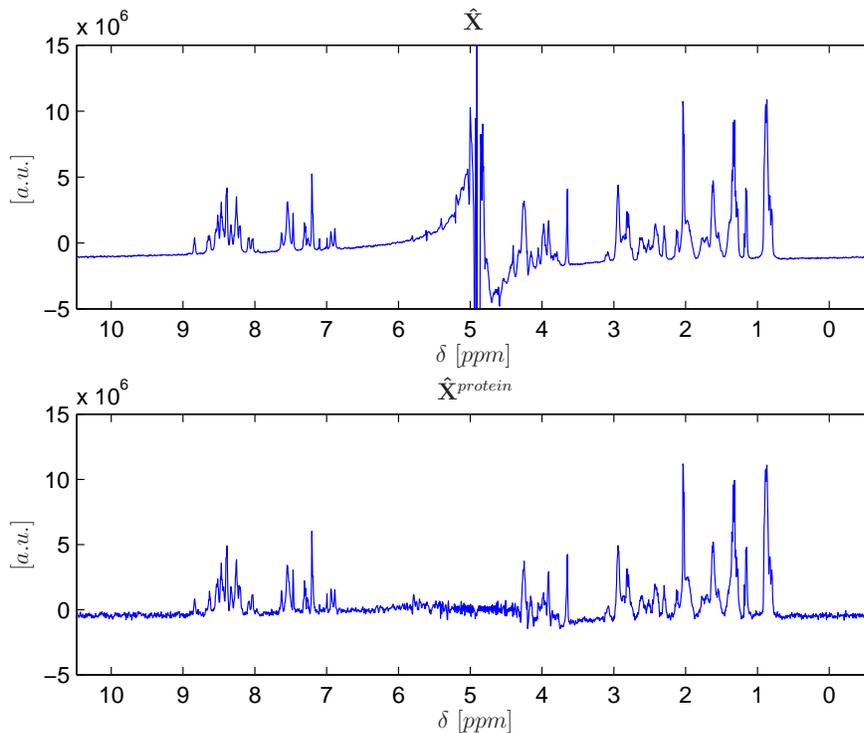


Figure 3.16: Comparison between the original (top) and the water-free P11 spectrum (bottom). Obviously, the water signal got almost perfectly removed.

are shown which depict the protein peak at 7.55 ppm and the water signal, respectively. In the bottom of Fig. 3.17 the same peak and the water signal are shown as they appeared in the spectrum recorded for the longest  $t_1$ -time. The latter signals were multiplied by a scalar factor such that their amplitudes equalled those of the corresponding signals recorded for the shortest  $t_1$ -time. Although the corresponding signals are similar for both  $t_1$ -times the differences in their shapes are highly visible and are responsible for the distribution of the water signal among several sources.

The second difference between the results obtained for the artificial Tm-CSP and the real life P11 data is that additional noise appeared in the water free P11 spectra. This phenomenon can be explained by investigating the individual sources (i.e. the rows of  $\hat{\mathbf{S}}^{est}$ ) which resulted from the MP-BSS analysis of the P11 spectra. The contributions of four such sources to the spectrum recorded for the shortest  $t_1$ -time are depicted in Fig. 3.18. As can be seen each of these sources is corrupted by significant levels of noise. If

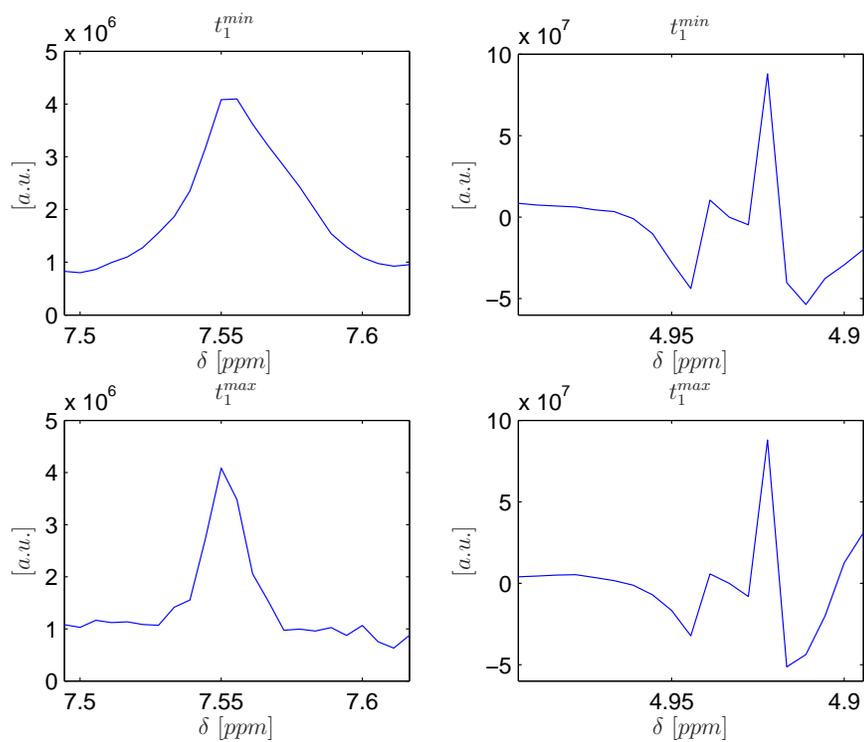


Figure 3.17: The shapes of the protein peak at 7.55 ppm (left) and the water signal (right) for different  $t_1$ -times. Top: the signals as they appeared in the spectrum recorded for the shortest  $t_1$ -time. Bottom: the same signals as recorded for the longest  $t_1$ -time. These signals were multiplied by a scalar factor such that they had identical amplitude as the corresponding signals above.

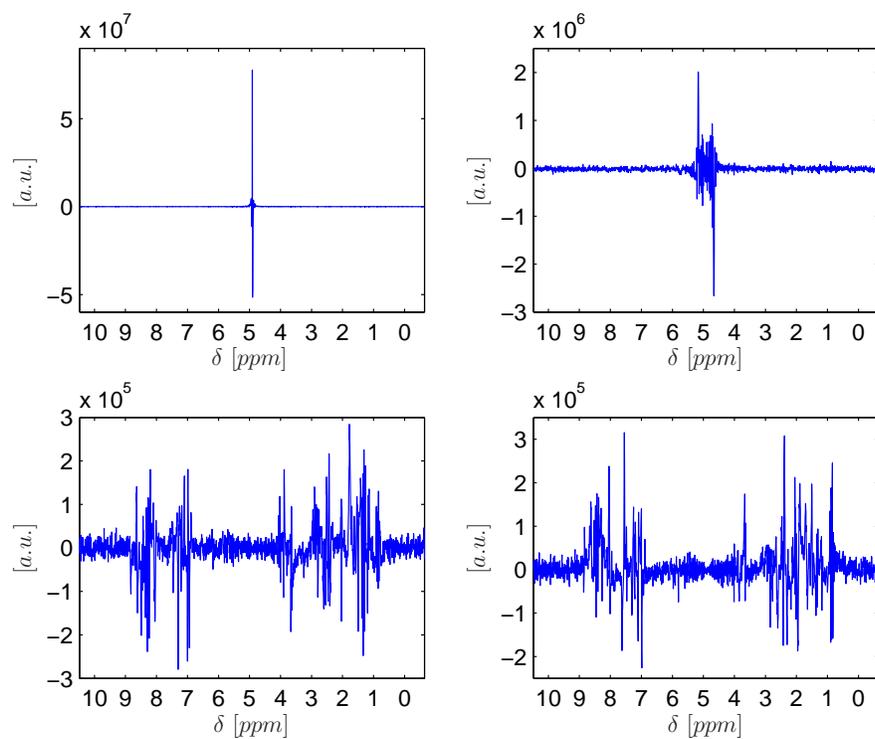


Figure 3.18: The contributions of four of the 128 estimated sources to the P11 spectrum recorded for the shortest  $t_1$ -time. Top: two typical water sources. Bottom: two typical protein sources. Note that all sources are corrupted by significant levels of noise.

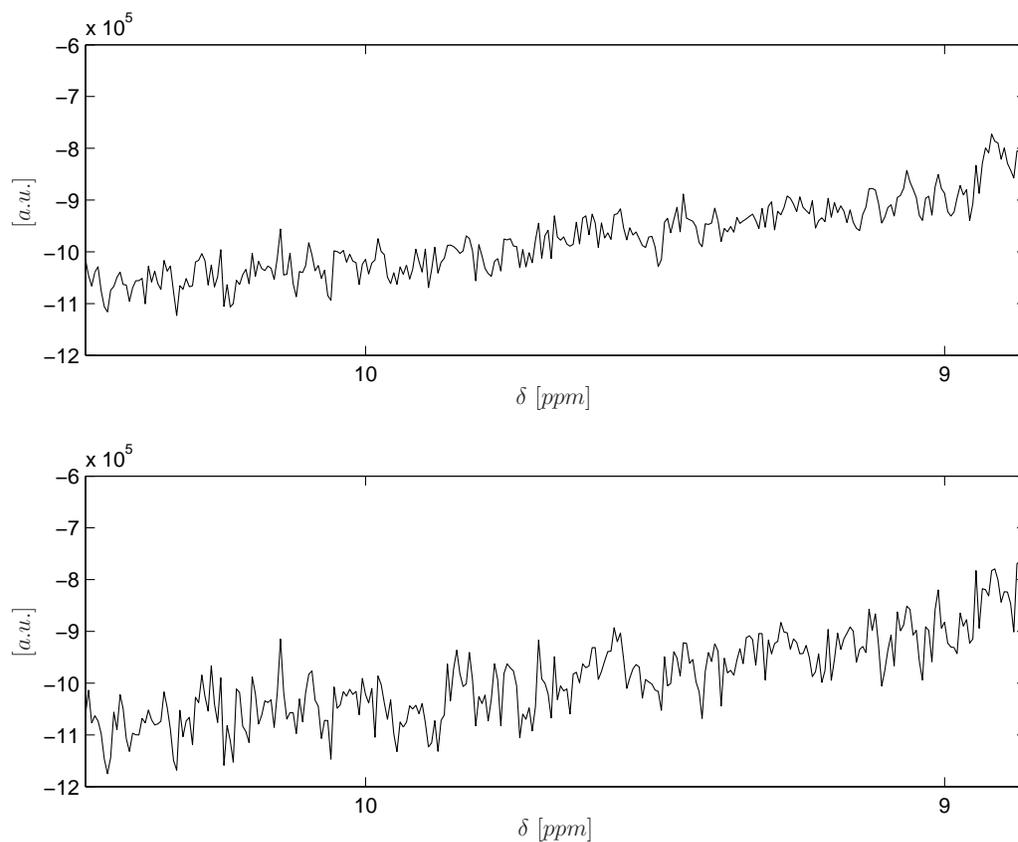


Figure 3.19: Increase of noise if water related signals are set to zero. Top: close-up of the original spectrum depicted in the top of Fig. 3.16. Bottom: close-up of the same region of a spectrum for which a single water related source was set to zero.

all sources are used to reconstitute the protein signal (i.e. if  $\mathbf{S}^{protein} = \mathbf{S}^{est}$  in Eq. (3.47)) these noise contributions cancel each other out to a large extent. However, if some of these sources are set to zero in order to remove the water signal from the spectra then this mutual compensation becomes less effective such that the resulting spectra are noisier than the original ones. This effect is illustrated in Fig. 3.19. In its top a close-up of the section [8.7 ppm 10.4 ppm] of the original spectrum as displayed in the top Fig. 3.16 is shown. In the bottom the same region is shown of a spectrum which was obtained by setting a single water related source (the one depicted in the upper left of Fig. 3.18) in  $\hat{\mathbf{S}}^{est}$  to zero before the matrix  $\hat{\mathbf{X}}^{protein}$  was computed. Obviously, the signal in the bottom of Fig. 3.19 is corrupted by a higher level of noise than that one in the top.

## 3.6 Conclusions

This chapter started with detailed description of the water signal and its impacts on protein spectra. It was explained how the data recorded during 2D-NOESY experiments fit into the linear mixture model and how the water signal can be detected and removed once the source and the mixing matrix have been estimated. Furthermore, MP-BSS was presented as an elegant tool to solve the BSS problem and its limitations with respect to 2D-NOESY data were elucidated in various simulations. To probe its suitability to analyze 2D-NOESY data MP-BSS was first applied to an artificially created TmCSP data set. Finally, MP-BSS was also applied to remove the water signal from real world 2D-NOESY data of the protein P11. While the water signal was removed almost perfectly, additional noise appeared in the spectra. Moreover, a large number of water sources were estimated which can be hard to detect manually. The latter two problems will be addressed in the next chapter.

# Chapter 4

## Denoising of MP-BSS data

As shown in the last chapter MP-BSS is powerful tool to remove the water signal from 2D-NOESY data sets. However, the resulting water free spectra suffer from increased noise levels which exacerbate the further analysis. This additional noise is difficult to remove as, for instance, classical denoising methods which average over small regions of the spectra change the shape and the amplitude of small protein peaks unacceptably. For this reason variants of Principal Component Analysis (PCA), called local PCA and Kernel PCA, are investigated in this section with which the spectra can be denoised while the protein peaks remain intact. Both methods map the observed data nonlinearly into feature spaces before classical PCA denoising is carried out. As will be seen this projection step leads to considerably better results than in classical PCA in which the data is analyzed in input space.

### 4.1 Principal Component Analysis

For convenience, the principles of PCA based denoising will be reviewed in this section as they also play a major role in local and Kernel PCA.

#### 4.1.1 Concept of PCA Denoising

PCA is a classic technique in statistical data analysis, feature extraction, and data compression, stemming from the early work of Pearson [43]. The starting point for PCA is an  $M$ -dimensional random vector  $\mathbf{x}$  of which  $T$  samples  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{T-1}$  are available [32]. The task is to find a lower dimensional (e.g.  $U < M$  dimensional) representation  $\mathbf{y}$  of  $\mathbf{x}$  which preserves as much of the variance of  $\mathbf{x}$  as possible.

The idea is illustrated in Fig. 4.1. In the left picture the crosses represent

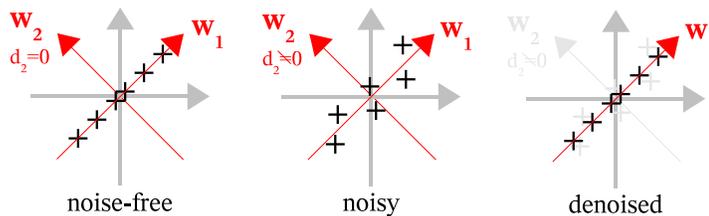


Figure 4.1: Denoising using linear PCA. Left: data stemming from a strictly linear process can be fully described by their projections onto  $\mathbf{w}_1$ . Middle: a noisy measurement of the data leads to nonvanishing projections onto  $\mathbf{w}_2$ . Right: the data are denoised by projecting them onto the eigenvector  $\mathbf{w}_1$  corresponding to the larger eigenvalue  $\lambda_1$  only.

the samples of the random vector  $\mathbf{x}$ . Using the standard coordinate system  $M = 2$  coordinates  $(x_1, x_2)$  are needed to describe their positions. However, a  $U = 1$  dimensional representation of the data points can be found in the rotated coordinate system as defined by the vectors  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . Then all the crosses lie on the  $\mathbf{w}_1$ -axis and it is sufficient to describe their positions by their projections onto  $\mathbf{w}_1$  only. The lower dimensional representation  $\mathbf{y}$  is then one-dimensional. The criterion for the determination of  $\mathbf{w}_1$  is that the variance of the data points projected onto it should be maximal.

If the crosses do not lie on a straight line as depicted in the middle of Fig. 4.1 neglecting their projections onto  $\mathbf{w}_2$  will result in a lossy compression. However, under all the compressions possible from a two-dimensional to a one-dimensional representation, projecting the data points onto  $\mathbf{w}_1$  is optimal in the sense that it preserves the highest amount of variance of the data.

Neglecting the projections onto  $\mathbf{w}_2$  may also lead to denoised data. Imagine the crosses represent a sample of a strictly linear process such that they should lie on a straight line as depicted in the left of Fig. 4.1. However, as measurements are usually corrupted by white Gaussian noise, the data points will deviate from that line, e.g. will be scattered around the  $\mathbf{w}_1$  direction as in the middle of Fig. 4.1. Particularly note that their nonvanishing projections onto  $\mathbf{w}_2$  are only caused by the presence of noise. Thus, denoising can be achieved by neglecting the projections onto  $\mathbf{w}_2$  and considering only projections onto  $\mathbf{w}_1$  (see right picture of Fig. 4.1).

More formally, the first step of the PCA transform is to center the vector  $\mathbf{x}$  by subtracting its mean  $\mathbf{m}_\mathbf{x}$

$$\mathbf{x} \leftarrow \mathbf{x} - \mathbf{m}_\mathbf{x}, \quad (4.1)$$

The first component  $y_1$  of the lower dimensional representation  $\mathbf{y}$  is then determined as a linear combination of the elements  $x_1, \dots, x_M$  of the vector  $\mathbf{x}$

$$y_1 = \sum_{k=1}^M \bar{w}_{k1} x_k = \mathbf{w}_1^H \mathbf{x}, \quad (4.2)$$

where the scalar coefficients  $w_{11}, \dots, w_{M1}$ , which are the elements of the vector  $\mathbf{w}_1$ , must be chosen such that the variance of  $y_1$  is maximal.  $y_1$  is then called the first principal component of  $\mathbf{x}$ . It must be noted, however, that the variance of  $y_1$  grows to the same degree as the norm of  $\mathbf{w}_1$  such that the additional constraint  $\|\mathbf{w}_1\| = 1$  must be imposed. Thus, a weight vector  $\mathbf{w}_1$  is sought which maximizes the following PCA criterion

$$J^{PCA}(\mathbf{w}_1) = E\{y_1^2\} = E\{(\mathbf{w}_1^H \mathbf{x})^2\} = \mathbf{w}_1^H E\{\mathbf{x}\mathbf{x}^H\} \mathbf{w}_1 = \mathbf{w}_1^H \mathbf{R}_x \mathbf{w}_1 \quad (4.3)$$

$$\text{under the constraint } \|\mathbf{w}_1\| = 1 \quad (4.4)$$

where  $\mathbf{R}_x$  denotes the  $M \times M$  correlation matrix of  $\mathbf{x}$ . Note that  $\mathbf{R}_x$  is Hermitian and positive semidefinite. It is well known from basic linear algebra that the solution to the PCA problem is given in terms of the unit-length eigenvectors  $\mathbf{e}_1, \dots, \mathbf{e}_M$ , of the matrix  $\mathbf{R}_x$ , whereas the eigenvectors are ordered such that the corresponding eigenvalues  $\lambda_1, \dots, \lambda_M$  satisfy  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$ . The solution maximizing (4.3) is given by

$$\mathbf{w}_1 = \mathbf{e}_1. \quad (4.5)$$

The PCA criterion  $J^{PCA}$  in Eq. (4.3) can be generalized to  $m$  principal components, with  $1 \leq m \leq M$ . Denoting by

$$y_m = \mathbf{w}_m^H \mathbf{x} \quad (4.6)$$

the  $m$ -th principal component and by  $\mathbf{w}_m$  the corresponding unit norm weight vector, the variance of  $y_m$  is now maximized under the constraint that  $y_m$  is uncorrelated with all the previously found principal components:

$$E\{y_m y_k\} = E\{(\mathbf{w}_m^H \mathbf{x})(\mathbf{w}_k^H \mathbf{x})\} = \mathbf{w}_m^H \mathbf{R}_x \mathbf{w}_k = 0, \quad k < m. \quad (4.7)$$

For the second principal component this leads to the constraint

$$\mathbf{w}_2^H \mathbf{R}_x \mathbf{w}_1 = d_1 \mathbf{w}_2^H \mathbf{e}_1 = 0, \quad (4.8)$$

as  $\mathbf{w}_1 = \mathbf{e}_1$  and  $\mathbf{e}_1$  is an eigenvector of  $\mathbf{R}_x$ . This means that the unit norm weight vector  $\mathbf{w}_2$ , which maximizes the variance  $E\{y_2^2\} = E\{(\mathbf{w}_2^H \mathbf{x})^2\}$ , is

**Data:**  $M \times T$  observation matrix  $\mathbf{X}$   
**Input:**  $U$  (number of dominant eigenvectors considered)  
**Result:** denoised observation matrix  $\mathbf{X}^{denoised}$

```

1 begin
2    $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{m}_X \mathbf{1}_T^\top$  where  $\mathbf{m}_X = \frac{1}{T} \mathbf{X} \mathbf{1}_T$ ;
3    $\mathbf{R}_X = \frac{1}{T-1} \mathbf{X} \mathbf{X}^H$ ;
4   EVD:  $\mathbf{R}_X \mathbf{E} = \mathbf{E} \mathbf{\Lambda}$ ;
6   Form  $\mathbf{E}^{red} = \{e_{ij}\}_{1 \leq i \leq M, 1 \leq j \leq U}$  from  $\mathbf{E} = \{e_{ij}\}_{1 \leq i, j \leq M}$ ;
7    $\mathbf{X}^{denoised} = \mathbf{E}^{red} ((\mathbf{E}^{red})^H \mathbf{X}) + \mathbf{m}_X \mathbf{1}_T^\top$ ;
8 end

```

**Algorithm 3:** Denoising based on PCA. The projection step (Eq. (4.11)), the coordinate transform (Eq. (4.13)) and the step in which the mean is added (Eq. (4.14)) are combined in the last step of the algorithm.

sought in the subspace orthogonal to the first eigenvector of  $\mathbf{R}_X$ . The solution is given by

$$\mathbf{w}_2 = \mathbf{e}_2. \quad (4.9)$$

Likewise, recursively it follows that

$$\mathbf{w}_k = \mathbf{e}_k \quad (4.10)$$

and the  $k$ -th principal component is given by

$$y_k = \mathbf{e}_k^H \mathbf{x}, \quad 1 \leq k \leq m. \quad (4.11)$$

Furthermore, from the result that the principal component basis vectors  $\mathbf{w}_i$  are eigenvectors  $\mathbf{e}_i$  of  $\mathbf{R}_X$  it follows that

$$E\{y_k^2\} = E\{\mathbf{e}_k^H \mathbf{x} \mathbf{x}^H \mathbf{e}_k\} = \mathbf{e}_k^H \mathbf{R}_X \mathbf{e}_k = \lambda_k. \quad (4.12)$$

Thus, the variances of the principal components are given by the eigenvalues of  $\mathbf{R}_X$ . In other words, the first  $U$  ( $U \in \{1, \dots, M\}$ ) principal components carry more variance than any other of the  $M - U$  alternative orthogonal directions. In data denoising it is usually assumed that the principal components corresponding to large eigenvalues represent the signal while principal components related with small eigenvalues carry noise. Denoising is thus achieved by projecting the data onto the  $U \leq M$  dominant eigenvectors (i.e. the eigenvectors  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_U$  related with the  $U$  largest eigenvalues) while projections onto the remaining ones are neglected.

The principal components  $y_k$  ( $k = 1, 2, \dots, U$ ), which form the vector  $\mathbf{y}$ , are related with the coordinate system spanned by the eigenvectors  $\mathbf{e}_k$  of the correlation matrix  $\mathbf{R}_{\mathbf{x}}$ . Hence,  $\mathbf{y}$  must be transformed back to the coordinate system in which the original vector  $\mathbf{x}$  was described. This is achieved by

$$\mathbf{y} \leftarrow \mathbf{E}^{red} \mathbf{y} \quad (4.13)$$

where  $\mathbf{E}^{red}$  denotes the  $M \times U$ -matrix having the dominant eigenvectors  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_U$ , in its columns. Eventually, the mean  $\mathbf{m}_{\mathbf{x}}$  must be added again to the denoised vectors in the last step of the PCA procedure:

$$\mathbf{y} \leftarrow \mathbf{y} + \mathbf{m}_{\mathbf{x}} \quad (4.14)$$

Finally, note that in real life experiments the expectation  $E\{.\}$  appearing in the above formulas must be estimated from the samples  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{T-1}$ , of the random vector  $\mathbf{x}$ . If these samples constitute the columns of the observation matrix  $\mathbf{X}$  the PCA algorithm as summarized in Alg. 3 is obtained.

#### 4.1.2 PCA Denoising of P11 Spectra

Following the procedure outlined in the last section PCA was used to denoise the water free P11 spectra obtained by MP-BSS (cf. Sec. 3.5.2). For this purpose, the mean

$$\mathbf{m}_{\hat{\mathbf{X}}^{protein}} = \frac{1}{2048} \hat{\mathbf{X}}^{protein} \mathbf{1}_{2048} \quad (4.15)$$

was subtracted from the water free observations stored in the  $128 \times 2048$ -matrix  $\hat{\mathbf{X}}^{protein}$  (cf. Eq (3.47)):

$$\hat{\mathbf{X}}^{protein} \leftarrow \hat{\mathbf{X}}^{protein} - \mathbf{m}_{\hat{\mathbf{X}}^{protein}} \mathbf{1}_{2048}^{\top} \quad (4.16)$$

For this centered data matrix the  $128 \times 128$ -correlation matrix

$$\mathbf{R}_{\hat{\mathbf{X}}^{protein}} = \frac{1}{2047} \hat{\mathbf{X}}^{protein} (\hat{\mathbf{X}}^{protein})^H \quad (4.17)$$

was computed and its eigenvalue decomposition

$$\mathbf{R}_{\hat{\mathbf{X}}^{protein}} \mathbf{E} = \mathbf{E} \mathbf{\Lambda} \quad (4.18)$$

was determined. Here,  $\mathbf{E}$  was an orthonormal matrix having the eigenvectors  $\mathbf{e}_m$ ,  $m = 1, 2, \dots, 128$ , in its columns and  $\mathbf{\Lambda}$  denoted a diagonal matrix with the corresponding eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{128}$  on its diagonal. Denoising was achieved by projecting  $\hat{\mathbf{X}}^{protein}$  onto  $U = 10$  dominant eigenvectors only

$$\hat{\mathbf{X}}^{denoised} = \mathbf{E}^{red} (\mathbf{E}^{red})^H \hat{\mathbf{X}}^{protein}, \quad (4.19)$$

whereas  $\mathbf{E}^{red}$  denoted the  $128 \times 10$  matrix having the eigenvectors  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{10}$  in its column and the  $128 \times 2048$ -matrix  $\hat{\mathbf{X}}^{denoised}$  contained the centered denoised signals in its rows. In this step

$$\frac{\sum_{u=1}^{10} \lambda_u}{\sum_{m=1}^{128} \lambda_m} \cdot 100\% \approx 80\% \quad (4.20)$$

of the variance of the data remained and the missing 20% were removed in order to denoise the spectra.

Finally, the mean was added back to  $\hat{\mathbf{X}}^{denoised}$  according to

$$\hat{\mathbf{X}}^{denoised} \leftarrow \hat{\mathbf{X}}^{denoised} + \mathbf{m}_{\hat{\mathbf{X}}^{protein}} \mathbf{1}_{2048}^T. \quad (4.21)$$

in the last step of the denoising procedure

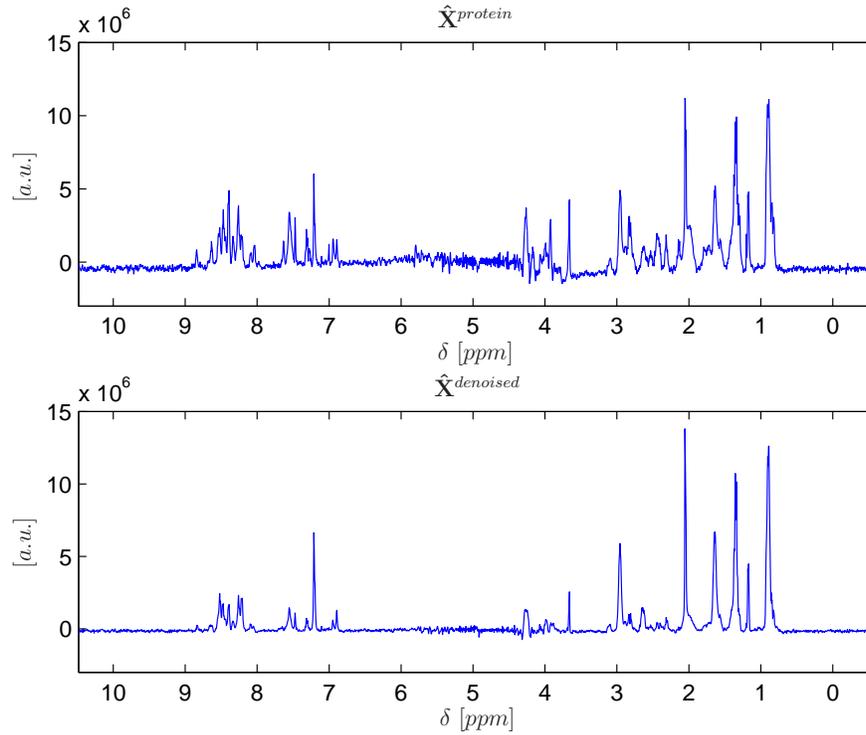


Figure 4.2: Denoising of the water free P11 spectra by PCA. Top: the noisy spectrum obtained after removing the water signal from the original P11 spectra. Bottom: the water free P11 spectrum after PCA based denoising. Clearly, the protein peaks got modified unacceptably or totally vanished from the spectrum.

The result is shown in Fig. 4.2. As can be seen the noise was reduced significantly, however, also a considerable fraction of the smaller peaks vanished from the spectra and some of the larger peaks got distorted unacceptably. Increasing the number of dominant eigenvectors considered in the projection step lead to better recovered peaks but the denoising effect was then rather marginal. Hence, it must be concluded that denoising based on standard PCA cannot be applied to protein spectra.

## 4.2 Local PCA

As the PCA approach described in the last section was not suitable to reduce the noise in the water free P11 spectra an extension to it, called local PCA [60], was investigated. Local PCA makes use of the principles of local linear projective noise reduction. The idea is to project noisy signals in a high-dimensional space which will be called feature space henceforth. In this space clusters of similar vectors are formed in which standard PCA denoising is carried out. Afterwards the data are transformed back to input space by inverting both clustering and embedding. As will be seen this approach is capable of removing the additional noise from the water free P11 spectra while keeping the protein peaks virtually untouched.

### 4.2.1 Concept of Local PCA Denoising

In local PCA it is assumed again that the noise can, at least locally, be modelled as a white stationary Gaussian process. Signals usually come from a deterministic or at least predictable source and can be described as a smooth function evaluated at discrete time steps small enough to capture the characteristics of the function. That implies, using a dynamical model for the data, that the signal embedded in delayed coordinates resides within a sub-manifold of the feature space spanned by these delayed coordinates. With local projective denoising techniques such as local PCA the task is to detect this signal manifold.

Note that delayed coordinates are an ideal tool for representing the signal information as for example in the context of chaotic dynamical systems the delayed coordinates of an observable component captures the full dynamical system [54]. There also exists a similar result in statistics for signals with a limited decaying memory [49].

In order to illustrate how local PCA is carried out formally assume that

**Data:** Observed signal  $x(t) = [x(0), x(1), \dots, x(T-1)]$   
**Input:**  $D$  (embedding dimension)  
 $\Delta t$  (delay)  
 $K$  (number of clusters)  
 $U$  (number of dominant eigenvectors considered)  
**Result:** denoised observation  
 $x^{denoised} = [x(0)^{denoised}, x(1)^{denoised}, \dots, x^{denoised}(T-1)]$

- 1 **begin**
- 2   Form trajectory matrix  $\mathbf{X}^{tr}$ ;
- 3   K-means clustering of columns of  $\mathbf{X}^{tr}$ ;
- 4   Form sub-trajectory matrix  $\mathbf{X}_k^{tr}$  from each cluster;
- 5   PCA denoising of each  $\mathbf{X}_k^{tr}$  by Alg. 3;
- 6   Invert clustering;
- 7   Invert embedding;
- 8 **end**

**Algorithm 4:** Denoising based on local PCA.

$T$  samples  $x(0), x(1), \dots, x(T-1)$  of a signal  $x(t)$ <sup>1</sup> are available, i.e.

$$x(t) = [x(0), x(1), \dots, x(T-1)]. \quad (4.22)$$

This signal is embedded into a  $D$  dimensional feature space by forming a trajectory matrix  $\mathbf{X}^{tr}$  whose rows consist of  $D$  signals

$$x_d = [x(d\Delta t), x(d\Delta t+1), \dots, x(T-1-(D-1-d)\Delta t)], \quad d = 0, 1, \dots, D-1 \quad (4.23)$$

which are mutually shifted by  $\Delta t \in \mathbb{N} \setminus \{0\}$  data points:

$$\mathbf{X}^{tr} = [x_{D-1}, x_{D-2}, \dots, x_0]^T = \quad (4.24)$$

$$\begin{bmatrix} x((D-1)\Delta t) & x((D-1)\Delta t+1) & x((D-1)\Delta t+2) & \dots & x(T-1) \\ x((D-2)\Delta t) & x((D-2)\Delta t+1) & x((D-2)\Delta t+2) & \dots & x(T-1-\Delta t) \\ x((D-3)\Delta t) & x((D-3)\Delta t+1) & x((D-3)\Delta t+2) & \dots & x(T-1-2\Delta t) \\ \vdots & \vdots & \vdots & & \vdots \\ x(\Delta t) & x(\Delta t+1) & x(\Delta t+2) & \dots & x(T-1-(D-2)\Delta t) \\ x(0) & x(1) & x(2) & \dots & x(T-1-(D-1)\Delta t) \end{bmatrix}$$

---

<sup>1</sup>Note that in the following  $t$  denotes an arbitrary discrete index and not necessarily a time index.

The columns

$$\mathbf{x}^{tr}(i) = \begin{bmatrix} x((D-1)\Delta t + i) \\ x((D-2)\Delta t + i) \\ x((D-3)\Delta t + i) \\ \vdots \\ x(\Delta t + i) \\ x(i) \end{bmatrix}, \quad 0 \leq i \leq T-1-(D-1)\Delta t \quad (4.25)$$

of this matrix are then considered as vectors in a  $D$  dimensional space.

For local PCA these vectors are grouped by similarity into  $K$  disjoint clusters by e.g.  $K$ -means [33]. Let

$$\{\mathbf{x}^{tr}(i_{1,k}), \mathbf{x}^{tr}(i_{2,k}), \dots, \mathbf{x}^{tr}(i_{L_k,k})\} \quad (4.26)$$

with

$$i_{j,k} \in \{0, 1, \dots, T-1-(D-1)\Delta t\}, \quad 1 \leq j \leq L_k$$

be the set of columns of  $\mathbf{X}^{tr}$  grouped into the  $k$ -th cluster ( $1 \leq k \leq K$ ) whereas  $L_k$  denotes the cluster size. The columns in each cluster are used to constitute  $K$  sub-trajectory matrices

$$\mathbf{X}_k^{tr} = [\mathbf{x}^{tr}(i_{1,k}), \mathbf{x}^{tr}(i_{2,k}), \dots, \mathbf{x}^{tr}(i_{L_k,k})] \quad (4.27)$$

which are analyzed one by one by PCA.

In particular, if the signal  $x(t)$  is to be denoised then each sub-trajectory matrix undergoes the denoising procedure as described in Sec. 4.1, i.e. each  $\mathbf{X}_k^{tr}$  is fed into Alg. 3 separately. After the PCA step  $K$  denoised sub-matrices

$$\mathbf{X}_k^{den} = [\mathbf{x}^{den}(i_{1,k}), \mathbf{x}^{den}(i_{2,k}), \dots, \mathbf{x}^{den}(i_{L_k,k})] \quad (4.28)$$

are available whose columns  $\mathbf{x}^{den}(i_{j,k})$  consist of the denoised vectors  $\mathbf{x}^{tr}(i_{j,k})$ . Based on these matrices a new overall trajectory matrix  $\mathbf{X}^{den}$  is formed by inverting the clustering carried out in the beginning of the local PCA procedure. In detail, if the  $l$ -th column of  $\mathbf{X}^{tr}$  was used to constitute the  $i$ -th column of  $\mathbf{X}_k^{tr}$  then the  $i$ -th column of  $\mathbf{X}_k^{den}$  forms the  $l$ -th column of  $\mathbf{X}^{den}$ .

Note that  $\mathbf{X}^{den} = \{x_{pq}^{den}\}_{1 \leq p \leq D, 1 \leq q \leq T-1-(D-1)\Delta t}$  has essentially the same structure as  $\mathbf{X}^{tr}$ . This means that if

$$x^{denoise}(t) = [x^{denoise}(0), x^{denoise}(1), \dots, x^{denoise}(T-1)] \quad (4.29)$$

denotes the signal  $x(t)$  after denoising then the element  $x^{denoise}(t=i)$  constitutes the elements

$$\begin{aligned} \{x_{p,q}^{den} | p \in \{n \in \mathbb{N} \setminus \{0\} | 1 \leq n \leq D \wedge n \geq D - \frac{i}{\Delta t} \wedge n \leq \frac{T-1-i}{\Delta t} + 1\}, \\ q = i + (p-D)\Delta t + 1\} =: \mathbb{L}_i \end{aligned} \quad (4.30)$$

$$\mathbf{X}^{den} = \begin{bmatrix} x^{denoise}(4) & \mathbf{x}^{denoise}(5) & x^{denoise}(6) & x^{denoise}(7) & x^{denoise}(8) & x^{denoise}(9) & x^{denoise}(10) \\ x^{denoise}(2) & x^{denoise}(3) & x^{denoise}(4) & \mathbf{x}^{denoise}(5) & x^{denoise}(6) & x^{denoise}(7) & x^{denoise}(8) \\ x^{denoise}(0) & x^{denoise}(1) & x^{denoise}(2) & x^{denoise}(3) & x^{denoise}(4) & \mathbf{x}^{denoise}(5) & x^{denoise}(6) \end{bmatrix}$$

Figure 4.3: Illustration of the set  $\mathbb{L}_i$ . Consider as example a process  $x(t) = [x(0), x(1), \dots, x(10)]$  which was embedded into a  $D = 3$  dimensional space using a delay of  $\Delta t = 2$ . After denoising the trajectory matrix  $\mathbf{X}^{den}$  is formed whose rows consist of shifted versions of the denoised signal  $x^{denoise}(t) = [x^{denoise}(0), x^{denoise}(1), \dots, x^{denoise}(10)]$ . The elements of  $\mathbf{X}^{den}$  constituting the set  $\mathbb{L}_5$  are displayed in blue. Note that only under optimal conditions all elements of  $\mathbb{L}_5$  are identical while they differ slightly in practice. Hence, they need to be averaged in order to obtain the final denoised data point  $x^{denoise}(5)$ .

of  $\mathbf{X}^{den}$  (see Fig. 4.3).

In practice, the elements of the set  $\mathbb{L}_i$  are usually not identical but need to be averaged according to

$$x^{denoise}(t = i) = \frac{1}{\#\mathbb{L}_i} \sum_{x_{p,q}^{den} \in \mathbb{L}_i} x_{p,q}^{den}, \quad (4.31)$$

whereas  $\#\mathbb{L}_i$  denotes the number of elements of the set  $\mathbb{L}_i$ , in order to get a good estimate of  $x^{denoise}$ .

Summarized three differences appear between PCA and local PCA denoising. First, in local PCA each recorded signal is denoised separately while in the classical approach all observed signals are analyzed simultaneously. Second, the signals are embedded into a higher dimensional space by means of the concept of delayed coordinates and are clustered by similarity before the actual PCA step is carried out. And finally, the averaging (Eq. (4.31)) at the end of the local PCA procedure helps to cancel out any noise contributions which may have remained after the projection step was performed in each cluster.

### 4.2.2 Selection of Parameters

One drawback of the local PCA denoising procedure described above is that four parameters, namely the embedding dimension  $D$ , the delay  $\Delta t$ , the number of cluster  $K$  and the number  $U$  of dominant eigenvectors, need to be tuned in order to obtain optimal results. While  $U$  can be determined

straightforwardly, an trial an error approach is needed to find the best combination of  $D$ ,  $\Delta t$  and  $K$ .

#### 4.2.2.1 Determination of $U$

For the automatic determination of the optimal number  $U_{opt}$  of dominant eigenvectors considered during the projection step a minimum description length (MDL) [45] approach has been discussed in the literature [59]. Closer investigations [45, 26] have shown, however, that this method often leads to too large values of  $U$  such that the signals are not denoised sufficiently. Hence, a different concept to estimate  $U_{opt}$  is proposed in the following which is applicable if the variance  $var_{noise}$  of the noise in the embedded signals can be estimated.

To illustrate this approach recall that in PCA denoising it is assumed that the actual (noise free) signal resides in a subspace of the feature space which is spanned by the  $U$  dominant eigenvectors. The remaining  $D - U$  eigenvectors are supposed to span another subspace in which the noise with which the signal was corrupted resides. Furthermore, it was shown in Eq. (4.12) that the variance of the signal projected onto the  $d$ -th eigenvector ( $1 \leq d \leq D$ ) is given by the  $d$ -th largest eigenvalue  $\lambda_d$ . Altogether this means that the sum of the  $D - U$  smallest eigenvalues equals the variance of the noise:

$$var_{noise} = \sum_{d=U+1}^D \lambda_d \quad (4.32)$$

Hence, if an estimate of  $var_{noise}$  is available the optimal value  $U_{opt}$  of  $U$  is that one for which the above equation is fulfilled as far as possible, i.e.

$$U_{opt} = \max \left\{ U \in \mathbb{N} \left| \sum_{d=U+1}^D \lambda_d \leq var_{noise} \right. \right\} \quad (4.33)$$

The same considerations hold in local PCA, however,  $U_{opt}$  must be determined separately in each cluster. Denoting by  $\lambda_{d,k}$  the eigenvalues determined in the  $k$ -th cluster this leads to the set of equations

$$U_{opt,k} = \max \left\{ U \in \mathbb{N} \left| \sum_{d=U+1}^D \lambda_{d,k} \leq var_{noise} \right. \right\} \quad (4.34)$$

whereas  $U_{opt,k}$  is the optimal number of dominant eigenvectors to project onto in the  $k$ -th cluster.

In order to elucidate how  $var_{noise}$  can be estimated in the case of NMR data sets consider the water free P11 spectrum  $\hat{x}^{protein}$  depicted in Fig. 4.4.

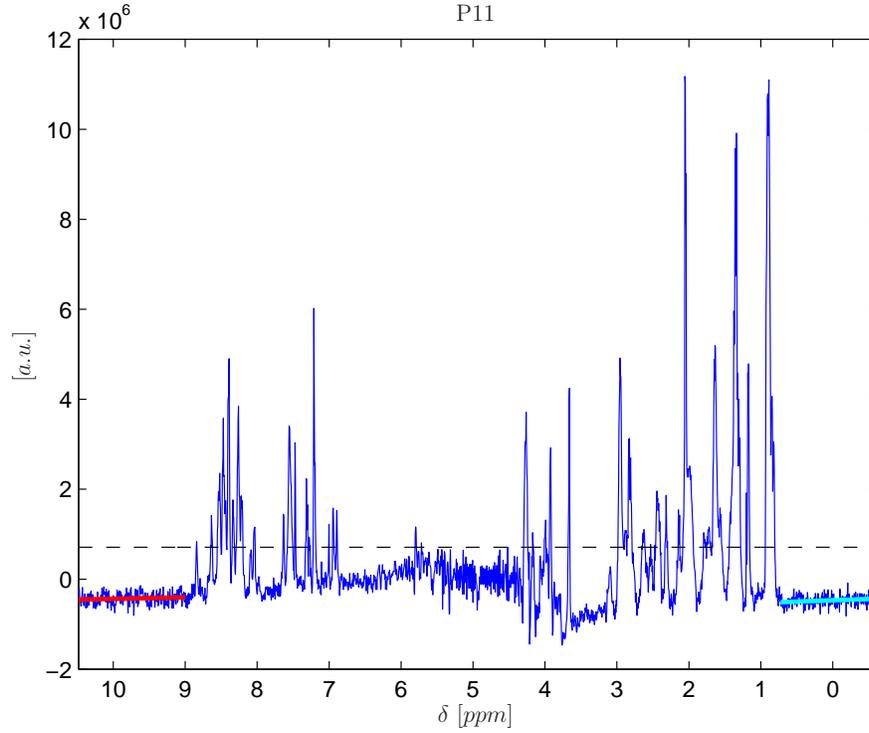


Figure 4.4: Denotations used to describe the parameter selection process in local PCA denoising considering as example the spectrum of P11. Shown in blue is the P11 spectrum recorded for the shortest evolution period  $t_1$  after the water signal was removed by MP-BSS. This spectrum is denoted by  $\hat{x}^{protein}$  in the text. The peak free areas  $\hat{x}_l^{protein} = \hat{x}^{Protein}(9 \text{ ppm} \leq \delta \leq 10.5 \text{ ppm})$  and  $\hat{x}_r^{protein} = \hat{x}^{Protein}(0.6 \text{ ppm} \leq \delta \leq -0.7 \text{ ppm})$  were used to estimate the variance of the noise. Concatenated they formed the sub-signal  $\hat{x}_{baseline}^{protein} = [\hat{x}_l^{protein}, \hat{x}_r^{protein}]$ . The first order polynomial fits  $\hat{x}_l^{poly}$  and  $\hat{x}_r^{poly}$  of these two sections are shown in red and cyan, respectively. These fits were concatenated and constituted the sub-signal  $\hat{x}^{poly} = [\hat{x}_l^{poly}, \hat{x}_r^{poly}]$ . Peaks larger than  $\tau_{act} = 7.1 \cdot 10^5$  (black dashed line) were used to evaluate the impact of the denoising procedure on the peaks. They were concatenated and formed the sub-signal  $\hat{x}_{peak}^{protein}$ .

Like any other  $t_2$ -Fourier transformed spectra  $\hat{x}^{protein}$  contains two peak free sections  $\hat{x}_l^{protein} = \hat{x}^{protein}(9 \text{ ppm} \leq \delta \leq 10.5 \text{ ppm})$  and  $\hat{x}_r^{protein} = \hat{x}^{protein}(0.6 \text{ ppm} \leq \delta \leq -0.7 \text{ ppm})$  at its left and right end. Ideally, these sections should have zero variance as they ought to contain a straight baseline only. In practice, however, a nonvanishing variance  $var_{baseline}$  is observed because of the presence of noise. Thus,  $var_{noise}$  can be estimated as

$$var_{noise} = D var_{baseline} \quad (4.35)$$

whereas the factor  $D$  allows for the embedding of the signal. Here, it is assumed that the noise is stationary in the sense that its variance is constant over the entire spectrum.

#### 4.2.2.2 Determination of $D$ , $\Delta t$ , and $K$

As described in the last section the optimal number  $U_{opt}$  of dominant eigenvectors can be determined directly by estimating the noise in the spectra. Unfortunately, no such straightforward approach exists for the determination of the best combination of the remaining parameters, i.e. the optimal values for the embedding dimension  $D$ , the delay  $\Delta t$  and the number of clusters  $K$  must be found by trial and error.

In order to compare the results obtained for various sets of these parameters a quantitative performance measure is needed which allows for the following two features: first, the noise reduction must be measured and second, any unwanted impacts of the denoising procedure onto the peaks must be quantized.

For the noise estimation a similar approach as in Sec. 4.2.2.1 is made in the sense that only the two peak free sections at the left and right end of the noisy and the denoised spectrum are considered. As a first step the peak-free sections  $\hat{x}_r^{protein}$  and  $\hat{x}_l^{protein}$  of the noisy spectrum are fitted one by one by a first order polynomial function

$$f_{poly}(t) = P_1 t + P_0 \quad (4.36)$$

which best describes the data in the least squares sense. The resulting fits, which will be denoted by  $\hat{x}_r^{poly}$  and  $\hat{x}_l^{poly}$ , respectively, in the sequel, are assumed to represent the theoretical baseline which would be have been observed if the spectrum was not corrupted by noise. Hence, the difference between these fits and the corresponding peak-free sections  $\hat{x}_l^{denoised}$  and  $\hat{x}_r^{denoised}$ , respectively, of the denoised spectrum can be used to estimate the remaining level of noise.

More formally, let  $\hat{x}_{baseline}^{denoised} = [\hat{x}_l^{denoised}, \hat{x}_r^{denoised}]$  be the concatenation of the two peak-free sections of the denoised signal and let  $\hat{x}^{poly} = [\hat{x}_l^{poly}, \hat{x}_r^{poly}]$

denote the subsignal combining the two polynomial fits. The noise in the spectrum can then be quantized by means of the signal to noise ratio

$$SNR_{noise} = SNR(\hat{x}^{poly}, \hat{x}_{baseline}^{denoised}). \quad (4.37)$$

Note, that in these steps as well as in the remainder of this section only the real part of the spectrum is considered.

In order to investigate the impact of the denoising procedure onto the peaks the following approach is carried out. First, the peaks are defined formally as the parts of the noisy spectrum which have an amplitude larger than a user defined threshold  $\tau_{act}$  (see Fig. 4.4). These parts are concatenated to constitute the sub-signal  $\hat{x}_{peak}^{protein} = [\hat{x}_{peak}^{protein}(0), \hat{x}_{peak}^{protein}(1), \dots, \hat{x}_{peak}^{protein}(L-1)]$ ,  $1 \leq L \leq T$ . After denoising, the same areas of the denoised spectrum are used to constitute the sub-signal  $\hat{x}_{peak}^{denoised} = [\hat{x}_{peak}^{denoised}(0), \hat{x}_{peak}^{denoised}(1), \dots, \hat{x}_{peak}^{denoised}(L-1)]$ .

Based on these two sub-signals the  $SNR(\hat{x}_{peak}^{protein}, \hat{x}_{peak}^{denoised})$  could now be used to quantify the impact of the denoising procedure onto the actual signals. However, it must be noted that two reasons may lead to differences between  $\hat{x}_{peak}^{protein}$  and  $\hat{x}_{peak}^{denoised}$ . On the one hand, the differences may simply originate from the fact that the noise was removed from  $\hat{x}_{peak}^{protein}$ . Obviously, such deviations are welcome and should not be penalized by the performance measure. On the other hand, the applied denoising procedure may also have lead to further distortions in the signal, e.g. some of the peaks in an NMR spectrum may have disappeared entirely or may have changed their shape or amplitude considerably. Such changes are bothersome and should hence be sensed by the performance measure. Thus, a performance measure is needed which can distinguish between unwanted and desired modifications of the denoised spectrum.

This problem can be addressed by disregarding differences between  $\hat{x}_{peak}^{protein}$  and  $\hat{x}_{peak}^{denoised}$  in the computation of  $SNR(\hat{x}_{peak}^{protein}, \hat{x}_{peak}^{denoised})$  which are smaller than the maximum amplitude  $n_{max}$  of the noise in the noisy spectrum. In the case of NMR data  $n_{max}$  is determined as the maximal difference between the two peak-free sections at the left and right end of the noisy spectrum and the corresponding polynomial fits:

$$n_{max} = \max(|\hat{x}_{baseline}^{protein} - \hat{x}_{poly}|). \quad (4.38)$$

where  $\hat{x}_{baseline}^{protein} = [\hat{x}_l^{protein}, \hat{x}_r^{protein}]$ .

This value is then subtracted from the differences between  $\hat{x}_{peak}^{protein}$  and  $\hat{x}_{peak}^{denoised}$  before the SNR is computed. Formally, this leads to the performance measure

$$SNR_{signal} = 20 \log_{10} \frac{\|\hat{x}_{peak}^{protein}\|}{\|g(\Delta\hat{x}_{peak})\|} [dB] \quad (4.39)$$

Parameter	Start value	Increment	End value
$D$	10	10	100
$\Delta t$	1	1	5
$K$	10	10	100

Table 4.1: The start and end values of the parameters  $D$ ,  $\Delta t$  and  $K$  as well as the step size in which they were incremented throughout the local PCA denoising procedure of P11.

where  $\Delta\hat{x}_{peak} := \hat{x}_{peak}^{protein} - \hat{x}_{peak}^{denoised}$ ,

$$g(\Delta\hat{x}_{peak}) = [g(\Delta\hat{x}_{peak}(0)), g(\Delta\hat{x}_{peak}(1)), \dots, g(\Delta\hat{x}_{peak}(L-1))]$$

and

$$g(\Delta\hat{x}_{peak}(l)) = \begin{cases} \text{sgn}(\Delta\hat{x}_{peak}(l))(|\Delta\hat{x}_{peak}(l)| - n_{max}) & \text{for } |\Delta\hat{x}_{peak}(l)| > n_{max} \\ 0 & \text{else} \end{cases} \quad (4.40)$$

Thus,  $SNR_{signal}$  considers only those differences between  $\hat{x}_{peak}^{protein}$  and  $\hat{x}_{peak}^{denoised}$  which cannot be explained by the removal of noise.

### 4.2.3 Local PCA Denoising of P11 Spectra

The local PCA denoising procedure described above was used to remove the noise from the water free P11 spectra (see Fig. 4.4) whereas the parameters shown in Tab. 4.1 were used. In order to estimate the variance of the noise the peak-free sections  $\hat{x}_l^{protein} = \hat{x}^{protein}(9 \text{ ppm} \leq \delta \leq 10.5 \text{ ppm})$  and  $\hat{x}_r^{protein} = \hat{x}^{protein}(-0.6 \text{ ppm} \leq \delta \leq -0.7 \text{ ppm})$  were considered whereas  $\hat{x}^{protein}$  denotes the first row of  $\hat{\mathbf{X}}^{Protein}$  (cf. Eq.(3.47)). Concatenated these two sections were 524 data points long and constituted the signal  $\hat{x}_{baseline}^{protein} = [\hat{x}_l^{protein}, \hat{x}_r^{protein}] = [\hat{x}_{baseline}^{protein}(0), \hat{x}_{baseline}^{protein}(1), \dots, \hat{x}_{baseline}^{protein}(523)]$ . This signal was used to estimate the variance of the noise in the spectrum whereas a value of  $var_{baseline} = 1.54 \cdot 10^{10}$  was obtained. Although this value was large enough to remove the noise from the peak-free areas  $\hat{x}_l^{protein}$  and  $\hat{x}_r^{protein}$  it turned out to be too small to sufficiently denoise the center, i.e. the section  $4.5 \text{ ppm} < \delta < 7.5 \text{ ppm}$  which was formerly occupied by the water signal. Hence, a three times larger value of  $var_{baseline} = 4.62 \cdot 10^{10}$  was used throughout the experiments.

Altogether, the P11 spectrum was denoised 500 times by local PCA (Alg. 4) whereas the individual parameters listed in Tab. 4.1 were used. In this process the number  $U_{opt,k}$  of dominant eigenvectors considered during

the projection step in each cluster was determined according to

$$U_{opt,k} = \max \left\{ U \in \mathbb{N} \left| \sum_{d=U+1}^D \lambda_{d,k} \leq 4.62 \cdot 10^{10} \right. \right\} \quad (4.41)$$

In order to compare quantitatively how well the noise was removed from the resulting 500 denoised spectra the polynomial fits of  $\hat{x}_l^{protein}$  and  $\hat{x}_r^{protein}$  had to be determined. For the fit of  $\hat{x}_l^{protein}$  the coefficients  $P_1$  and  $P_0$  in Eq. 4.36 were set to 80 and  $-4.5 \cdot 10^5$ , respectively, while for  $\hat{x}_r^{protein}$   $P_1 = 90$  and  $P_0 = -6.3 \cdot 10^5$  were used. These fits were concatenated to form the sub-signal  $\hat{x}^{poly}$  which was inserted together with  $\hat{x}_{baseline}^{denoised}$  into

$$SNR_{noise} = SNR(\hat{x}^{poly}, \hat{x}_{baseline}^{denoised}) \quad (4.42)$$

in order to quantify the remaining noise in the spectrum.

Furthermore, it was determined how far the local PCA denoising procedure affected the actual peaks. For this purpose, signals larger than  $\tau_{act} = 7.1 \cdot 10^5$  were considered and concatenated in order to form the subsignal  $\hat{x}_{peak}^{protein}$  which consisted of 331 data points (see Fig. 4.4). After denoising, the corresponding signals were used to constitute the 331 data points long subsignal  $\hat{x}_{peak}^{denoised}$ . Furthermore, the largest amplitude of the noise was estimated by determining the largest difference between  $\hat{x}^{poly}$  and  $\hat{x}_{baseline}^{protein}$  whereas a value of  $n_{max} = 3.7 \cdot 10^5$  was obtained. The difference signal  $\Delta \hat{x}_{peak} = \hat{x}_{peak}^{protein} - \hat{x}_{peak}^{denoised}$  was computed and inserted together with  $\hat{x}_{peak}^{protein}$  and  $n_{max}$  into Eq. 4.39 to determine  $SNR_{signal}$ .

The results are summarized in Fig. 4.5. As can be seen good denoising results (i.e. large  $SNR_{noise}$  and  $SNR_{baseline}$  values) were only obtained if at least  $K = 30$  clusters were used. For  $K < 30$  the number of clusters was too small to separate unressembling parts of the spectrum. For instance, for  $K > 30$  clusters appeared which contained only the baseline at the left of the spectrum (i.e.  $\hat{x}_l^{protein}$ ) while for smaller  $K$ 's no such clear separation between peak-free and peak containing areas was observed.

However, for  $K > 50$  the values of  $SNR_{noise}$  were decreasing again. This is not surprising as in such cases the number of cluster members in each cluster became small such that single outliers sufficed to distort the estimates made during the PCA procedure significantly.

Concerning the embedding dimension it was observed that for larger values of  $D$   $SNR_{noise}$  increased at the expense of  $SNR_{signal}$ . Quite the opposite observation was made for increasing delays  $\Delta t$  which lead to better  $SNR_{signal}$  values but at the same time to worse  $SNR_{noise}$ 's (see Fig. 4.6).

Obviously, it was hard to tell from the plot in Fig. 4.5 which combination of the parameters  $K$ ,  $D$ , and  $\Delta t$  lead to an optimal result, i.e. to both a high

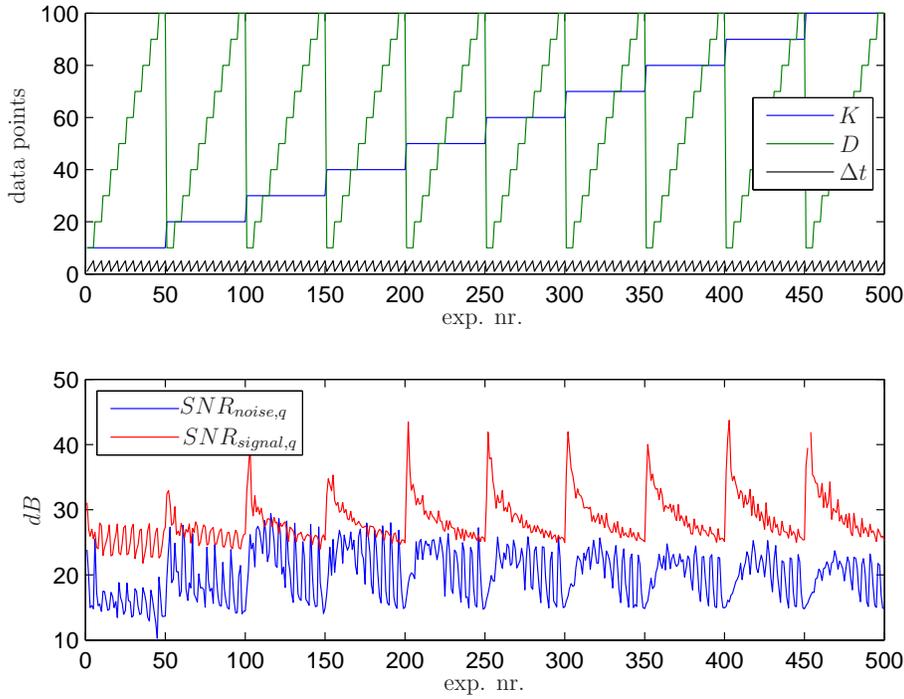


Figure 4.5: Results of the local PCA denoising procedure of P11. Altogether, 500 experiments were carried out. Top: the parameters used in the individual experiments plotted over the experiment numbers. Bottom:  $SNR_{noise,q}$  and  $SNR_{signal,q}$ ,  $1 \leq q \leq 500$ , as obtained for the parameter sets shown above (also plotted over the experiment numbers).

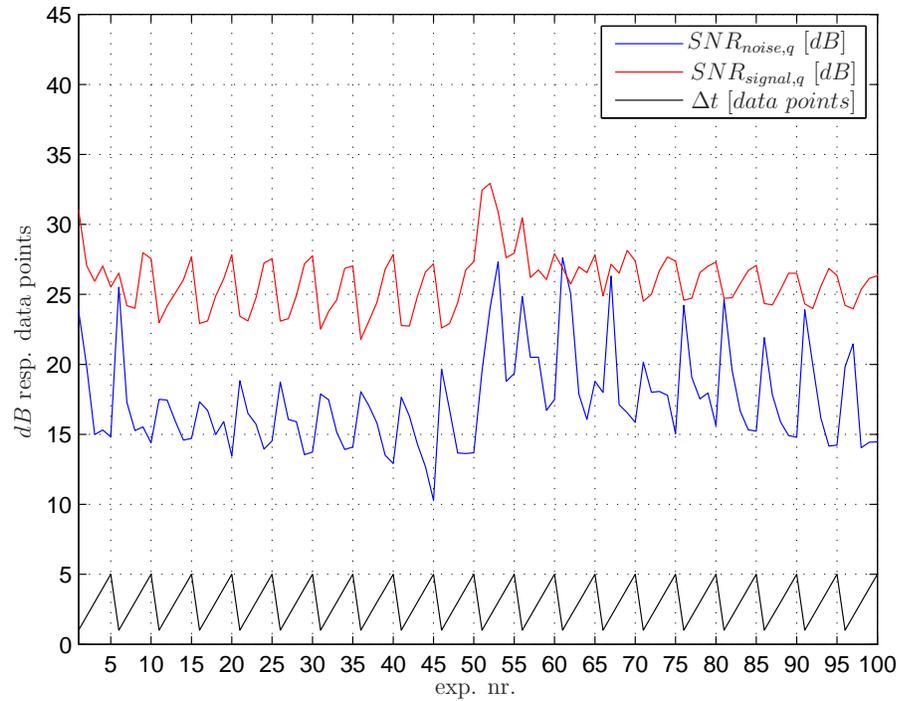


Figure 4.6: Dependence of  $SNR_{noise,q}$  and  $SNR_{q,signal}$  on the delay  $\Delta t$  for the first 100 experiments. Obviously,  $SNR_{denoised}$  increased for larger  $\Delta t$ 's at the expense of  $SNR_{noise}$ . The same behavior was observed in the other 400 experiments, however, sometimes the correlation was less obvious.

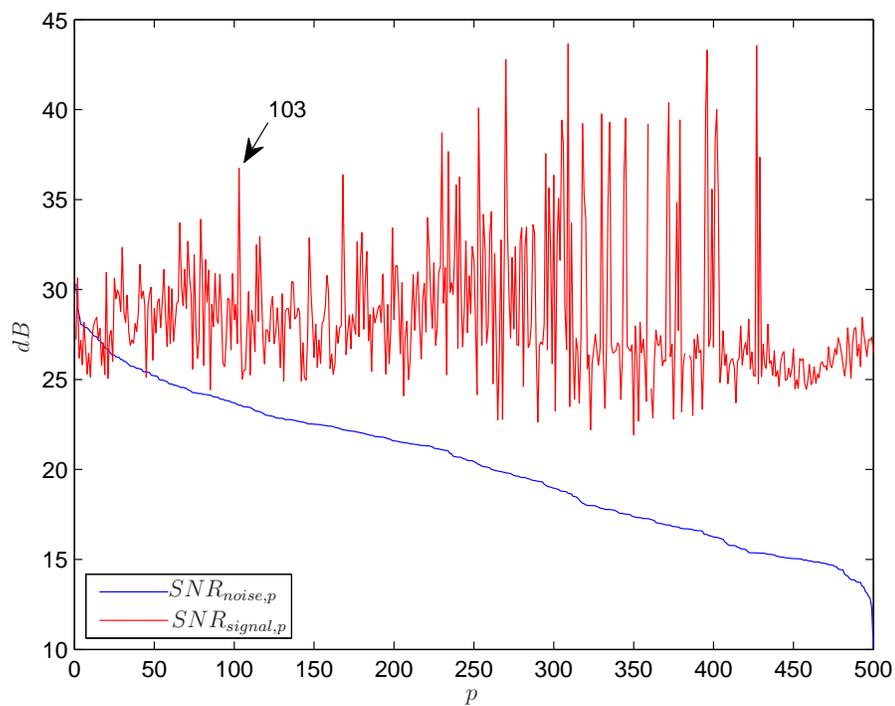


Figure 4.7: The rearranged values  $SNR_{noise,p}$  and  $SNR_{signal,p}$ . The best compromise between a large  $SNR_{noise,p}$  and an acceptable  $SNR_{signal,p}$  is achieved for  $p_{103}$  (see also Fig. 4.8).

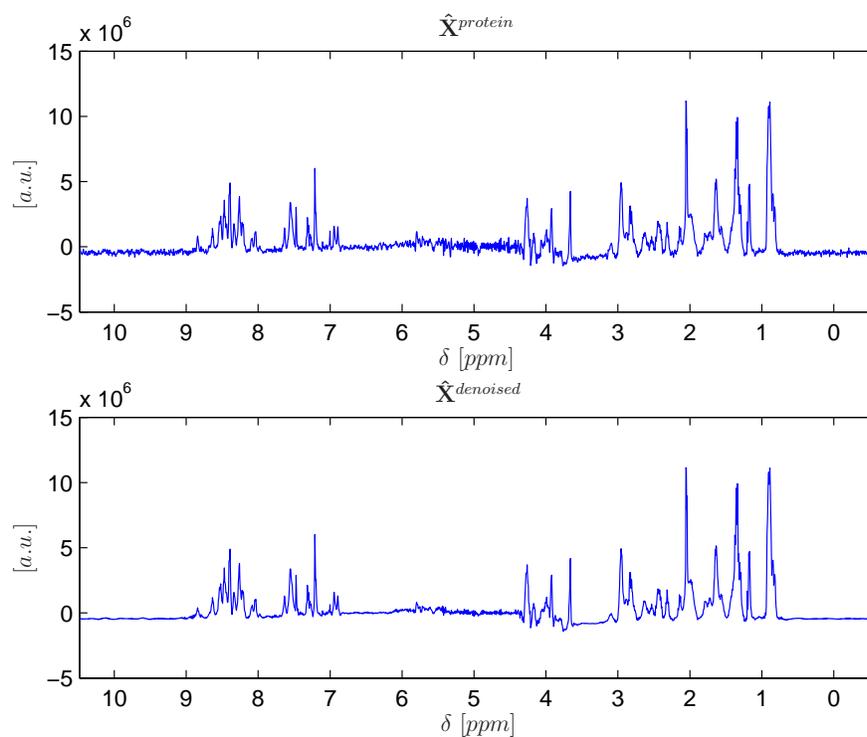


Figure 4.8: Comparison between the water free P11 spectrum before (top) and after local PCA denoising (bottom). Obviously, the noise got removed significantly while even small peaks remained undistorted in the denoised spectrum. For local PCA the parameters  $K = 40$ ,  $D = 10$ , and  $\Delta t = 3$  were used.

$SNR_{noise}$  and at the same time to an acceptable  $SNR_{signal}$ . For this reason, a permutation  $p_1, p_2, \dots, p_{500}$  of the experiment numbers  $q = 1, 2, \dots, 500$  was determined for which the individual  $SNR_{noise,q}$  values computed throughout the 500 experiments were arranged in descending order, i.e.

$$SNR_{noise,p_1} \leq SNR_{noise,p_2} \leq \dots \leq SNR_{noise,p_{500}} \quad (4.43)$$

The same permutation was used to sort the individual  $SNR_{signal,q}$  values whereupon both performance measures were plotted in Fig. 4.7. As can be seen, a good compromise between  $SNR_{signal,153} = 37 \text{ dB}$  and  $SNR_{noise,153} = 24 \text{ dB}$  was obtained for experiment number  $p_{103} = 153$ , for which the parameter set  $K = 40$ ,  $D = 10$ , and  $\Delta t = 3$  was used.

For comparison  $SNR_{noise}$  values were also computed for the original spectrum (as shown in the top of Fig. 3.16) as well as for the noisy water free P11 spectrum whereas values of  $32 \text{ dB}$  and  $11 \text{ dB}$  were obtained, respectively. These values show that local PCA denoising succeeded in reducing the noise contribution in the water free P11 spectra considerably even if the high value computed for the original spectrum could not be reached.

The denoised spectrum is also shown in Fig. 4.8. Clearly, the noise got removed significantly and in contrast to standard PCA denoising even small peaks were well-conserved.

Hence, it may be concluded that local PCA is an eligible method to remove the noise from  $t_2$ -Fourier transformed NMR spectra. The optimal number of dominant eigenvectors used during the PCA step can be estimated well *a priori* while optimal values for the number of clusters  $K$ , the embedding dimension  $D$  and the delay  $\Delta t$  must be determined by trial and error. Note, however, that local PCA denoising is a rapid procedure such that the 500 experiments described above took less than 15 minutes of computation time on a normal PC. Together with the proposed performance measures which quantize the obtained results this means that the optimal parameter sets can be determined readily.

### 4.3 Kernel PCA

Another extension to standard PCA as described in Sec. 4.1 is Kernel PCA [51, 38], a procedure in which the observed data is first mapped into a possibly high dimensional feature space by a nonlinear mapping before the eigenvectors and eigenvalues are computed. In fact, the nonlinear mapping is not carried out explicitly, though, as all necessary operations conducted in feature space can be expressed in terms of the input space vectors by means of positive definite kernel functions. This so-called kernel trick can generally be

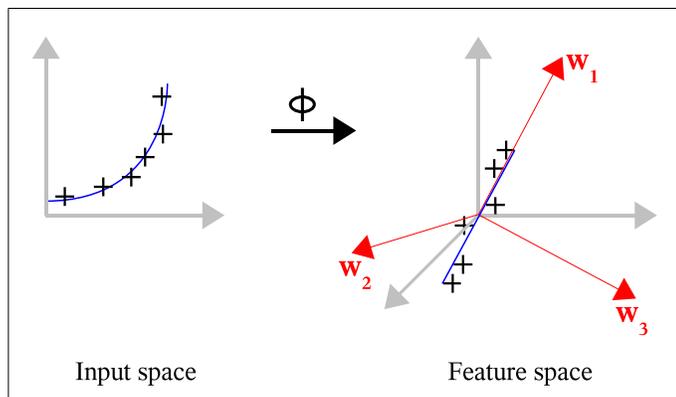


Figure 4.9: Nonlinear PCA: Sample data stemming from a nonlinear process are nonlinearly mapped to a high dimensional feature space where linear PCA based denoising is then applied. In feature space, the mapped data points are assumed to show a more linear structure and projections onto more eigenvectors may be considered than in input space.

applied to all algorithms which are solely expressed in terms of dot products, i.e. which do not use the variables themselves explicitly.

Historically, the Kernel trick was already discussed in the literature in the 1960s [2], however, little use of it was made until the 1990s when Vapnik applied it to generalize Support Vector machines [58]. Nowadays, it is a popular method to adapt algorithms which were originally devised to solve linear problems to nonlinear settings [37, 24, 3, 27].

### 4.3.1 Concept of Kernel PCA

#### 4.3.1.1 Nonlinear PCA

Generally, two advantages are attributed to kernel PCA in comparison to its classical counterpart. On the one hand mapping the data into a higher dimensional feature space allows the computation of more eigenvectors such that e.g. in denoising tasks a better separation between the sub-spaces containing the actual signal and the noise, respectively, can be expected. On the other hand it is assumed that the mapped data can be better represented by the eigensystem computed in feature space if an appropriate non-linear mapping is used.

Fig. 4.9 sketches the idea. In the left of Fig. 4.9 the crosses represent  $T$  noisy samples  $\mathbf{x}_t$  of the nonlinear process indicated by the solid line. The problem being two-dimensional only two eigenvectors can be determined by PCA. Following the PCA based denoising procedure presented in Sec. 4.1 pro-

jections onto the eigenvector corresponding to the smaller eigenvalue would be neglected and only the projection onto the eigenvector related with the larger eigenvalue would be considered. But projecting a curved line onto one eigenvector only results in a straight line such that after denoising the nonlinear process would appear to be linear.

The basic idea of nonlinear PCA is to map the data first by a nonlinear mapping  $\Phi$  to a high dimensional feature space where linear PCA is then performed (see right of Fig. 4.9). This leads to two advantages. On the one hand, if an eligible feature map  $\Phi$  can be found the data should show a more linear structure in feature space than in input space. On the other hand, with the feature space being high dimensional, more eigenvectors can be determined than in input space. So e.g. the projections onto the first two eigenvectors in the right of Fig. 4.9 could be used to describe the linear and any remaining nonlinear structure of the mapped data while the projection onto the third eigenvector corresponding to the smallest eigenvalue can be neglected in order to reduce noise. After denoising in feature space the data is finally mapped back to input space.

Formally, the first step of nonlinear PCA is to map the input space data  $\mathbf{x}_t \in \mathcal{X}$ ,  $t = 0, 1, \dots, T-1$ , by a nonlinear feature map  $\Phi$  to feature space  $\mathcal{H}$  [51]:

$$\Phi : \mathcal{X} \rightarrow \mathcal{H} \quad (4.44)$$

$$\mathbf{x} \mapsto \Phi(\mathbf{x}). \quad (4.45)$$

Here,  $\mathcal{H}$  is a dot product space of arbitrary large dimension.

In analogy to linear PCA, nonlinear PCA now boils down to the eigenvalue decomposition of the feature space correlation matrix which is estimated from the mapped samples  $\Phi(\mathbf{x}_t)$ ,  $t = 0, 1, \dots, T-1$ , as

$$\mathbf{C}_{\Phi(\mathbf{x})} \approx \frac{1}{T-1} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t) \Phi(\mathbf{x}_t)^\top, \quad (4.46)$$

where, for simplicity, the  $\Phi(\mathbf{x}_t)$ 's are assumed to be centered in feature space. In other words, the eigenvalue problem

$$\lambda_k \mathbf{v}_k = \mathbf{C}_{\Phi(\mathbf{x})} \mathbf{v}_k, \quad k = 0, 1, \dots, F-1 \quad (4.47)$$

for  $\lambda_k \neq 0$  must be solved in feature space  $\mathcal{H}$  whereas  $F$  denotes the feature space dimension. Note that the solutions  $\mathbf{v}_k$  lie in the span of  $\Phi(\mathbf{x}_0)$ ,  $\Phi(\mathbf{x}_1)$ ,  $\dots$ ,  $\Phi(\mathbf{x}_{T-1})$  as

$$\lambda_k \mathbf{v}_k = \mathbf{C}_{\Phi(\mathbf{x})} \mathbf{v}_k = \frac{1}{T-1} \sum_{t=0}^{T-1} \langle \Phi(\mathbf{x}_t), \mathbf{v}_k \rangle \Phi(\mathbf{x}_t), \quad (4.48)$$

where  $\langle \cdot, \cdot \rangle$  denotes the dot product in  $\mathcal{H}$ . This has two useful consequences. First, instead of Eq. (4.47) the following set of equations may be considered:

$$\lambda_k \langle \Phi(\mathbf{x}_t), \mathbf{v}_k \rangle = \langle \Phi(\mathbf{x}_t), \mathbf{C}\mathbf{v}_k \rangle \text{ for all } t = 0, 1, \dots, T-1. \quad (4.49)$$

Second, there exist coefficients  $a_{tk}$ , ( $t = 0, 1, \dots, T-1$ ) such that

$$\mathbf{v}_k = \sum_{t=0}^{T-1} a_{tk} \Phi(\mathbf{x}_t). \quad (4.50)$$

Combining Eq. (4.50) and Eq. (4.49) leads to

$$\begin{aligned} \lambda_k \sum_{t=0}^{T-1} a_{tk} \langle \Phi(\mathbf{x}_n), \Phi(\mathbf{x}_t) \rangle = \\ \frac{1}{T-1} \sum_{t=0}^{T-1} a_{tk} \left\langle \Phi(\mathbf{x}_n), \sum_{u=0}^{T-1} \Phi(\mathbf{x}_u) \langle \Phi(\mathbf{x}_u), \Phi(\mathbf{x}_t) \rangle \right\rangle \end{aligned} \quad (4.51)$$

for all  $n = 0, 1, \dots, T-1$ . By defining the  $T \times T$  Gram matrix  $\mathbf{K}$  with elements

$$K_{tu} = \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_u) \rangle, \quad 0 \leq t, u \leq T-1 \quad (4.52)$$

Eq. (4.51) can be simplified to

$$(T-1)\lambda_k \mathbf{K}\mathbf{a}_k = \mathbf{K}^2\mathbf{a}_k, \quad (4.53)$$

where  $\mathbf{a}_k$  denotes the column vector with entries  $a_{1k}, \dots, a_{Tk}$ .

Eq. (4.53) may seem to be hard to tackle at first glance, however, it was shown in [51] that its solutions can be determined by solving the dual problem

$$(T-1)\lambda_k \mathbf{a}_k = \mathbf{K}\mathbf{a}_k \quad (4.54)$$

for nonzero eigenvalues  $\lambda_k$ , i.e. by computing the eigenvalue decomposition of the matrix  $\mathbf{K}$  whereas the eigenvectors  $\mathbf{a}_k$  are assumed to be orthonormal. Inserting the  $\mathbf{a}_k$ 's into Eq. (4.50) actually leads to the desired eigensystem of  $\mathbf{C}_{\Phi(\mathbf{x})}$ , however, an additional normalization step is still needed to assure that the  $\mathbf{v}_k$ 's have unit norm. The corresponding normalization condition can be found readily as

$$\begin{aligned} 1 \stackrel{!}{=} \langle \mathbf{v}_k, \mathbf{v}_k \rangle &= \sum_{t,u=0}^{T-1} a_{tk} a_{uk} \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_u) \rangle = \sum_{t,u=0}^{T-1} a_{tk} a_{uk} K_{tu} = \langle \mathbf{a}_k, \mathbf{K}\mathbf{a}_k \rangle \\ &= (T-1)\lambda_k \langle \mathbf{a}_k, \mathbf{a}_k \rangle. \end{aligned} \quad (4.55)$$

such that the eigenvectors  $\mathbf{a}_k$  of  $\mathbf{K}$  must be divided by  $\sqrt{(T-1)\lambda_k}$  before they can be inserted into Eq. (4.50).

Once the principal axes  $\mathbf{v}_k$  have been determined in feature space  $\mathcal{H}$  the nonlinear principal components  $\psi_k$  of a mapped data point  $\Phi(\mathbf{x})$  can be computed as in classical PCA (cf. Eq. (4.11)), i.e.

$$\psi_k = \langle \mathbf{v}_k, \Phi(\mathbf{x}) \rangle = \sum_{t=0}^{T-1} a_{tk} \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}) \rangle \quad (4.56)$$

#### 4.3.1.2 Centering in Feature Space

So far, it has been assumed that the mapped vectors  $\Phi(\mathbf{x}_t)$ ,  $t = 0, 1, \dots, T-1$  are centered in feature space while in practice their mean usually differs from zero. However, the derivations made in the last section hold analogously if instead of the  $\Phi(\mathbf{x}_t)$ 's the centered counterparts

$$\tilde{\Phi}(\mathbf{x}_t) := \Phi(\mathbf{x}_t) - \frac{1}{T} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t) \quad (4.57)$$

are inserted into the above equations. In particular, Eq. (4.54), in which the eigensystem of the feature space vectors is actually determined, is replaced by the eigenvalue problem

$$(T-1)\tilde{\lambda}_k \tilde{\mathbf{a}}_k = \tilde{\mathbf{K}} \tilde{\mathbf{a}}_k \quad (4.58)$$

whereas  $\tilde{\mathbf{K}}$  is the Gram matrix of the centered feature space vectors with elements

$$\tilde{K}_{tu} = \langle \tilde{\Phi}(\mathbf{x}_t), \tilde{\Phi}(\mathbf{x}_u) \rangle. \quad (4.59)$$

Straightforward calculations [51] show that  $\tilde{\mathbf{K}}$  can be derived from  $\mathbf{K}$  by

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{I}_T \mathbf{K} - \mathbf{K} \mathbf{I}_T + \mathbf{I}_T \mathbf{K} \mathbf{I}_T \quad (4.60)$$

whereas  $\mathbf{I}_T$  is a  $T \times T$  matrix whose elements equal all  $1/T$ .

After the eigenvalue problem Eq. (4.58) has been solved the eigenvectors  $\tilde{\mathbf{v}}_k$  as well as the nonlinear principal components  $\tilde{\psi}_k$  of a test point  $\Phi(\mathbf{x})$  can be determined as described in the last section, i.e.

$$\tilde{\mathbf{v}}_k = \sum_{t=0}^{T-1} \tilde{a}_{tk} \tilde{\Phi}(\mathbf{x}_t) \quad (4.61)$$

and

$$\begin{aligned}
\tilde{\psi}_k &= \langle \tilde{\mathbf{v}}_k, \tilde{\Phi}(\mathbf{x}) \rangle = \sum_{t=0}^{T-1} \tilde{a}_{tk} \langle \tilde{\Phi}(\mathbf{x}_t), \tilde{\Phi}(\mathbf{x}) \rangle = \\
&\stackrel{\text{Eq. (4.57)}}{=} \sum_{t=0}^{T-1} \tilde{a}_{tk} \left\{ \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}) \rangle - \frac{1}{T} \sum_{u=0}^{T-1} \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_u) \rangle \right. \\
&\quad \left. - \frac{1}{T} \sum_{w=0}^{T-1} \langle \Phi(\mathbf{x}_w), \Phi(\mathbf{x}) \rangle + \frac{1}{T^2} \sum_{u=0}^{T-1} \sum_{w=0}^{T-1} \langle \Phi(\mathbf{x}_u), \Phi(\mathbf{x}_w) \rangle \right\}, \tag{4.62}
\end{aligned}$$

respectively. Note that in Eq. (4.62) it is assumed that the mapped vector  $\tilde{\Phi}(\mathbf{x})$  has the same distribution as the  $\tilde{\Phi}(\mathbf{x}_t)$ 's which were used to determine the principal axes such that

$$\tilde{\Phi}(\mathbf{x}) = \Phi(\mathbf{x}) - \sum_{u=0}^{T-1} \Phi(\mathbf{x}_u) \tag{4.63}$$

can be used.

### 4.3.1.3 Kernel Trick

The nonlinear PCA procedure described so far essentially consisted of two steps: first, the data had to be mapped to feature space by the nonlinear function  $\Phi$  where, second, standard PCA was carried out. The explicit mapping of the input vectors  $\mathbf{x}_t$ , can be circumvented, however, if so-called positive definite kernel functions are used advantageously. These functions are defined as follows:

**Definition 12** (positive definite kernel [47]). *Let  $\mathcal{X}$  be a nonempty set. A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  which for all  $T \in \mathbb{N}$  and all  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{T-1} \in \mathcal{X}$  gives rise to a positive definite Gram matrix  $\mathbf{K}$  with elements  $K_{tu} = k(\mathbf{x}_t, \mathbf{x}_u)$ ,  $t, u = 0, 1, \dots, T-1$ , is called a positive definite kernel.*

Positive definite kernels have the appealing property that they compute dot products between mapped vectors  $\Phi(\mathbf{x}_t)$  and  $\Phi(\mathbf{x}_u)$  in terms of the corresponding input space vectors  $\mathbf{x}_t$  and  $\mathbf{x}_u$ , respectively, as outlined by the following theorem:

**Theorem 5** (kernel induced feature space). *Let  $\mathcal{X}$  be a topological space and let  $k$  be a continuous positive definite kernel on  $\mathcal{X} \times \mathcal{X}$ . There exists a Hilbert space  $\mathcal{H}$  and a continuous mapping  $\Phi : \mathcal{X} \rightarrow \mathcal{H}$  such that for all  $\mathbf{x}_t, \mathbf{x}_u \in \mathcal{X}$*

$$k(\mathbf{x}_t, \mathbf{x}_u) = \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_u) \rangle. \tag{4.64}$$

**Data:** Samples  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{T-1}$   
**Input:** Positive definite kernel  $k(\mathbf{x}_t, \mathbf{x}_u)$ , test point  $\mathbf{x}$ .  
**Result:** Nonlinear principal components  $\tilde{\boldsymbol{\psi}} = [\tilde{\psi}_0, \tilde{\psi}_1, \dots, \tilde{\psi}_{F-1}]$  of  $\mathbf{x}$

- 1 **begin**
- 2     Compute kernel matrix  $\mathbf{K}$  with  $K_{u,t} = k(\mathbf{x}_u, \mathbf{x}_t)$ ;
- 3     Compute centered kernel matrix  $\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{I}_T \mathbf{K} - \mathbf{K} \mathbf{I}_T + \mathbf{I}_T \mathbf{K} \mathbf{I}_T$ ;
- 4     EVD of  $\tilde{\mathbf{K}}$ :  $(T-1)\tilde{\lambda}_k \tilde{\mathbf{a}}_k = \tilde{\mathbf{K}} \tilde{\mathbf{a}}_k$ ;
- 5     Normalize  $\tilde{\mathbf{a}}_k$ :  $\tilde{\mathbf{a}}_k \leftarrow \frac{1}{(T-1)\tilde{\lambda}_k} \tilde{\mathbf{a}}_k$ ;
- 6     Compute nonlinear principle components  $\tilde{\psi}_k$  of  $\mathbf{x}$ :
- 7

$$\tilde{\psi}_k = \sum_{t=0}^{T-1} \tilde{a}_{tk} \left\{ k(\mathbf{x}_t, \mathbf{x}) - \frac{1}{T} \left( \sum_{u=0}^{T-1} k(\mathbf{x}_t, \mathbf{x}_u) - \sum_{w=0}^{T-1} k(\mathbf{x}_w, \mathbf{x}) \right) + \frac{1}{T^2} \sum_{u,w=0}^{T-1} k(\mathbf{x}_u, \mathbf{x}_w) \right\};$$

- 8 **end**

**Algorithm 5:** Kernel PCA algorithm obtained by replacing the dot products  $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$  in Eq. (4.52) and Eq. (4.62), respectively, by a positive definite kernel function  $k(\mathbf{x}_i, \mathbf{x}_j)$ .

Theo. 5 is a short form of Mercer's theorem of functional analysis which deals with positive definite kernels in the context of integral operators. For a more detailed discussion of Mercer's theorem the interested reader is referred to [47, 14, 16, 39].

For feature space algorithms, in which the mapped input space vectors  $\Phi(\mathbf{x}_t)$  only appear in dot products, Theo. 5 has the useful consequence that an explicit mapping of the input space vectors to feature space becomes superfluous. However, instead of searching for an eligible feature map  $\Phi$  an appropriate kernel function must be found now. Often, the replacement of feature space dot products  $\langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_u) \rangle$  by kernel functions  $k(\mathbf{x}_t, \mathbf{x}'_u)$  is referred to as the kernel trick.

Obviously, this trick can be applied to the nonlinear PCA procedure described above as in it the variables appear only in terms of dot products (cf. Eq. (4.59) in connection with Eq. (4.60) and Eq. (4.52), Eq. (4.62)). Thus, the kernelized nonlinear PCA procedure as outlined in Alg. 5 can be formulated.

### 4.3.2 Kernel PCA Denoising and the Pre-Image Problem

Certainly, one of the most important questions arising in Kernel PCA is which of the infinitely many positive definite kernels is eligible for solving a particular problem. For denoising purposes the Gaussian kernel

$$k_G(\mathbf{x}_t, \mathbf{x}_u) = \exp\left(-\frac{\|\mathbf{x}_t - \mathbf{x}_u\|^2}{2\sigma^2}\right) \quad (4.65)$$

is usually selected which owes most of its popularity to the fact that in the feature space induced by it vectors  $\Phi(\mathbf{x}_t)$  and  $\Phi(\mathbf{x}_u)$  are linearly independent for all  $\mathbf{x}_t \neq \mathbf{x}_u$ , i.e. the corresponding feature space is infinite dimensional.

In practice, this means that if  $T$  samples are available the kernel matrix  $\tilde{\mathbf{K}}$  will be of size  $T \times T$  such that  $T$  eigenvalues and corresponding eigenvectors can be determined. In most applications,  $T$  will exceed the input space dimension  $M$  considerably such that in feature space a better separation between eigenvectors carrying the actual signal and eigenvectors carrying noise can be expected. Denoising in feature space is achieved again as in the case of PCA or local PCA by projecting the mapped vectors  $\tilde{\Phi}(\mathbf{x}_t)$  onto the  $U \leq T$  dominant eigenvectors only:

$$\tilde{\Phi}(\mathbf{x}_t)^{denoised} = \sum_{u=0}^{U-1} \tilde{\psi}_u \tilde{\mathbf{v}}_u = \sum_{u=0}^{U-1} \tilde{\psi}_u \sum_{t=0}^{T-1} \tilde{a}_{tu} \tilde{\Phi}(\mathbf{x}_t) = \sum_{t=0}^{T-1} b_t \tilde{\Phi}(\mathbf{x}_t) \quad (4.66)$$

where  $b_t = \sum_{u=0}^{U-1} \tilde{\psi}_u \tilde{a}_{tu}$ .

As a last step of the kernel PCA denoising procedure, the vectors  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  have to be mapped back from feature space to input space in order to obtain the actual denoised signals  $\mathbf{x}_t^{denoised}$ . In this step the problem arises that the denoised feature space vectors  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  may not have an exact pre-image in input space, i.e. no vector  $\mathbf{z}$  in input space exists for which  $\Phi(\mathbf{z}) = (\tilde{\Phi}(\mathbf{x}_t))^{denoised}$  holds strictly.

The cause for this problem becomes obvious by reconsidering Eq. (4.66) in which it is shown that the denoised vector  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  is a linear combination of the mapped (centered) input space vectors  $\tilde{\Phi}(\mathbf{x}_t)$ . Although the mapped vectors  $\tilde{\Phi}(\mathbf{x}_t)$  apparently have a pre-image in input space this does not hold necessarily for linear combinations of them as is illustrated in Fig. 4.10 in which the Gaussian kernel  $k_G$  is considered as example. Loosely speaking, the Gaussian kernel maps each input space vector  $\mathbf{x}_t$  to a Gaussian bump centered at  $\mathbf{x}_t$  and  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  consists of a linear combination of all those bumps. However, it is well-known that no Gaussian function exists which

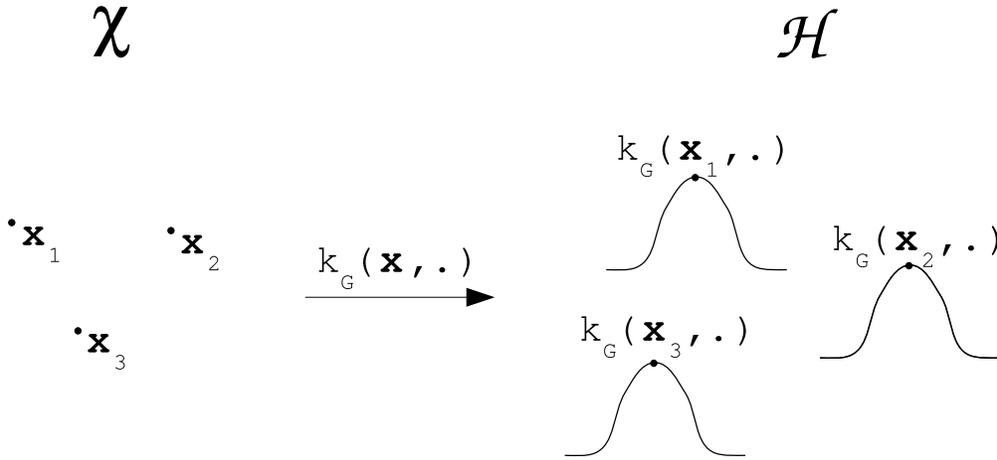


Figure 4.10: Illustration of the effect of the Gaussian kernel. The input space vectors  $\mathbf{x}_t$  are mapped to Gaussian bumps  $k(\mathbf{x}_t, \cdot)$ . The denoised vectors (not shown) consist of linear combinations of these bumps and have no pre-image in input space (see text).

equals a linear combination of Gaussians centered at different points and hence no exact pre-image of  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  exists in input space.

The problem can be addressed by searching for approximate pre-images of  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  in input space. For this purpose, several approaches have been discussed in the literature [38, 5, 50, 36] of which the two most popular ones will be discussed and compared in the following.

#### 4.3.2.1 Approximate Pre-Images by Minimum Feature Space Distance

One way to solve the pre-image problem approximately is to search for an input space vector  $\mathbf{z} \in \mathcal{X}$  whose centered mapped feature space image  $\tilde{\Phi}(\mathbf{z})$  is as close to the denoised vector  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  as possible. Formally, this leads to the optimization problem [50]

$$\text{minimize } f(\mathbf{z}) := \|\tilde{\Phi}(\mathbf{z}) - \tilde{\Phi}(\mathbf{x}_t)^{denoised}\|^2 \text{ w.r.t. } \mathbf{z} \quad (4.67)$$

with  $\tilde{\Phi}(\mathbf{z}) = \Phi(\mathbf{z}) - \frac{1}{T} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t)$ . In the following, this problem will be solved by gradient descent in which  $\mathbf{z}$  is updated according to

$$\mathbf{z} \leftarrow \mathbf{z} - \gamma \frac{df(\mathbf{z})}{d\mathbf{z}}, \quad \gamma \in \mathbb{R}^+ \setminus \{0\}. \quad (4.68)$$

In this procedure, the optimal  $\gamma$  will be determined by line search.

**Data:** expansion coefficients  $b_u$  of  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$ , number of samples  $T$   
**Input:** thresholds  $\epsilon_{LS}, \epsilon_{grad} \in \mathbb{R}^+$   
**Result:** pre-image  $\mathbf{z}$  of  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$

```

1 begin
2   Initialize  $\mathbf{z} = \mathbf{x}_t$ ;
3   repeat
4      $\mathbf{z}_{old} = \mathbf{z}$ ;
5     Compute  $\frac{df(\mathbf{z})}{d\mathbf{z}}$  (cf. Eq. (4.71));
6     Initialize  $A, B$  at random such that  $g_A \leq 0 \wedge g_B \geq 0$ ;
7     Compute  $g_{\frac{A+B}{2}}$  (cf. Eqs. (4.73), (4.74));
8     while  $|g_{\frac{A+B}{2}}| > \epsilon_{LS}$  do
9       if  $g_{\frac{A+B}{2}} < 0$  then
10         $A \leftarrow \frac{A+B}{2}$ ;
11      else
12         $B \leftarrow \frac{A+B}{2}$ ;
13      end
14    end
15     $\gamma = \frac{A+B}{2}$ ;
16     $\mathbf{z} \leftarrow \mathbf{z} - \gamma \frac{df(\mathbf{z})}{d\mathbf{z}}$ ;
17  until  $\|\mathbf{z}_{old} - \mathbf{z}\| \leq \epsilon_{grad}$ ;
18 end

```

**Algorithm 6:** Determination of approximate pre-images by gradient descent in combination with line search.

To start with note that for the Gaussian kernel  $k_G$  (see Eq. (4.65))  $f(\mathbf{z})$  can be expressed as

$$f(\mathbf{z}) = -\frac{1}{T} \sum_{t=0}^{T-1} k_G(\mathbf{z}, \mathbf{x}_t) - \sum_{u=0}^{T-1} b_u \left\{ k_G(\mathbf{z}, \mathbf{x}_u) - \frac{1}{T} \sum_{w=0}^{T-1} k_G(\mathbf{z}, \mathbf{x}_w) \right\} + C \quad (4.69)$$

where  $C$  is a constant independent of  $\mathbf{z}$  (see App. B).

By making use of

$$\frac{dk_G}{d\mathbf{z}} = \frac{d}{d\mathbf{z}} \exp\left(-\frac{\|\mathbf{z} - \mathbf{x}\|^2}{2\sigma^2}\right) = -\frac{1}{\sigma^2} k_G(\mathbf{z} - \mathbf{x}) \quad (4.70)$$

the gradient of  $f(\mathbf{z})$  is obtained readily as

$$\begin{aligned} \frac{df(\mathbf{z})}{d\mathbf{z}} &= \frac{1}{T\sigma^2} \sum_{t=0}^{T-1} k_G(\mathbf{z}, \mathbf{x}_t)(\mathbf{z} - \mathbf{x}_t) \\ &\quad + \frac{1}{\sigma^2} \sum_{u=0}^{T-1} b_u \left\{ k_G(\mathbf{z}, \mathbf{x}_u)(\mathbf{z} - \mathbf{x}_u) - \frac{1}{T} \sum_{w=0}^{T-1} k_G(\mathbf{z}, \mathbf{x}_w)(\mathbf{z} - \mathbf{x}_w) \right\} \end{aligned} \quad (4.71)$$

Furthermore, the optimal  $\gamma$  is needed for which  $f\left(\mathbf{z} - \gamma \frac{df(\mathbf{z})}{d\mathbf{z}}\right)$  reaches its minimum. For this purpose the gradient

$$g_A := \left. \frac{d}{d\gamma} f\left(\mathbf{z} - \gamma \frac{df(\mathbf{z})}{d\mathbf{z}}\right) \right|_{\gamma=A} \quad (4.72)$$

has to be computed if the optimal  $\gamma$  is to be found by means of line search. After some lengthy but straightforward calculations  $g_A$  can be expressed as follows for the Gaussian kernel  $k_G$ :

$$g_A = \frac{1}{T} \sum_{t=0}^{T-1} D_A(\mathbf{z}, \mathbf{x}_t) - \sum_{u=0}^{T-1} b_u \left\{ D_A(\mathbf{z}, \mathbf{x}_u) - \frac{1}{T} \sum_{w=0}^{T-1} D_A(\mathbf{z}, \mathbf{x}_w) \right\} \quad (4.73)$$

with

$$\begin{aligned} D_A(\mathbf{z}, \mathbf{x}) &:= \frac{1}{\sigma^2} k_G\left(\mathbf{z} - \gamma \frac{df(\mathbf{z})}{d\mathbf{z}}, \mathbf{x}\right) \\ &\quad \cdot \left( \left\langle \mathbf{z}, \frac{df(\mathbf{z})}{d\mathbf{z}} \right\rangle - \gamma \left\langle \frac{df(\mathbf{z})}{d\mathbf{z}}, \frac{df(\mathbf{z})}{d\mathbf{z}} \right\rangle - \left\langle \frac{df(\mathbf{z})}{d\mathbf{z}}, \mathbf{x} \right\rangle \right) \end{aligned} \quad (4.74)$$

Based on  $g_A$  the optimal value of  $\gamma$  can be determined by line search as follows: first, two scalars  $A, B \in \mathbb{R}$  with  $A < B$  are chosen such that  $g_A < 0$  and  $g_B > 0$  holds, i.e.  $A$  is left and  $B$  is right of the minimum to be found. Next,  $g_{\frac{A+B}{2}}$  is computed. If  $g_{\frac{A+B}{2}} > 0$  then  $B$  is set to  $\frac{A+B}{2}$  while otherwise  $A$  is updated by  $\frac{A+B}{2}$ . This procedure is repeated until  $g_{\frac{A+B}{2}}$  becomes smaller than a user defined threshold  $\epsilon_{LS}$  close to zero. Once this threshold is reached  $\gamma$  is set to the current value of  $\frac{A+B}{2}$ .

After the gradient of  $f(\mathbf{z})$  and the optimal  $\gamma$  have been computed  $\mathbf{z}$  is updated according to Eq. (4.68) and the entire procedure is reiterated until the changes in  $\mathbf{z}$  become negligible (see Alg. 6).

It turns out that the proposed method is sensitive to the initialization of  $\mathbf{z}$  in the beginning of the gradient descent method. In particular, if  $\mathbf{z}$  is

initialized such that it is far away from the actual pre-image the procedure often converges prematurely. Hence, it is proposed in [38] to initialize  $\mathbf{z}$  with the input space vector to be denoised (i.e. with  $\mathbf{x}_t$  if the pre-image of  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  is to be determined) as this vector should be spatially close to its denoised counterpart.

### 4.3.2.2 Approximate Pre-Images by Metric Multidimensional Scaling

Apart from iterative methods devised to map denoised feature space vectors  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  back to input space another popular approach [36] (see Alg. 7) is found in the literature in which the pre-image problem is solved by making use of concepts stemming from metric multidimensional scaling (MDS) [35].

In this approach squared feature space distances  $d_t^\Phi(\Phi(\mathbf{x}_i)^{denoised}, \Phi(\mathbf{x}_t))$  between  $\Phi(\mathbf{x}_i)^{denoised}$  and the mapped input space vectors  $\Phi(\mathbf{x}_t)$  are computed first and then transferred to squared input space distances  $d_t(\mathbf{x}^{denoised}, \mathbf{x}_t)$  between the pre-image and the input space vectors  $\mathbf{x}_t$ . Based on these  $d_t$ 's the approximate pre-image is determined as that input space vector  $\mathbf{z}$  for which the distance constraints in input space are fulfilled as far as possible. Note that often no  $\mathbf{z}$  can be found that strictly satisfies all distance constraints as usually no exact pre-image of  $\Phi(\mathbf{x}_i)^{denoised}$  exists in input space.

**Feature Space Distances.** As a first step of the metric MDS procedure the squared feature space distances

$$\begin{aligned} d_t^\Phi(\Phi(\mathbf{x}_i)^{denoised}, \Phi(\mathbf{x}_t)) &= \|\Phi(\mathbf{x}_i)^{denoised} - \Phi(\mathbf{x}_t)\|^2 = \\ &= \langle \Phi(\mathbf{x}_i)^{denoised}, \Phi(\mathbf{x}_i)^{denoised} \rangle - 2\langle \Phi(\mathbf{x}_i)^{denoised}, \Phi(\mathbf{x}_t) \rangle + \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_t) \rangle \end{aligned} \quad (4.75)$$

between the denoised feature space vector  $\Phi(\mathbf{x}_i)^{denoised} = \tilde{\Phi}(\mathbf{x}_i)^{denoised} + \frac{1}{T} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t)$  and the mapped input space vectors  $\Phi(\mathbf{x}_t)$ ,  $i, t = 0, 1, \dots, T-1$ , have to be determined.

After inserting Eq. (4.66) in the above equation the kernel trick is applied again such that the individual summands can be expressed as follows:

$$\begin{aligned} \langle \Phi(\mathbf{x}_i)^{denoised}, \Phi(\mathbf{x}_i)^{denoised} \rangle &= \mathbf{b}^\top \mathbf{K} \mathbf{b} + \frac{2}{T} \mathbf{b}^\top \mathbf{K} \mathbf{1}_T + \frac{1}{T^2} \mathbf{1}_T^\top \mathbf{K} \mathbf{1}_T \\ \langle \Phi(\mathbf{x}_i)^{denoised}, \Phi(\mathbf{x}_t) \rangle &= \mathbf{b}^\top \mathbf{k}_t + \frac{1}{T} \mathbf{1}_T^\top \mathbf{k}_t \\ \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_t) \rangle &= k(\mathbf{x}_t, \mathbf{x}_t) \end{aligned} \quad (4.76)$$

**Data:** expansion coefficients  $\mathbf{b}$  of  $\Phi(\mathbf{x}_i)^{denoised}$   
samples  $\mathbf{x}_t$   
kernel matrix  $\mathbf{K}$   
number of samples  $T$   
**Input:**  $L$ : number of next neighbors  
**Result:** pre-image  $\mathbf{z}$  of  $\Phi(\mathbf{x}_i)^{denoised}$

- 1 **begin**
- 2     Compute feature space distances  $d_t^\Phi$  (cf. Eq. (4.75), Eqs. (4.76));
- 3     Select  $L$  nearest neighbors  $\Phi(\mathbf{x}_{l_i}), i = 0, 1, \dots, L - 1$ , of  $\Phi(\mathbf{x}_i)^{denoised}$ ;
- 4     Determine input space distance constraints  $\mathbf{d} = [d_{l_0}, d_{l_1}, \dots, d_{l_{L-1}}]^\top$  (cf. Eq. (4.79));
- 5     Form  $\mathbf{X}_{neigh} = [\mathbf{x}_{l_0}, \mathbf{x}_{l_1}, \dots, \mathbf{x}_{l_{L-1}}]$ ;
- 6     SVD:  $\mathbf{X}_{neigh}\mathbf{H} = \mathbf{E}\mathbf{\Lambda}\mathbf{V}^\top = \mathbf{E}\mathbf{G}$  with  $\mathbf{H} = \mathbf{I}_L - \frac{1}{L}\mathbf{1}_L\mathbf{1}_L^\top$ ;
- 7     Form  $\mathbf{d}_0 = [\|\mathbf{g}_0\|^2, \|\mathbf{g}_1\|^2, \dots, \|\mathbf{g}_{L-1}\|^2]^\top$  where  $\mathbf{g}_i$  is  $i$ -th column of  $\mathbf{G}$ ;
- 8     Determine pre-image:  $\tilde{\mathbf{z}} = -\frac{1}{2}\mathbf{\Lambda}\mathbf{V}^\top(\mathbf{d} - \mathbf{d}_0)$ ;
- 9     Return to input space:  $\mathbf{z} = \mathbf{E}\tilde{\mathbf{z}} + \frac{1}{L}\sum_{i=0}^{L-1}\mathbf{x}_{l_i}$ ;
- 10 **end**

**Algorithm 7:** Determination of approximate pre-images by metric MDS for the Gaussian kernel.

whereas the vector  $\mathbf{b} = [b_0, b_1, \dots, b_{T-1}]^\top$  consists of the expansion coefficients of  $\tilde{\Phi}(\mathbf{x}_t)^{denoised}$  (cf. Eq. (4.66)) and  $\mathbf{k}_t$  denotes the  $t$ -th column of the kernel matrix  $\mathbf{K}$ .

**Input Space distances.** Once the squared feature space distances  $d_t^\Phi$  have been obtained the corresponding input space distances between the actual pre-image and the vectors  $\mathbf{x}_t, t = 0, 1, \dots, T - 1$ , need to be computed. In order to show how the feature space distances are related with their input space counterparts assume for the moment that an exact pre-image  $\mathbf{z}$  of  $\Phi(\mathbf{x}_i)^{denoised}$  exists in input space. In this case the squared feature space distance  $d_t^\Phi$  between  $\Phi(\mathbf{z})$  and  $\Phi(\mathbf{x}_t)$  reads

$$d_t^\Phi = \|\Phi(\mathbf{z}) - \Phi(\mathbf{x}_t)\|^2 = \langle \Phi(\mathbf{z}), \Phi(\mathbf{z}) \rangle - 2\langle \Phi(\mathbf{z}), \Phi(\mathbf{x}_t) \rangle + \langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_t) \rangle. \quad (4.77)$$

Replacing the dot products in the above equation by the Gaussian kernel  $k_G$  leads to

$$d_t^\Phi = 2 - 2 \exp\left(-\frac{d_t}{2\sigma^2}\right) \quad (4.78)$$

as  $k_G(\mathbf{x}, \mathbf{x}) = 1$ . Here,  $d_t = \|\mathbf{z} - \mathbf{x}_t\|^2$  denotes the squared input space distance between the pre-image and the input space vector  $\mathbf{x}_t$ . Obviously,  $d_t$  is determined by

$$d_t = -2\sigma^2 \ln \left( \frac{1}{2}(2 - d_t^\Phi) \right). \quad (4.79)$$

**Approximate Pre-Image Determination.** Next, a vector  $\mathbf{z}$  is sought in input space which fulfills the distance constraints  $d_t$ ,  $t = 0, 1, \dots, T - 1$ , as far as possible. It turns out, however, that poor results are obtained if all  $T$  constraints are considered such that usually only the distances between  $\mathbf{z}$  and the pre-images of the  $L < T$  nearest feature space neighbors of  $\Phi(\mathbf{x}_i)^{denoised}$  are taken into account. Note that similar strategies are used in locally linear embedding [46] where only the local neighborhood structure needs to be preserved and in metric MDS [15] in which smaller dissimilarities are given more weight.

In order to determine an approximate pre-image  $\mathbf{z}$  which best meets the  $L$  distance constraints in the least-square sense a relation devised by Gower [23] will be used below. For this relation the pre-images  $\mathbf{x}_{l_0}, \mathbf{x}_{l_1}, \dots, \mathbf{x}_{l_{L-1}}$  of the  $L$  nearest feature space neighbors of  $\Phi(\mathbf{x}_i)^{denoised}$  need are centered and a new coordinate system is defined in their span.

More formally, the  $L$  vectors  $\mathbf{x}_{l_i} \in \mathbb{R}^M$  are used to constitute the columns of the  $M \times L$  neighbor matrix  $\mathbf{X}_{neigh}$ . In the following it will be convenient to describe the centering of the  $\mathbf{x}_{l_i}$ 's by means of the  $L \times L$  matrix

$$\mathbf{H} = \mathbf{I}_L - \frac{1}{L} \mathbf{1}_L \mathbf{1}_L^\top \quad (4.80)$$

whereas  $\mathbf{I}_L$  denotes the  $L \times L$  identity matrix. With this matrix centered  $\mathbf{x}_{l_i}$ 's are obtained by  $\mathbf{H}\mathbf{X}_{neigh}^\top$ , i.e. the column sums of  $\mathbf{H}\mathbf{X}_{neigh}^\top$  are zero. Assuming that the  $\mathbf{x}_{l_i}$  span a  $Q$  dimensional space the singular value decomposition (SVD)

$$\mathbf{X}_{neigh}\mathbf{H} = \mathbf{E}\mathbf{\Lambda}\mathbf{V}^\top = \mathbf{E}\mathbf{G} \quad (4.81)$$

of the  $M \times L$  matrix  $(\mathbf{H}\mathbf{X}_{neigh}^\top)^\top = \mathbf{X}_{neigh}\mathbf{H}$  can be determined whereas the columns  $\mathbf{e}_j$  constituting the  $M \times Q$  matrix  $\mathbf{E} = [\mathbf{e}_0, \mathbf{e}_1, \dots, \mathbf{e}_{Q-1}]$  are orthonormal and  $\mathbf{G} = [\mathbf{g}_0, \mathbf{g}_1, \dots, \mathbf{g}_{L-1}]$  is an  $Q \times L$  matrix with columns  $\mathbf{g}_i$  being the projections of  $\mathbf{x}_{l_i}$  onto the  $\mathbf{e}_j$ 's.

Thus, the squared distances of the  $\mathbf{x}_{l_i}$ 's to the origin (i.e. to their centroid) equal  $\|\mathbf{g}_i\|^2$ . These distances are collected into the vector  $\mathbf{d}_0 = [\|\mathbf{g}_0\|^2, \|\mathbf{g}_1\|^2, \dots, \|\mathbf{g}_{L-1}\|^2]^\top$ .

Let  $\mathbf{d} = [d_{l_0}, d_{l_1}, \dots, d_{l_{L-1}}]^\top$  be the vector containing the  $L$  distances between the pre-image and its  $L$  neighbors as determined in Eq. (4.79). Based

on  $\mathbf{d}_0$  and  $\mathbf{d}$  the approximate pre-image which best satisfies the distance constraints  $\mathbf{d}$  in the least squares sense can be determined according to Gower [23] (see also [36]) as

$$\tilde{\mathbf{z}} = -\frac{1}{2}\mathbf{\Lambda}\mathbf{V}^\top(\mathbf{d} - \mathbf{d}_0). \quad (4.82)$$

However,  $\tilde{\mathbf{z}}$  is still expressed in terms of the coordinate system spanned by the  $\mathbf{x}_{l_i}$ . Hence, it must be transformed back to input space by

$$\mathbf{z} = \mathbf{E}\tilde{\mathbf{z}} + \frac{1}{L} \sum_{i=0}^{L-1} \mathbf{x}_{l_i} \quad (4.83)$$

in order to obtain the actual pre-image of  $\Phi(\mathbf{x}_i)^{denoised}$ .

Note that for the simulations and experiments described below the implementation of the metric MDS pre-image method provided by the ‘‘The Spider’’ toolbox [1] was used.

### 4.3.3 Kernel PCA Denoising of P11 Spectra

Before Kernel PCA could be applied to denoise water free P11 spectra the general question had to be answered which kind of overall denoising strategy should be used. Furthermore, a befitting value of the width  $\sigma$  of the Gaussian kernel had to be determined together with the optimal number  $U$  of dominant eigenvectors to project onto. Furthermore, the two pre-image algorithms presented in the last section had to be compared. The results of these investigations will be presented in this section.

#### 4.3.3.1 Overall Denoising Strategy

Generally, two different approaches can be pursued when real world NMR data are to be denoised by means of Kernel PCA. On the one hand all the spectra recorded during the NMR experiment can be denoised collectively, i.e. an analogues procedure as in standard PCA is carried out. In detail this means that the columns of the (water-free) data matrix  $\hat{\mathbf{X}}^{protein}$  (cf. Eq. (3.47)) are fed as samples  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{T-1}$  into the Kernel PCA algorithm (cf. Alg. 5) on which the denoising procedure is based.

On the other hand the individual rows of  $\hat{\mathbf{X}}^{protein}$  can be denoised one by one as in local PCA. For this approach a trajectory matrix  $\hat{\mathbf{X}}^{tr}$  (see Eq. (4.2.1)) is formed for each row of  $\hat{\mathbf{X}}^{protein}$  and the columns of  $\hat{\mathbf{X}}^{tr}$  are fed as the samples  $\mathbf{x}_t$  into the Kernel PCA algorithm. In this case, the procedure has to be repeated  $M$  times if  $M$  different spectra were recorded during the 2D-NOESY experiment.

<p><b>Data:</b> <math>M \times 2048</math> matrix <math>\hat{\mathbf{X}}^{protein}</math></p> <p><b>Input:</b> <i>Embedding:</i></p> <p><math>D</math>: number of delay</p> <p><math>\Delta t</math>: delay</p> <p><i>Kernel PCA:</i></p> <p><math>\sigma</math>: width of Gaussian kernel</p> <p><math>U</math>: number of dominant eigenvectors</p> <p><i>Pre-image problem:</i></p> <p><math>L</math>: number of next neighbors</p> <p><b>Result:</b> <math>\hat{\mathbf{X}}^{denoised}</math>: denoised matrix <math>\hat{\mathbf{X}}^{protein}</math></p> <pre> 1 begin 2   for <math>m = 0, 1, \dots, M</math> do 3     for <math>l = 1, 2, \dots, 4</math> do 4       Form subsegment: <math>\hat{x}^{seg}(t) = [\hat{x}^{seg}(0), \hat{x}^{seg}(1), \dots, \hat{x}^{seg}(511)]</math>                     <math>= \{\hat{\mathbf{X}}_{ij}^{protein}\}_{i=m, j=512(l-1), 512(l-1)+1, \dots, 512l-1}</math> ; 5       Form <math>D \times (512 - D\Delta t)</math> trajectory matrix                     <math>\hat{\mathbf{X}}^{tr} = [\hat{\mathbf{x}}_0^{tr}, \hat{\mathbf{x}}_1^{tr}, \dots, \hat{\mathbf{x}}_{511-D\Delta t}^{tr}]</math> from <math>\hat{x}^{seg}</math> ; 6       Denoising by Kernel PCA                     <math>\rightarrow \tilde{\Phi}(\hat{\mathbf{x}}_t^{tr}), t = 0, 1, \dots, 511 - D\Delta t</math> ; 7       Determine pre-image <math>\hat{\mathbf{z}}_t</math> of <math>\tilde{\Phi}(\hat{\mathbf{x}}_t^{tr}), t = 0, 1, \dots, 511 - D\Delta t</math> ; 8       Form matrix <math>\hat{\mathbf{X}}^{den} = [\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_{511-D\Delta t}]</math> ; 9       Constitute                     <math>\hat{x}^{denoised}(t) = [\hat{x}^{denoised}(0), \hat{x}^{denoised}(1), \dots, \hat{x}^{denoised}(511)]</math>                     from <math>\hat{\mathbf{X}}^{den}</math> by inverting the embedding ; 10      <math>\{\hat{\mathbf{X}}_{ij}^{denoised}\}_{i=m, j=512(l-1), 512(l-1)+1, \dots, 512l-1} = \hat{x}^{denoised}(t)</math> ; 11    end 12  end 13 end </pre>
---

**Algorithm 8:** Overall strategy applied in Kernel PCA denoising of NMR spectra which consist of 2048 data points. Each of the  $M$  spectra is divided into four 512 data points long subsegments (line 4) in order to avoid numerical instabilities in the EVD of the kernel matrix. These subsegments are embedded into a  $D$  dimensional space (line 5). The columns of the resulting trajectory matrix are fed into the Kernel PCA algorithm for denoising (line 6). After the pre-image has been determined (line 7) the embedding is inverted (lines 8, 9) and the denoised signal is used to constitute the rows of the matrix  $\hat{\mathbf{X}}^{denoised}$  (line 10).

In the experiments carried out in this investigation it turned out that the second approach is better suited to denoise NMR spectra than the first one. In particular, if the columns of the matrix  $\hat{\mathbf{X}}^{protein}$  were fed into the kernel PCA algorithm a significant reduction of noise always came along with unacceptable distortions of the smaller peaks. Hence, the combination of embedding and Kernel PCA is recommended for NMR spectra.

Furthermore, numerical problems may occur when Kernel PCA is used to denoise NMR spectra as these usually consist of 2048 data points. The latter means that the kernel matrix  $\mathbf{K}$  (see Eq. (4.52)) is of size  $2048 \times 2048$  such that numerical instabilities have to be expected when its eigenvalue decomposition is computed. Hence, the spectra were divided into four subsegments of size 512 data points (see Fig. 4.11) and each of these segments was provided separately to the Kernel PCA procedure. Together with the embedding approach this lead to the overall denoising strategy outlined in Alg. 8.

#### 4.3.3.2 Choice of Pre-Image Algorithm

When Kernel PCA denoising is to be applied to a new problem it is generally hard to anticipate which of the two pre-image algorithms presented above will lead to better results. The iterative method presented in Sec. 4.3.2.1, for instance, has the advantage that it is virtually free of critical parameters as the thresholds  $\epsilon_{LS}$  and  $\epsilon_{grad}$  only appear in the termination conditions of the algorithm. Hence, setting them to small enough values should guarantee good results. However, as any other gradient descent based method the iterative approach is prone to get stuck in local minima and may thus converge prematurely. In contrast, no convergence problems exist in the MDS based method (Sec. 4.3.2.2), however, the optimal number  $L$  of distance constraints needs to be set *a priori*.

It turned out that the convergence problem is severer than finding an optimal  $L$  when NMR spectra are to be denoised by Kernel PCA. To illustrate this the water free P11 spectrum depicted in the top of Fig. 4.11 was denoised by Kernel PCA in combination with the iterative pre-image method. For this purpose, the spectrum was divided into four subsegments of length 512 data points and each subsegment was used to form a trajectory matrix consisting of  $D = 11$  delayed signals shifted by  $\Delta t = 1$  data points. These trajectory matrices were fed one by one into the Kernel PCA algorithm whereas a Gaussian kernel of width  $\sigma = 5 \cdot 10^5$  was used. For denoising the data were projected onto the  $U = 4$  dominant eigenvectors only and the pre-images were determined by gradient descent whereas the thresholds  $\epsilon_{LS} = 10^{-20}$  and  $\epsilon_{grad} = 10^{-10}$  were used. Finally, the embedding was inverted such that

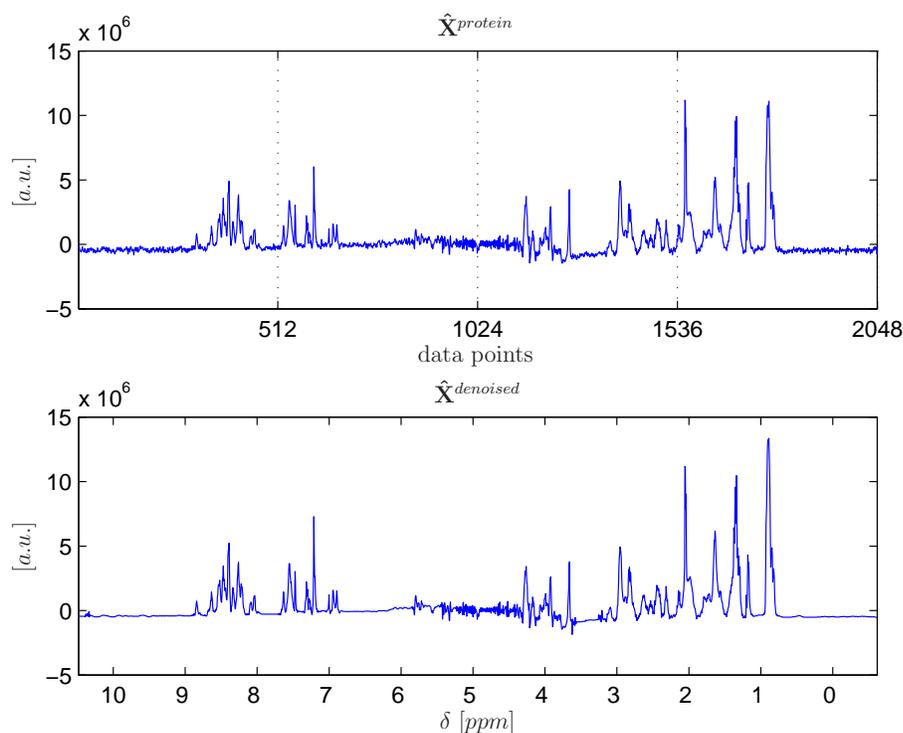


Figure 4.11: Deceptive result of the iterative pre-image algorithm. Top: the water free P11 spectrum obtained by MP-BSS plotted over data points instead of the chemical shift. Horizontal grid lines separate the 512 data points long segments which were denoised one by one by means of Kernel PCA. Bottom: the denoised spectrum obtained by Kernel PCA in combination with the iterative pre-image approach. The noise was only removed from the baseline but not from the peaks (see also Fig. 4.12).

the denoised spectrum shown in the bottom of Fig. 4.11 was obtained.

At first glance, the denoising procedure seemed to have succeeded as a smooth baseline together with well preserved peaks was observed. A closer inspection of the spectrum (see Fig. 4.12) revealed, though, that the noise was only removed from the peak-free segments of the spectrum but not from the peaks themselves. The latter are, however, the actual carrier of information such that it would be more important to remove the noise from them than from the baseline. Hence, the denoising procedure did not lead to any *significant* improvements in the spectrum.

The origin of the problem lay in the particular initialization of the iterative pre-image algorithm. Following the procedure outlined in Sec. 4.3.2.1

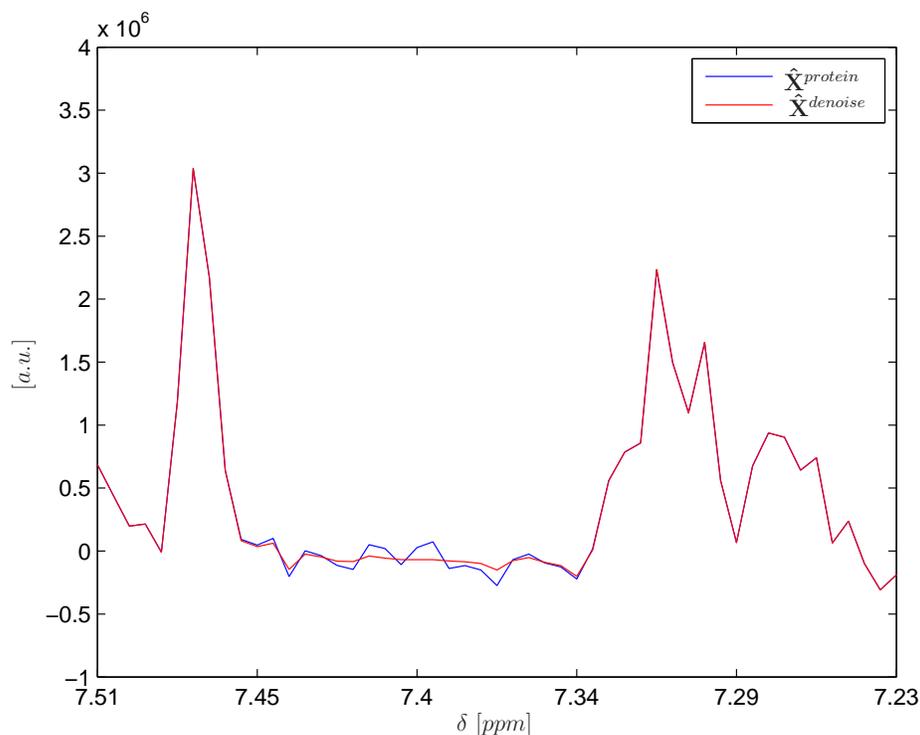


Figure 4.12: Close-up of the denoised P11 spectrum shown in Fig. 4.11. As can be seen the denoising procedure only affected the peak-free segment  $7.34 \text{ ppm} \leq \delta \leq 7.45 \text{ ppm}$  of the spectrum but did not denoise the peaks.

the noisy input space vectors were used to initialize the gradient descent algorithm in the hope to start from a point in search space which was already close to the actual pre-image such that the risk of getting stuck in a local minimum could be reduced.

Note, however, that if inappropriate values for  $\sigma$  and  $U$  are used in the Kernel PCA algorithm the denoised spectrum will be distorted significantly such that, for instance, its peaks become much smaller than the ones in the noisy spectrum. This means that the noisy vector and the corresponding denoised pre-image are far apart in search space such that the gradient descent procedure is likely to converge prematurely.

Apparently, such a premature convergence occurred already at the very beginning of the iterative pre-image method if peak-containing segments of the P11 spectrum were denoised. Thus, the iterative pre-image just reproduced the noisy data but did not reveal the actual pre-images of the denoised feature space vectors.

Hence, the MDS based pre-image procedure is recommended if NMR data sets are to be denoised albeit the resulting spectra hardly resembled its noisy counterpart in the experiment at hand. However, this was not a shortcoming of MDS based pre-imaging but was caused by the too small value of  $\sigma = 5 \cdot 10^5$  which impeded a reasonable Kernel PCA of the spectrum. Thus, MDS based pre-imaging clearly indicated ill-chosen parameters while the iterative approach lead to deceptive results because of initialization and premature convergence problems.

### 4.3.3.3 Determination of optimal parameters

As already pointed out in the last section an ill-chosen width  $\sigma$  of the Gaussian kernel can hinder a successful Kernel PCA denoising of NMR spectra. This is of no surprise as e.g. for extremely small  $\sigma$ 's the function

$$k_G(\mathbf{x}_t, \mathbf{x}_u) = \exp\left(-\frac{\|\mathbf{x}_t - \mathbf{x}_u\|^2}{2\sigma^2}\right), \quad t, u = 0, 1, \dots, T-1 \quad (4.84)$$

vanishes for  $t \neq u$  such that the resulting kernel matrix  $\mathbf{K}$  resembles the identity matrix. In contrast, a too large value of  $\sigma$  leads to a matrix  $\mathbf{K}$  whose elements consist of ones only. Obviously, the kernel matrix cannot represent the structure of the data in these cases such that a  $\sigma$  is needed which is far away from both extremes.

A good way to determine such a  $\sigma$  for NMR spectra was to require that the denominator in Eq. (4.84) should equal the average squared distance between the noisy input space vectors. In detail, this average was computed as follows: first, the entire 1D spectrum to be denoised was embedded into a  $D$  dimensional space whereas a delay of  $\Delta t$  was used. In this step the same values of  $D$  and  $\Delta t$  were used as later on when the 512 data points long subsegments of the spectrum underwent the denoising procedure. Let  $\hat{\mathbf{X}}^{tr} = [\hat{\mathbf{x}}_0^{tr}, \hat{\mathbf{x}}_1^{tr}, \dots, \hat{\mathbf{x}}_{T-1-D\Delta t}^{tr}]$  be the trajectory matrix obtained by embedding the entire 1D spectrum.  $\sigma$  was then set such that the average squared distance between the  $\hat{\mathbf{x}}_t^{tr}$  equalled  $2\sigma^2$ , i.e.

$$\sigma = \sqrt{\frac{1}{2} \frac{1}{T - D\Delta t} \sum_{t,u=0}^{T-1-D\Delta t} \|\hat{\mathbf{x}}_t^{tr} - \hat{\mathbf{x}}_u^{tr}\|^2} \quad (4.85)$$

The resulting value was kept fixed throughout the denoising of all four subsegments in which the entire spectrum was divided. Note, however, that a new  $\sigma$  was computed for each of the 1D spectra recorded in a 2D-NOESY experiment.

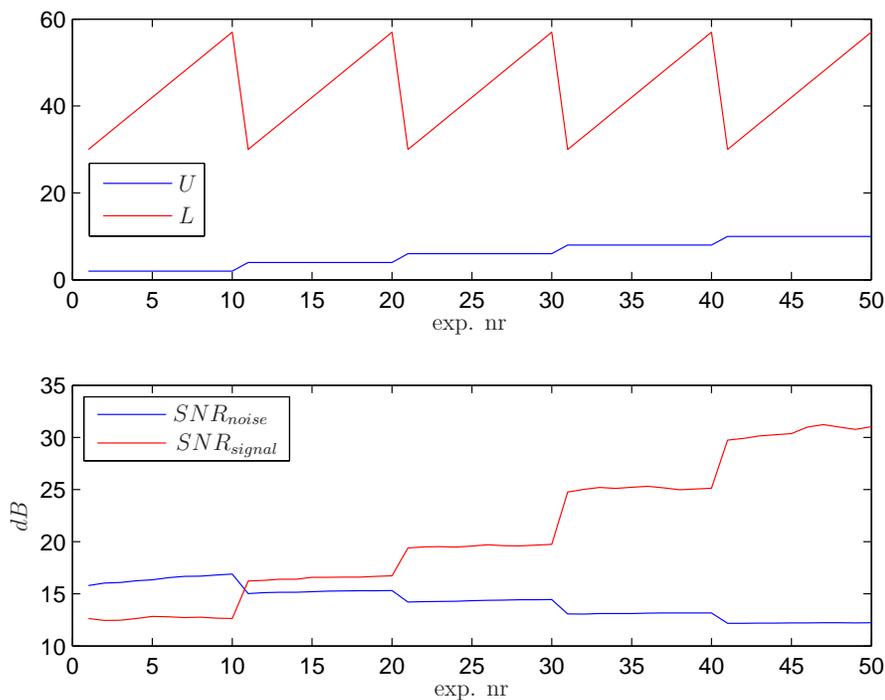


Figure 4.13: Determination of optimal  $U$  and  $L$ . Top: the parameter sets of  $U$  and  $L$  plotted over the experiment numbers. Bottom: the performance measures  $SNR_{signal}$  and  $SNR_{noise}$  plotted over the experiment numbers. The best compromise between a relatively low  $SNR_{noise}$  and a quite large  $SNR_{signal}$  were obtained in experiment 20 in which  $U = 4$  dominant eigenvectors as well as  $L = 57$  next neighbors were considered.

In this context it should be mentioned that Kernel PCA denoising procedure seemed to be quite robust against the particular choice of the embedding parameters  $D$  and  $\Delta t$ . For NMR spectra good results were obtained if these two parameters were set to  $D = 10$  and  $\Delta t = 1$ , respectively. However, increasing e.g.  $D$  to 20 lead to virtually identical results if all other parameters were kept fixed (e.g. the differences of the denoised spectra for  $D = 10$  and  $D = 20$  were in the range of  $10^4$ , a value which is negligible compared to the maximal peak amplitudes of about  $11 \cdot 10^6$ ).

In contrast, the number  $U$  of dominant eigenvectors considered during the projection step of the Kernel PCA procedure as well as the number  $L$  of next feature space neighbors used to compute the pre-image by means of metric MDS turned out to have a stronger influence on the outcome of the denoising

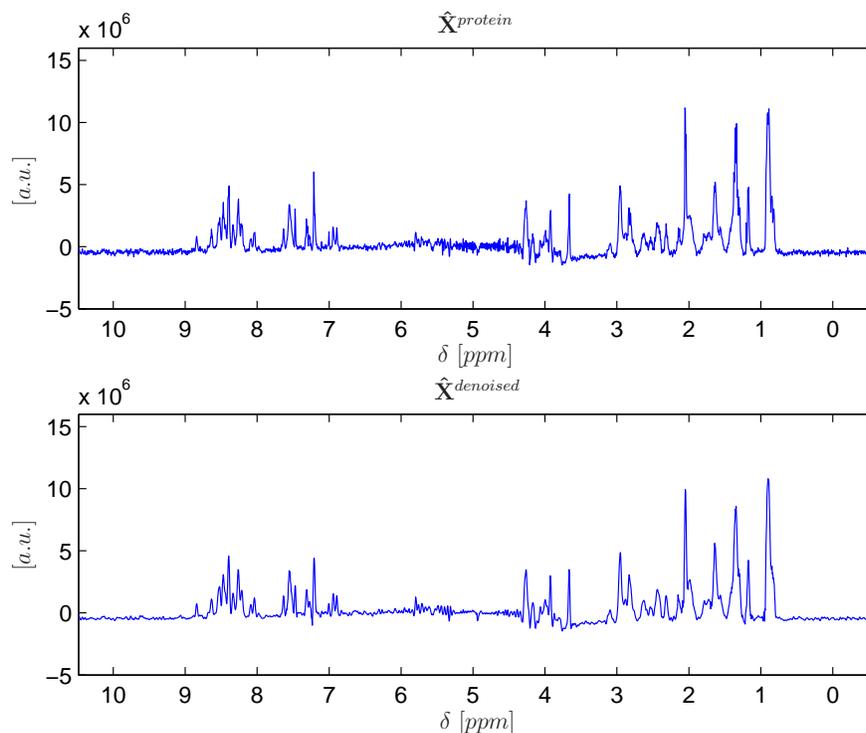


Figure 4.14: Kernel PCA denoising of the P11 spectrum recorded for the shortest  $t_1$ -time. Top: the noise water free spectrum obtained by means of MP-BSS. Bottom: the denoised spectrum. The noise was reduced noticeably but the peaks at about 7.3 ppm were distorted.

procedure. Unfortunately, optimal values of these two parameters could not be found beforehand such that they had to be determined by trial and error. For this purpose 50 experiments were carried out in which  $U$  was increased from 2 to 10 in steps of size two and  $L$  was increased from 30 to 57 in steps of size three. In order to quantify the results the the same performance measures  $SNR_{signal}$  (cf. Eq. (4.39)) and  $SNR_{noise}$  (cf. Eq. (4.42)) were used as in local PCA denoising. The results are summarized in Fig. 4.13. As expected the signal got less distorted if more dominant eigenvectors were considered, while at the same time, though, the denoising performance decreased. The number of next neighbors was of less importance but there seemed to be a minor trend that increasing  $L$  lead to better  $SNR_{noise}$  and  $SNR_{signal}$  values. Note, however, that generally the denoised spectra were distorted significantly if less than 15 next neighbors were used during the pre-image procedure.

Overall, the best compromise between both a high  $SNR_{noise} = 15$  dB

Parameter	Value
$D$	10
$\Delta t$	1
$U$	4
$L$	57
$\sigma$	$4.34 \cdot 10^7$

Table 4.2: The parameter set used in Kernel PCA denoising for which the best result as shown in the bottom of Fig. 4.14 was obtained.

and large  $SNR_{signal} = 17 \text{ dB}$  was obtained in experiment nr. 20 for which  $U$  was set to four and  $L$  to 57, respectively (see Tab. 4.2 for a complete list of parameters). The corresponding spectrum is depicted in Fig. 4.14. Obviously, the noise was reduced noticeably and the majority of peaks (apart those at about  $7.3 \text{ ppm}$ ) remained undistorted. It must be pointed out, though, that a much better result was obtained with the local PCA denoising procedure (see Fig. 4.8) with which a  $SNR_{noise}$  of  $24 \text{ dB}$  and a  $SNR_{signal}$  of  $37 \text{ dB}$  was achieved.

Using the same parameters as shown in Tab. 4.2 Kernel PCA combined with the iterative pre-image algorithm was applied to denoise the P11 spectra for comparison. In contrast to the experiment described in Sec. 4.3.3.2 the used value for  $\sigma$  was now more reasonable such that the noise was also partly removed from the protein peaks. Still, the pre-image algorithm seemed to have converged prematurely from time to time as approximately 10 % of the peaks did not change their appearance at all. This led to a better  $SNR_{signal}$  value of  $19 \text{ dB}$  in comparison with the MDS pre-image method, however, the overall noise level remained larger ( $SNR_{noise} = 14 \text{ dB}$ ). This confirms the conclusion drawn in Sec. 4.3.3.2 that the MDS-based pre-image method is preferable over its iterative counterpart in the context of NMR data.

## 4.4 Conclusions

In this chapter postprocessing algorithms were investigated which were capable of reducing the additional noise appearing in NMR spectra after the water signal has been removed. First the principles of classical PCA denoising were reviewed and applied to NMR data. However, PCA led to unsatisfying results as especially smaller peaks in the P11 spectra got distorted considerably or vanished entirely. Hence, an extension to PCA, called local PCA, was investigated in which the data were first embedded into a higher dimensional feature space and clustered by similarity before standard

PCA was carried out. A drawback of local PCA was the large number of parameters which needed to be tuned in order to obtain optimal results. Hence, procedures and performance measures were presented with which the optimal set of parameters could be determined automatically and within reasonable computational time.

For comparison also Kernel PCA was investigated. In this approach the data were mapped nonlinearly into a higher dimensional feature space whereas, however, explicit projections could be circumvented with the aid of appropriate kernel functions. After standard PCA was carried out the denoised feature space vectors had to be mapped back to input space whereas two different approaches, an iterative and an metric MDS based one were tried. It was observed that the iterative method lead to spurious results such that the metric MDS pre-image algorithm was recommended for NMR data. Using the same performance measures as in local PCA the optimal parameters appearing in the Kernel PCA and the MDS pre-image procedure could be determined automatically again.

In general, it turned out that local PCA lead to better results than Kernel PCA when it was applied to NMR data. In particular, local PCA better removed the noise from the spectra and hardly lead to any distortions of the protein peaks. Additionally, local PCA was computationally far less demanding as no pre-image problem needed to be solved.

Overall, local PCA well-preserved the protein peaks and reduced the additional noise in the protein spectra impressively. Hence, it is the method of choice for the postprocessing of spectra which were analyzed by MP-BSS before.

# Chapter 5

## BSS Combined Denoising and Automated Water Assignment

So far the water signal has been removed from NMR spectra by the following two step procedure: first, the recorded spectra were decomposed into protein and water related signals by BSS. The water related sources were then removed whereas the noise level in the resulting spectra increased considerably. In the second step of the proposed procedures this additional noise was removed again by variants of PCA denoising.

This approach is reversed in the chapter at hand, i.e. now the original spectra are denoised first before they are analyzed by BSS. For this purpose, the algorithm dAMUSE [56] is used which elegantly comprises both the denoising and the BSS step in a single algorithm. For the denoising part a similar embedding strategy is followed as in local PCA while the BSS step is analogous to the MP-BSS procedure.

Furthermore a new algorithm called Autoassign is presented in this chapter which automatizes the task of detecting water related signals in the sources obtained by BSS. With the aid of this algorithm also the optimal set of the parameters appearing in dAMUSE is determined. Technically, Autoassign estimates the water signal in the time domain by means of singular spectrum analysis (SSA) whereupon the water related sources are detected by an genetic algorithm (GA) based optimization procedure.

### 5.1 dAMUSE

In this section a variant of dAMUSE [56] is presented with which both BSS and denoising can be performed simultaneously. Similar to the well-known AMUSE algorithm [57] a covariance matrix and an autocovariance matrix are

jointly diagonalized in dAMUSE, whereas, however, the observed data are first embedded into a higher dimensional feature space by means of delayed coordinates. After the covariance and the autocovariance matrix have been computed in feature space their GEVD is determined by the two step EVD procedure presented in Sec. 1.3. In this process the dimension of the problem is reduced after the first EVD step in order to denoise the data. Eventually, the mixing matrix is estimated by the second EVD.

A drawback of dAMUSE is that it has a filtering indeterminacy in addition to the scaling and permutation indeterminacies inherent in the BSS problem. However, this shortcoming does not affect the water removal procedure as will be seen later on.

Note, that in the following a slight deviation from the classical dAMUSE algorithm is made as instead of an autocorrelation matrix a correlation matrix of filtered signals is used in analogy to the MP-BSS approach. Furthermore, the data are Fourier transformed after they have been embedded into feature space.

### 5.1.1 Embedding

Consider the linear mixture model (see also Eq. (3.6))

$$\mathbf{X} = \mathbf{A}\mathbf{S} \quad (5.1)$$

whereas  $\mathbf{X}$  and  $\mathbf{S}$  are  $M \times T$ -matrices and  $\mathbf{A}$  is of size  $M \times M$ . In order to increase the dimension of the problem a trajectory matrix  $\mathbf{X}_m^{tr}$  (see Eq. (4.2.1)),  $m = 1, 2, \dots, M$ , is formed for each of the  $M$  rows of  $\mathbf{X}$ . These matrices are of size  $D \times T - (D - 1)\Delta t$  whereas  $D$  denotes the embedding dimension and  $\Delta t$  is the shift between consecutive rows of  $\mathbf{X}_m^{tr}$ .

In contrast to local or Kernel PCA the  $\mathbf{X}_m^{tr}$ 's are not analyzed one by one but are used to constitute the overall  $MD \times T - (D - 1)\Delta t$  trajectory matrix

$$\mathbf{X}^{tr} = [\mathbf{X}_1^{tr}, \mathbf{X}_2^{tr}, \dots, \mathbf{X}_M^{tr}]^\top \quad (5.2)$$

in dAMUSE.

Following the same procedure the matrix

$$\mathbf{S}^{tr} = [\mathbf{S}_1^{tr}, \mathbf{S}_2^{tr}, \dots, \mathbf{S}_M^{tr}]^\top \quad (5.3)$$

is constituted whereas the  $\mathbf{S}_m^{tr}$ 's,  $m = 1, 2, \dots, M$ , denote the individual  $D \times T - (D - 1)\Delta t$  trajectory matrices formed for each of the  $M$  rows of  $\mathbf{S}$ .

Eventually, the  $MD \times MD$  matrix

$$\mathbf{A}^{tr} = \begin{bmatrix} a_{11}\mathbf{I}_D & a_{12}\mathbf{I}_D & \dots & a_{1M}\mathbf{I}_D \\ a_{21}\mathbf{I}_D & a_{22}\mathbf{I}_D & \dots & a_{2M}\mathbf{I}_D \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}\mathbf{I}_D & a_{M2}\mathbf{I}_D & \dots & a_{MM}\mathbf{I}_D \end{bmatrix} \quad (5.4)$$

is built. Here, the  $a_{ij}$ 's denote the elements of the mixing matrix  $\mathbf{A}$  while  $\mathbf{I}_D$  is the  $D \times D$  identity matrix.

With these matrices the linear mixture model can be expressed in embedding space (also referred to as feature space in the following) as in input space as

$$\mathbf{X}^{tr} = \mathbf{A}^{tr} \mathbf{S}^{tr}. \quad (5.5)$$

### 5.1.2 Estimation of $\mathbf{A}^{tr}$ and $\mathbf{S}^{tr}$

Essentially, the mixing matrix  $\mathbf{A}^{tr}$  and the source matrix  $\mathbf{S}^{tr}$  are estimated as in MP-BSS after the observed signals have been embedded in feature space. For this purpose, the rows of the matrix  $\mathbf{X}^{tr}$  are filtered and stored in the matrix  $\mathbf{X}_F^{tr}$  whereupon the  $MD \times MD$  covariance matrices

$$\begin{aligned} \mathbf{C}_{\mathbf{X}^{tr}} &= \frac{1}{P-1} (\mathbf{X}^{tr} - \mathbf{m}_{\mathbf{X}^{tr}} \mathbf{1}_P^\top) (\mathbf{X}^{tr} - \mathbf{m}_{\mathbf{X}^{tr}} \mathbf{1}_P^\top)^H \\ \mathbf{C}_{\mathbf{X}_F^{tr}} &= \frac{1}{P-1} (\mathbf{X}_F^{tr} - \mathbf{m}_{\mathbf{X}_F^{tr}} \mathbf{1}_P^\top) (\mathbf{X}_F^{tr} - \mathbf{m}_{\mathbf{X}_F^{tr}} \mathbf{1}_P^\top)^H, \end{aligned} \quad (5.6)$$

with  $P = T - (D-1)\Delta t$ , are computed.

Let  $\mathbf{S}_F^{tr}$  be the  $MD \times T - (D-1)\Delta t$ -matrix obtained by filtering the rows of  $\mathbf{S}^{tr}$  and denote by  $\mathbf{C}_{\mathbf{S}^{tr}}$  and  $\mathbf{C}_{\mathbf{S}_F^{tr}}$  the corresponding  $MD \times MD$  covariance matrices computed in analogy to Eq. (5.6). Following the derivations in Eq. (3.25) it can be shown readily that the matrix pencils  $(\mathbf{C}_{\mathbf{X}^{tr}}, \mathbf{C}_{\mathbf{X}_F^{tr}})$  and  $(\mathbf{C}_{\mathbf{S}^{tr}}, \mathbf{C}_{\mathbf{S}_F^{tr}})$  are congruent with congruence matrix  $\mathbf{A}^{tr}$ . Hence, the eigenvector matrices  $\mathbf{E}_{\mathbf{X}^{tr}}$  and  $\mathbf{E}_{\mathbf{S}^{tr}}$  of  $(\mathbf{C}_{\mathbf{X}^{tr}}, \mathbf{C}_{\mathbf{X}_F^{tr}})$  and  $(\mathbf{C}_{\mathbf{S}^{tr}}, \mathbf{C}_{\mathbf{S}_F^{tr}})$ , respectively, are related by

$$\mathbf{E}_{\mathbf{S}^{tr}} = (\mathbf{A}^{tr})^H \mathbf{E}_{\mathbf{X}^{tr}} \quad (5.7)$$

according to Theo. 4 (see Sec. 1.3).

Assuming that the sources (i.e. the rows of the matrix  $\mathbf{S}$ , cf. Eq. 5.1) are uncorrelated the matrices  $\mathbf{C}_{\mathbf{S}^{tr}}$  and  $\mathbf{C}_{\mathbf{S}_F^{tr}}$  are of block diagonal structure, i.e.

$$\mathbf{C}_{\mathbf{S}^{tr}} = \begin{bmatrix} \mathbf{C}_{\mathbf{S}^{tr},11} & 0 & \dots & 0 \\ 0 & \mathbf{C}_{\mathbf{S}^{tr},22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{C}_{\mathbf{S}^{tr},MM} \end{bmatrix} \quad (5.8)$$

and

$$\mathbf{C}_{\mathbf{S}_F^{tr}} = \begin{bmatrix} \mathbf{C}_{\mathbf{S}_F^{tr},11} & 0 & \dots & 0 \\ 0 & \mathbf{C}_{\mathbf{S}_F^{tr},22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{C}_{\mathbf{S}_F^{tr},MM} \end{bmatrix} \quad (5.9)$$

with block matrices

$$\mathbf{C}_{\mathbf{S}^{tr},mm} = \frac{1}{P-1} (\mathbf{S}_m^{tr} - \mathbf{m}_{\mathbf{S}_m^{tr}} \mathbf{1}_P^\top) (\mathbf{S}_m^{tr} - \mathbf{m}_{\mathbf{S}_m^{tr}} \mathbf{1}_P^\top)^H \quad (5.10)$$

respectively

$$\mathbf{C}_{\mathbf{S}_F^{tr},mm} = \frac{1}{P-1} (\mathbf{S}_{F,m}^{tr} - \mathbf{m}_{\mathbf{S}_{F,m}^{tr}} \mathbf{1}_P^\top) (\mathbf{S}_{F,m}^{tr} - \mathbf{m}_{\mathbf{S}_{F,m}^{tr}} \mathbf{1}_P^\top)^H \quad (5.11)$$

on their diagonal.

Accordingly, also the eigenvector matrix

$$\mathbf{E}_{\mathbf{S}^{tr}} = \begin{bmatrix} \mathbf{E}_{\mathbf{S}^{tr},11} & 0 & \dots & 0 \\ 0 & \mathbf{E}_{\mathbf{S}^{tr},22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{E}_{\mathbf{S}^{tr},MM} \end{bmatrix} \quad (5.12)$$

of the matrix pencil  $(\mathbf{C}_{\mathbf{S}^{tr}}, \mathbf{C}_{\mathbf{S}_F^{tr}})$  has block matrix structure whereas  $\mathbf{E}_{\mathbf{S}^{tr},mm}$  is the  $D \times D$  eigenvector matrix of the pencil  $(\mathbf{C}_{\mathbf{S}^{tr},mm}, \mathbf{C}_{\mathbf{S}_F^{tr},mm})$ .

Following the same approach as in MP-BSS the source matrix  $\mathbf{S}_{est}^{tr}$  is estimated in dAMUSE by multiplying the observation matrix  $\mathbf{X}^{tr}$  by the Hermitian transpose  $\mathbf{E}_{\mathbf{X}^{tr}}^H$  of the eigenvector matrix of the observation matrix pencil. This leads to

$$\mathbf{S}_{est}^{tr} = \mathbf{E}_{\mathbf{X}^{tr}}^H \mathbf{X} = \mathbf{E}_{\mathbf{X}^{tr}}^H \mathbf{A} \mathbf{S} \stackrel{\text{Eq. (5.7)}}{=} \mathbf{E}_{\mathbf{S}^{tr}}^H \mathbf{S}^{tr}. \quad (5.13)$$

In the sequel of this section it will be shown that the rows of  $\mathbf{S}_{est}^{tr}$  contain FIR-filtered versions of the original sources. In these derivations the superscript “*tr*” will be omitted for simplicity, i.e.  $\mathbf{S} \equiv \mathbf{S}^{tr}$ ,  $\mathbf{S}^{est} \equiv \mathbf{S}_{est}^{tr}$ ,  $\mathbf{E}_{\mathbf{S}} \equiv \mathbf{E}_{\mathbf{S}^{tr}}$  and  $\mathbf{E}_{ii} \equiv \mathbf{E}_{\mathbf{S}^{tr},ii}$  will be used.

To start with the matrix  $\mathbf{S}^{est}$  is divided into  $M$  blocks  $\mathbf{S}_m^{est}$

$$\mathbf{S}^{est} = \begin{bmatrix} \mathbf{S}_1^{est} \\ \mathbf{S}_2^{est} \\ \vdots \\ \mathbf{S}_M^{est} \end{bmatrix} \quad (5.14)$$

whereas the  $m$ -th submatrix is given by

$$\mathbf{S}_m^{est} = \begin{bmatrix} \mathbf{s}_{m,1}^{est}(0) & \mathbf{s}_{m,1}^{est}(1) & \dots & \mathbf{s}_{m,1}^{est}(T-1-(D-1)\Delta t) \\ \mathbf{s}_{m,2}^{est}(0) & \mathbf{s}_{m,2}^{est}(1) & \dots & \mathbf{s}_{m,2}^{est}(T-1-(D-1)\Delta t) \\ \vdots & & & \\ \mathbf{s}_{m,D}^{est}(0) & \mathbf{s}_{m,D}^{est}(1) & \dots & \mathbf{s}_{m,D}^{est}(T-1-(D-1)\Delta t) \end{bmatrix}. \quad (5.15)$$

The source matrix  $\mathbf{S}$  is of similar structure

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_M \end{bmatrix} \quad (5.16)$$

whereas the  $m$ -th block  $\mathbf{S}_m$  consists of the trajectory matrix of the  $m$ -th original source signal  $s_m(t)$ :

$$\mathbf{S}_m = \begin{bmatrix} s_m((D-1)\Delta t) & s_m((D-1)\Delta t + 1) & \dots & s_m(T-1) \\ s_m((D-2)\Delta t) & s_m((D-2)\Delta t + 1) & \dots & s_m(T-1-\Delta t) \\ \vdots & \vdots & & \vdots \\ s_m(0) & s_m(1) & \dots & s_m(T-1-(D-1)\Delta t) \end{bmatrix} \quad (5.17)$$

Given the block diagonal structure of  $\mathbf{E}_S$   $\mathbf{S}_m$  and  $\mathbf{S}^{est}$  are related by the following equation

$$\mathbf{S}_m^{est} = \mathbf{E}_{mm} \mathbf{S}_m \quad (5.18)$$

such that the  $n$ -th row of  $\mathbf{S}_m^{est}$  is determined by

$$\mathbf{s}_{m,n}^{est}(t) = \sum_{k=1}^D (\mathbf{E}_{mm})_{n,k} s_m(t + (D-k)\Delta t). \quad (5.19)$$

Eq. (5.19) defines a convolution operation between row  $m$  of  $\mathbf{E}_{mm}$  and source signal  $s_m$ . Thus, the rows of the matrix  $\mathbf{E}_{mm}$  can be interpreted as impulse responses of finite impulse response (FIR) filters and the rows of the matrix  $\mathbf{S}^{est}$  contain filtered versions of the underlying sources.

In summary, the algorithm dAMUSE yields uncorrelated component signals which are filtered versions of the underlying source signals. Hence similar to blind deconvolution methods in addition to scaling and permutation indeterminacies there appears a filtering indeterminacy here. Likewise, the original mixing matrix  $\mathbf{A}$  cannot be determined up to permutation and scaling indeterminacies only as  $\mathbf{E}_S$  is block diagonal in dAMUSE (cf. Eq. 5.13). Still, the matrix  $\mathbf{A}_{est}^{tr} := (\mathbf{E}_{\mathbf{X}^{tr}}^H)^{-1}$  is often referred to as the estimated mixing matrix in the context of dAMUSE.

### 5.1.3 Denoising

As pointed out in the last section dAMUSE essentially boils down to the GEVD of the observation matrix pencil  $(\mathbf{C}_{\mathbf{X}^{tr}}, \mathbf{C}_{\mathbf{X}_F^{tr}})$  in which the eigenvector matrix  $\mathbf{E}_{\mathbf{X}^{tr}}$  is computed. This GEVD is performed by the two step procedure summarized in Alg. 1 whereas, however, only the  $U$  dominant eigenvectors and corresponding eigenvalues are taken into account after the first EVD. In other words, this means that the noise is removed from the embedded data by PCA before the second EVD is carried out.

In detail, let

$$\mathbf{C}_{\mathbf{X}_F^{tr}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H \quad (5.20)$$

be the EVD of the matrix  $\mathbf{C}_{\mathbf{X}_F^{tr}}$  whereas  $\mathbf{V}$  denotes the  $MD \times MD$  eigenvector matrix and  $\mathbf{\Lambda}$  the corresponding  $MD \times MD$  eigenvalue matrix.

In order to denoise the data the  $MD \times U$  matrix  $\mathbf{V}_U$  and the  $U \times U$  diagonal matrix  $\mathbf{\Lambda}_U$  are formed whereas the columns of  $\mathbf{V}_U$  consist of the  $U$  dominant eigenvectors and the diagonal elements of  $\mathbf{\Lambda}_U$  are constituted by the  $U$  largest eigenvalues, respectively.

Considering that with these two matrices  $\mathbf{C}_{\mathbf{X}_F^{tr}}$  can be well-approximated according to

$$\mathbf{C}_{\mathbf{X}_F^{tr}} = \mathbf{V}_U \mathbf{\Lambda}_U^{1/2} \mathbf{\Lambda}_U^{1/2} \mathbf{V}_U^H \quad (5.21)$$

the GEVD procedure outlined in Alg. 1 can be continued. Accordingly, the  $U \times MD$  transformation matrix

$$\mathbf{W} = \mathbf{\Lambda}_U^{-1/2} \mathbf{V}_U^H \quad (5.22)$$

is formed and the EVD

$$\mathbf{W}\mathbf{C}_{\mathbf{X}^{tr}}\mathbf{W}^H\mathbf{E} = \mathbf{E}\mathbf{D} \quad (5.23)$$

is computed whereas  $\mathbf{E}$  denotes the eigenvector and  $\mathbf{D}$  the corresponding eigenvalue matrix.

Eventually, the  $MD \times U$  eigenvector matrix  $\mathbf{E}_{\mathbf{X}^{tr}}$  and the  $U \times U$  eigenvalue matrix  $\mathbf{D}_{\mathbf{X}^{tr}}$  of the matrix pencil  $(\mathbf{C}_{\mathbf{X}^{tr}}, \mathbf{C}_{\mathbf{X}_F^{tr}})$  are determined by

$$\mathbf{E}_{\mathbf{X}^{tr}} = \mathbf{V}_U \mathbf{\Lambda}_U^{-1/2} \mathbf{E} \quad (5.24)$$

and

$$\mathbf{D}_{\mathbf{X}^{tr}} = \mathbf{D} \quad (5.25)$$

respectively.

Hence, the estimated source matrix

$$\mathbf{S}_{est}^{tr} = \mathbf{E}_{\mathbf{X}^{tr}}^H \mathbf{X}^{tr} = \mathbf{V}_U \mathbf{\Lambda}_U^{-1/2} \mathbf{E} \mathbf{X}^{tr} \quad (5.26)$$

<p><b>Data:</b>  <b>X:</b> <math>M \times T</math> observation matrix  <b>Input:</b>  <i>D</i>: number of delays  <math>\Delta t</math>: delay  <i>U</i>: number of dominant eigenvectors considered  <b>Result:</b>  <math>\mathbf{A}_{est}^{tr}</math>: <math>MD \times U</math> estimated mixing matrix  <math>\mathbf{S}_{est}^{tr}</math>: <math>U \times T - 1 - (D - 1)\Delta t</math> estimated source matrix</p> <pre> 1 begin 2   Form trajectory matrix <math>\mathbf{X}_m^{tr}</math> for each row of <math>\mathbf{X}</math>; 3   <math>\mathbf{X}^{tr} := [\mathbf{X}_1^{tr}, \mathbf{X}_2^{tr}, \dots, \mathbf{X}_M^{tr}]^\top</math>; 4   Form <math>\mathbf{X}_F^{tr}</math> by filtering the rows of <math>\mathbf{X}^{tr}</math>; 5   Compute covariance matrices <math>\mathbf{C}_{\mathbf{X}^{tr}}</math>, <math>\mathbf{C}_{\mathbf{X}_F^{tr}}</math> of <math>\mathbf{X}^{tr}</math>, <math>\mathbf{X}_F^{tr}</math>; 6   EVD: <math>\mathbf{C}_{\mathbf{X}^{tr}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H</math>; 7   Consider only <i>U</i> dominant eigenvectors: <math>\mathbf{V} \rightsquigarrow \mathbf{V}_U</math>, <math>\mathbf{\Lambda} \rightsquigarrow \mathbf{\Lambda}_U</math>; 8   <math>\mathbf{W} := \mathbf{\Lambda}_U^{-1/2}\mathbf{V}_U^H</math>; 9   EVD: <math>\mathbf{W}\mathbf{C}_{\mathbf{X}^{tr}}\mathbf{W}^H\mathbf{E} = \mathbf{E}\mathbf{D}</math>; 10  <math>\mathbf{A}_{est}^{tr} = (\mathbf{V}_U^U\mathbf{\Lambda}_U^{-1/2}\mathbf{E})\#</math>; 11  <math>\mathbf{S}_{est}^{tr} = \mathbf{V}_U^U\mathbf{\Lambda}_U^{-1/2}\mathbf{E}\mathbf{X}^{tr}</math>; 12 end </pre>
---

**Algorithm 9:** The algorithm dAMUSE.

consists of  $U$  recovered signals only while the estimated mixing matrix

$$\mathbf{A}_{est}^{tr} = (\mathbf{V}_U\mathbf{\Lambda}_U^{-1/2}\mathbf{E})\# \quad (5.27)$$

where “#” denotes the Moore-Penrose pseudoinvers, is of size  $MD \times U$ .

Combined with the embedding procedure described in Sec. 5.1.1 this leads to the dAMUSE algorithm outlined in Alg. 9

#### 5.1.4 dAMUSE analysis of NMR spectra

In the context of data stemming from 2D-NOESY experiments the following application of dAMUSE has proven to be successful. Based on the  $M \times T$ -matrix  $\mathbf{X}$  whose rows contain the individual time domain signals recorded throughout the experiment (i.e. the FIDs recorded for different evolution times  $t_1$ ) the  $DM \times T - (D - 1)\Delta t$  trajectory matrix  $\mathbf{X}^{tr}$  is formed by embedding each row of  $\mathbf{X}$  into a  $D$  dimensional feature space whereas a delay

of  $\Delta t$  is used. This matrix is concatenated with the  $DM \times (D - 1)\Delta t$ -matrix  $\mathbf{Z}$  whose elements are all zero

$$\mathbf{X}^{tr} \leftarrow [\mathbf{X}^{tr}, \mathbf{Z}] \quad (5.28)$$

This zero-filling is actually not required for the dAMUSE analysis of the spectra but it will be of advantage for the automatic assignment of water related components presented in Sec. 5.2.

Next, the  $MD \times T$ -matrix  $\hat{\mathbf{X}}^{tr}$  is computed by Fourier transforming each row of  $\mathbf{X}^{tr}$

$$\hat{\mathbf{X}}^{tr} = \mathcal{F}(\mathbf{X}^{tr}) \quad (5.29)$$

Once this matrix is obtained the dAMUSE algorithm is continued as usual, i.e. the matrix  $\hat{\mathbf{X}}_F^{tr}$  is formed by filtering each row of  $\hat{\mathbf{X}}^{tr}$  in frequency domain by a Gaussian shaped filter whereupon the estimated mixing matrix  $\mathbf{A}_{est}^{tr}$  and source matrix  $\hat{\mathbf{S}}_{est}^{tr}$  are obtained by the GEVD of the covariance matrices  $\mathbf{C}_{\hat{\mathbf{X}}^{tr}}$  and  $\mathbf{C}_{\hat{\mathbf{X}}_F^{tr}}$  (cf. Alg. 9 lines 4 et sqq.).

## 5.2 Automated water assignment

The procedure used in dAMUSE to remove the water signal from NMR spectra is similar to the one outlined in Sec. 3.2. Once the mixing matrix has been estimated the contributions of the recovered sources to the undelayed spectrum recorded for the smallest  $t_1$  time are computed whereupon water related sources are nilled out deliberately. The water free spectra are then obtained by mixing the remaining sources by the estimated mixing matrix.

However, up to  $D$  times more estimated signals have to be analyzed in dAMUSE than in standard BSS algorithms (such as MP-BSS) which do not embed the data in a higher dimensional feature spaces. Thus, a far larger number of signals has to be analyzed manually in order to determine the sources belonging to the water signal. This task tends to be quite tedious such that the algorithm Autoassign will be presented in this section which performs the water assignment automatically.

With the aid of autoassign also the optimal set of the parameters appearing in dAMUSE will be determined later on whereas similar strategies as in local or kernel PCA denoising will be applied.

### 5.2.1 Autoassign

Essentially, Autoassign consists of two steps: in the first step the water signal appearing in the FID recorded for the shortest  $t_1$ <sup>1</sup> is estimated by SSA. This estimated water signal is  $t_2$ -Fourier transformed and represents the target signal for the optimization procedure carried out in the second step. In the latter, the contributions of the individual sources to the first spectrum are determined. Autoassign then tries to find the set of those contributions which, when summed up, best describe the target signal.

#### 5.2.1.1 SSA estimation of the water signal

In order to get a good estimate of the water signal with which the first FID is corrupted singular spectrum analysis (SSA) can be used advantageously. Generally, SSA is just another term for the PCA of the trajectory matrix of a single signal.

Accordingly, the water signal is estimated by embedding the first FID  $x_1(t)$  into a  $D_{H_2O}$  dimensional space using the concept of delayed coordinates, i.e. a  $D_{H_2O} \times T - (D - 1)\Delta t_{H_2O}$  trajectory matrix  $\mathbf{X}_1^{tr}$  is formed whereas  $T$  denotes the number of data points of which the FID consists and  $\Delta t_{H_2O}$  denotes the shift between consecutive rows of  $\mathbf{X}_1^{tr}$ .

This matrix is fed into the PCA algorithm Alg. 3, whereas only  $U_{H_2O} = 1$  dominant eigenvector is considered in the projection step. The resulting matrix, which will be denoted by  $\mathbf{X}_{H_2O}^{tr}$  henceforth, is a trajectory matrix containing rough estimates of the signals stored in  $\mathbf{X}_1^{tr}$  in its rows. From this matrix the estimated water signal  $x_{H_2O}(t)$  is obtained by inverting the embedding into delayed coordinates (this equals the last step carried out in local PCA, see Eq. (4.30) et sqq.).

$x_{H_2O}(t)$  contains a raw approximation of the original FID as only the most dominant eigenvector is considered in PCA. In particular,  $x_{H_2O}(t)$  should only consist of the dominant low frequency component which gives the FID its basic shape.

This shape is defined by the signal originating from the water protons as these have the highest concentration in the sample, precess at a frequency close to zero and usually have the largest  $T_2$ -time. Hence,  $x_{H_2O}(t)$  is supposed to be a good estimate of the pure water signal which would be observed if no proteins were present in the sample.

---

<sup>1</sup>In the sequel of the chapter the FID and the corresponding spectrum obtained for the shortest evolution time  $t_1$  will be referred to as the first FID and first spectrum, respectively.

**Data:**  
 $x_1(t)$ : first FID

**Input:**  
 $D_{H_2O}$ : number of delays  
 $(\Delta t_{H_2O}=1$ : delay)  
 $(U_{H_2O}=1$ : number of dominant eigenvectors considered)

**Result:**  
 $\hat{x}_{H_2O}(\omega)$ : estimated water resonance

**1 begin**  
**2** Form trajectory matrix  $\mathbf{X}_1^{tr}$  of  $x_1(t)$ ;  
**3** Determine  $\mathbf{X}_{H_2O}^{tr}$  by PCA (Alg. 3 with  $U = U_{H_2O} = 1$ ) of  $\mathbf{X}_1^{tr}$ ;  
**4** Form  $x_{H_2O}(t)$  by inverting embedding of  $\mathbf{X}_{H_2O}^{tr}$ ;  
**5** F.T.:  $\hat{x}_{H_2O}(\omega) = \mathcal{F}(x_{H_2O}(t))$ ;  
**6 end**

**Algorithm 10:** SSA based estimation of the water resonance. Note that  $\Delta t_{H_2O} = 1$  and  $U_{H_2O} = 1$  are kept fixed for the water estimation.

As a last step of the water estimation procedure (cf. Alg. 10) the  $t_2$ -Fourier transformed counterpart

$$\hat{x}_{H_2O}(\omega_2) = \mathcal{F}(x_{H_2O}(t)) \quad (5.30)$$

of  $x_{H_2O}$  is computed. Given that  $T$  samples of the first FID are available also the spectrum  $\hat{x}_{H_2O}(\omega)$  will consist of  $T$  data points. In the following it will be more convenient to interpret the signal  $\hat{x}_{H_2O}$  as an  $1 \times T$  dimensional row vector  $\hat{\mathbf{x}}_{H_2O}$  whose elements consist of the  $T$  samples of  $\hat{x}_{H_2O}$ .

### 5.2.1.2 Assignment of water related sources

The vector  $\hat{\mathbf{x}}_{H_2O}$  determined above contains the water resonance as appearing in the first spectrum. Hence, the contributions of the estimated sources to the first spectrum are needed for the water assignment. Taking into account that the first spectrum forms the  $D$ -th row of the matrix  $\hat{\mathbf{X}}^{tr}$  these contributions can be computed readily as

$$\hat{\mathbf{S}}^{1st} = \text{diag}(a_{D,1}^{tr}, a_{D,2}^{tr}, \dots, a_{D,U}^{tr}) \hat{\mathbf{S}}_{est}^{tr}. \quad (5.31)$$

Here,  $a_{D,i}^{tr}$  denotes the  $i$ -th element of the  $D$ -th row of  $\mathbf{A}_{est}^{tr}$  while the  $u$ -th row of the  $U \times T$ -matrix  $\hat{\mathbf{S}}^{1st}$  contains the contribution of the  $u$ -th estimated source (i.e. the  $u$ -th row of  $\hat{\mathbf{S}}_{est}^{tr}$ ) to the first spectrum (i.e. to the  $D$ -th row of  $\hat{\mathbf{X}}^{tr}$ ).

Parameter	Value
Population size $N_{ind}$	800
Individual size $U$	variable
Nr. of subpopulations $n_{sub}$	8
Individuals per subpopulation $N_{sub}$	100
Selective pressure $\mu$	1.5
Nr. of offsprings $N_{off}$ per subpopulation	100
Mutation probability $p_{mut}$	0.01
Nr. of reinserted offsprings $N_{rein}$ per subpopulation	20
Nr. of generations between migrations $T_{mig}$	100
Nr. of migrants per subpopulation $N_{migrate}$	20

Table 5.1: Summary of parameters of the GA used to minimize the target function  $F$  (cf. Eq. (5.33)).

In the following, it will be convenient to describe the assignment of the individual sources to the water signal by means of the binary decision vector  $\mathbf{d} = [d_1, d_2, \dots, d_U]^\top$ , whose  $u$ -th element indicates if the  $u$ -th estimated source belongs to the water signal ( $d_u = 1$ ) or not ( $d_u = 0$ ).

Assuming that all sources are assigned correctly the  $1 \times T$  row vector

$$\hat{\mathbf{x}}_{approx} = \mathbf{d}^\top \hat{\mathbf{S}}^{1st} \quad (5.32)$$

should equal the estimated water signal  $\hat{\mathbf{x}}_{H_2O}$ . Hence, the following optimization problem can be formulated for the assignment of the water related sources, or, in other words, for the determination of the optimal decision vector  $\mathbf{d}$ :

$$\text{minimize } F(\mathbf{d}) := \|\hat{\mathbf{x}}_{H_2O} - \hat{\mathbf{x}}_{approx}\|^2 = \|\hat{\mathbf{x}}_{H_2O} - \mathbf{d}^\top \hat{\mathbf{S}}^{1st}\|^2 \text{ w.r.t. } \mathbf{d}. \quad (5.33)$$

Apparently, the above optimization problem is discontinuous because of the discrete structure of  $\mathbf{d}$  and can thus not be solved by means of standard optimization procedures such as gradient descent. Hence, a GA (see App. C) is used for the minimization of Eq. (5.33).

### 5.2.1.3 GA based optimization

In the following the major characteristics of the GA used to optimize Eq. (5.33) will be described briefly. Readers who are not familiar with the technical terms appearing in the following are referred to App. C which offers a short introduction into the concept of GAs.

For a reliable minimization of the target function Eq. (5.33) a GA consisting of a population of  $N_{ind} = 800$  individuals is used. Each individual is made up of a  $1 \times U$  binary vector  $d^{i} = [d_1^{i}, d_2^{i}, \dots, d_U^{i}]^\top$ ,  $i = 1, 2, \dots, 800$ , whereas  $U$  denotes the number of estimated sources determined by dAMUSE. These individuals are initialized at random at the beginning of the GA optimization procedure and are divided into  $n_{sub} = 8$  subpopulations of size  $N_{sub} = 100$ .

For each of the 800 individuals the target function value  $F(\mathbf{d}^{i})$  is determined. The individuals in each subpopulation are ranked in the ascending order of their target function values before their fitness is determined by means of the fitness function

$$\mathfrak{F}_{rank}^{\{k\}} = 2 - \mu + 2(\mu - 1) \frac{r^{\{k,j\}} - 1}{N_{sub} - 1}, \quad k = 1, 2, \dots, 100, \quad j = 1, 2, \dots, 8 \quad (5.34)$$

whereas  $r^{\{k,j\}}$  denotes the rank of the  $k$ -th individual in the  $j$ -th subpopulation and  $\mu$  is the selective pressure which is set to 1.5.

Based on these fitness values  $N_{off} = 100$  parent individuals are selected in each subpopulation by stochastic universal sampling (SUS) whereupon 100 offsprings are formed by uniform cross-over. During mutation each bit of the newly created individuals is flipped with a probability  $p_{mut}$  of 1%.

$N_{rein} = 20$  of the resulting offsprings are selected according to their fitness by SUS and are reinserted into their corresponding subpopulation whereas the 20 least fit individuals are replaced.

After each  $T_{mig} = 100$  generations migration is carried out between the subpopulations whereas the complete net structure scheme is used. In this process  $N_{migrate} = 20$  emigrants are chosen by fitness based SUS in each subpopulation which replace the least fit individuals in their new host populations after migration.

Eventually, the GA is terminated if the best individual does not change after  $T_{mig} = 100$  generations.

Note that in the proposed GA only the size  $U$  of the individuals is adapted while all other parameters are kept fixed no matter how many sources are estimated by dAMUSE (see also Tab. 5.1). This may seem surprising at first glance as usually at least the number of individuals has to be increased with the number of parameters (i.e. the number  $U$  of elements of  $\mathbf{d}$ ) of the problem at hand. It turns out, however, that for the usual cases in which 128 spectra are available and in which the embedding dimension is kept at a reasonable level (i.e.  $D < 4$ ) the proposed GA always leads to the same result if it is run several times with different initializations of the individuals. Hence, it can be concluded that the GA converges to the global minimum of  $F(\mathbf{d})$  if the

proposed parameters are used. However, the GA needs more generations to converge if the number  $U$  of elements of  $\mathbf{d}$  increases.

Finally, note that for the simulations and experiments described below the ‘‘Genetic Algorithm Toolbox’’ [10] was used.

## 5.3 dAMUSE Analysis of P11 Spectra

In this section it will be shown how the water approximation determined by SSA may be used advantageously to find the optimal values of the embedding dimension  $D$ , the delay  $\Delta t$  and the width  $\sigma$  of the Gaussian filter which appear in dAMUSE. Furthermore, strategies similar to the ones developed in the context of local and Kernel PCA denoising will be presented with which a befitting number of dominant eigenvectors  $U$  considered after the first EVD in dAMUSE may be estimated.

### 5.3.1 Water Estimate

In order to obtain a good estimate of the water signal with which the first P11 spectrum was corrupted SSA was applied. In this procedure the first FID was embedded into a feature space of dimension  $D_{H_2O} = 5, 6, \dots, 20$  whereas a fixed shift of  $\Delta t_{H_2O} = 1$  was used. The 16 resulting trajectory matrices were fed one by one into the PCA algorithm and projected onto the most dominant eigenvector only. This lead to 16 different time domain estimates  $\mathbf{x}_{approx}^{\{i\}}$ ,  $i = 1, 2, \dots, 16$ , of the water signal (see Fig. 5.1). The Fourier transforms

$$\hat{\mathbf{x}}_{approx}^{\{i\}} = \mathcal{F}(\mathbf{x}_{approx}^{\{i\}}) \quad (5.35)$$

of these estimates were computed and subtracted from the first P11 spectrum. Eventually, the resulting spectra were analyzed by visual inspection.

The best result was achieved for a  $D_{H_2O} = 12$  dimensional feature space. In this case the protein peaks were well preserved while the water signal was reduced drastically (see Fig. 5.2). For smaller values of  $D$  peaks close to the water signal vanished while for larger values the water signal became more and more dominant. Note that the removal of the water signal by SSA did not lead to any increase in noise in the spectrum.

### 5.3.2 Determination of $D$ , $\Delta t$ , and $\sigma$

During Autoassign the sources are determined which best describe the estimated water signal  $\hat{\mathbf{x}}_{approx}$ . In this process the lowest values of the target function  $F$  are expected for cases in which dAMUSE perfectly separates the

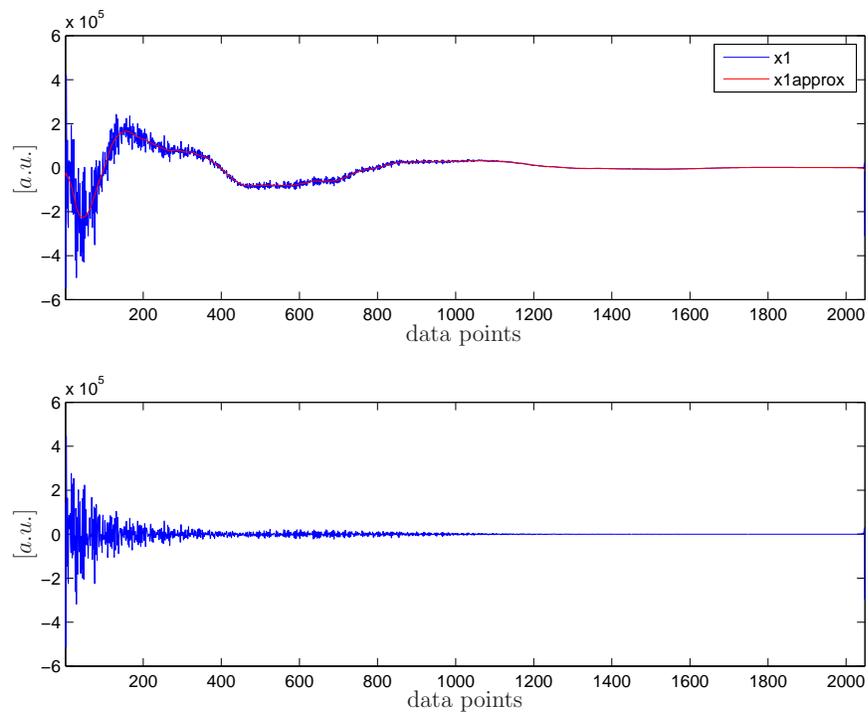


Figure 5.1: Water approximation by SSA. Top: the original first FID of P11 ( $x_1$ ) and its SSA approximation  $\mathbf{x}_{approx}$  as obtained for  $D = 12$ . Bottom: the P11-FID after the water signal was subtracted. Obviously the low frequency component which was responsible for the FID's basic shape was removed.

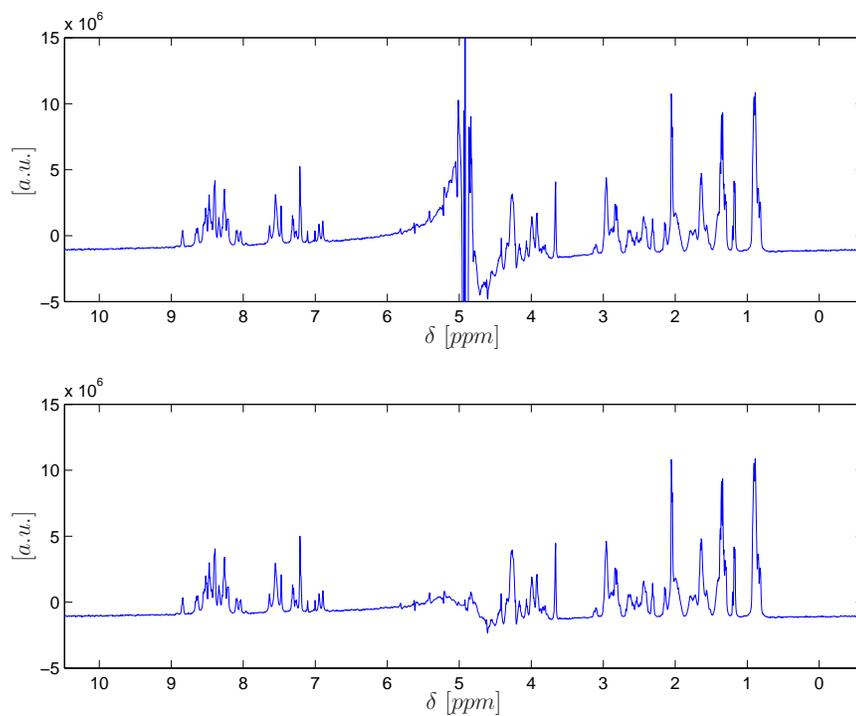


Figure 5.2: Water signal removal by SSA. Top: the original P11 spectrum. Bottom: the same spectrum after the water signal estimated by SSA was removed. For the SSA procedure the water signal was embedded into a  $D = 12$  dimensional space.

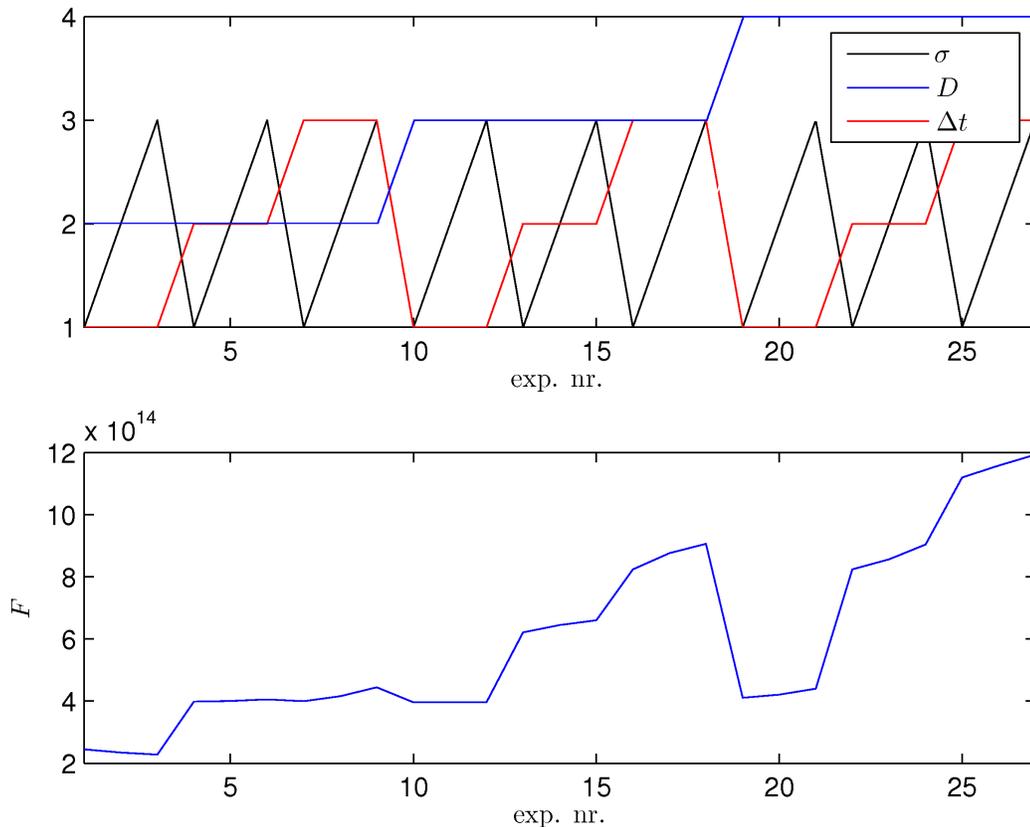


Figure 5.3: Quantitative comparison between results obtained by dAMUSE for various values of the parameters  $\sigma$ ,  $D$ , and  $\Delta t$ . Top: the parameters used in the individual experiments. Bottom: the resulting target function values  $F$  obtained after convergence of Autoassign.

water and the protein signals. Hence, the results obtained by dAMUSE for various sets of the parameters  $D$ ,  $\Delta t$  and  $\sigma$  can be compared quantitatively by means of the corresponding target function values of  $F$ .

Following this approach dAMUSE was run 27 times whereas the parameters  $D$ ,  $\Delta t$ , and  $\sigma$  were varied as shown in Tab. 5.2 (see also top of Fig. 5.3) and whereas always all available eigenvectors were used after the first EVD. This lead to 27 estimated mixing matrices  $\mathbf{A}_{est,i}^{tr}$  and source matrices  $\mathbf{S}_{est,i}^{tr}$ ,  $i = 1, 2, \dots, 27$ , respectively, which were used to constitute 27 contribution matrices  $\hat{\mathbf{S}}_i^{1st}$  according to Eq. (5.31). These matrices were fed together with the estimated water signal  $\hat{\mathbf{x}}_{approx}$  (as determined in the preceding section) into Autoassign. The resulting target function values obtained after the convergence of Autoassign are shown in the bottom of Fig. 5.3. Note that for each data set Autoassign was run 10 times whereas, however, always the

Parameter	Start value	Increment	End value
$D$	2	1	4
$\Delta t$	1	1	3
$\sigma$	1	1	3

Table 5.2: Values of the parameters  $D$ ,  $\Delta t$ , and  $\sigma$  used during the investigation of dAMUSE.

same sources were assigned to the water signal. This indicates that the used GA in fact converged to the global minimum of the target function  $F$ .

As can be seen in Fig. 5.3 best results were obtained for small values  $D = 2$  and  $\Delta t = 1$  of the embedding dimension and the shift, respectively, whereas the absolute minimum was reached for  $\sigma = 3$ . It seemed as if  $F$  decreased for increasing  $\sigma$  if  $D = 1$  and  $\Delta t = 1$  were used. Hence,  $\sigma$  was also set to 4, 5 and 6 additionally, however,  $F$  slightly increased again for these values (data not shown). Fig. 5.4 shows the spectrum corresponding to the lowest target function value which will be referred to as  $\hat{x}_{dAMUSE}^{noisy}$  henceforth. Clearly, the water signal got reduced significantly while the noise level increased as in MP-BSS.

In summary, this means that dAMUSE satisfactorily removed the water signal from the P11 spectra if  $D = 1$ ,  $\Delta t = 1$  and  $\sigma = 3$  were used. Note, that up to now all eigenvectors were taken into account after the first EVD such that the noise in the spectrum increased as in MP-BSS.

### 5.3.3 Estimation of $U$

For the determination of the optimal number  $U_{opt}$  of dominant eigenvectors considered after the first EVD in dAMUSE an equivalent approach as in local PCA (cf. Sec. 4.2.2.1) is used. First, the variance of the noise is estimated by computing the variance  $var_{baseline}$  of the peak-free sections at the left and right end of the noisy spectrum  $\hat{x}_{dAMUSE}^{noisy}$  (i.e. the spectrum obtained by dAMUSE if all available eigenvectors are considered after the first EVD (see preceding section)).

Next, dAMUSE is rerun, however, this time only a limited number  $U_{opt}$  of dominant eigenvectors is taken into account. Denoting by  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{MD}$  the eigenvalues of the correlation matrix  $\mathbf{C}_{\hat{\mathbf{x}}_F^{tr}}$   $U_{opt}$  is determined in analogy to Eq. (4.33) by

$$U_{opt} = \max \left\{ U \in \mathbb{N} \left| \sum_{d=U+1}^{MD} \lambda_d \leq MD var_{baseline} \right. \right\} \quad (5.36)$$

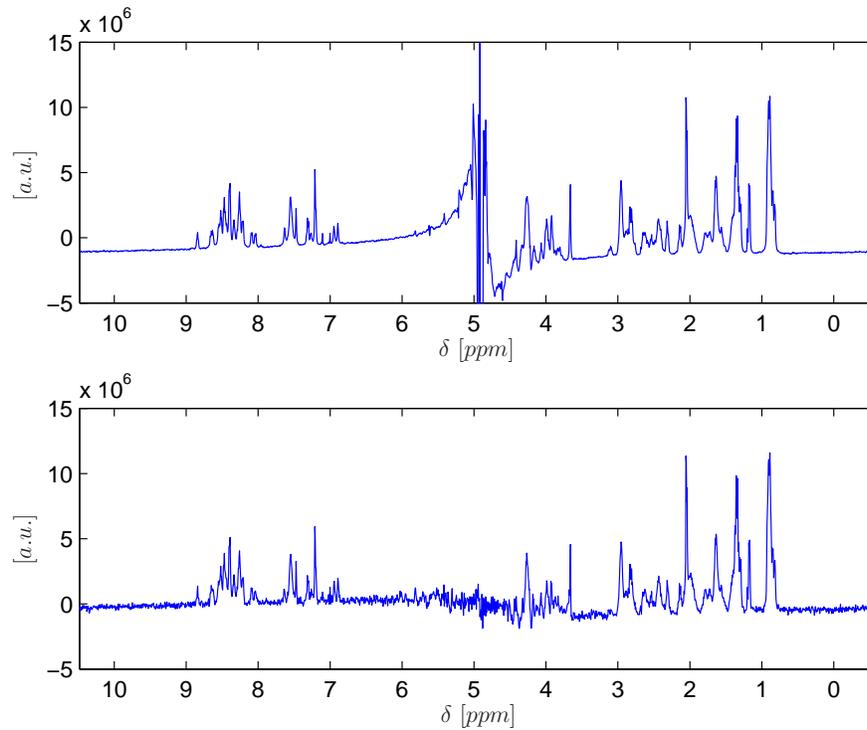


Figure 5.4: dAMUSE analysis of P11 without denoising. Top: the original first P11 spectrum. Bottom: the P11 spectrum after the water signal was removed by dAMUSE and Autoassign (called  $\hat{x}_{dAMUSE}^{noisy}$  in the text). In dAMUSE  $D = 2$ ,  $\Delta t = 1$  and  $\sigma = 3$  were used while all  $U = 128$  eigenvectors were considered after the first EVD.

i.e. the fraction of variance is removed from the data which could originate from the noise.

This procedure was applied to the P11 spectra whereas a value  $var_{noise} = 2.4 \cdot 10^{10}$  was estimated. It turned out, however, that this value was too large such that in the resulting spectra smaller peaks got removed together with the noise. Hence, a ten times smaller value had to be used for  $var_{noise}$  such that eventually  $U_{opt} = 111$  dominant eigenvalues were considered after the first EVD.

The corresponding spectrum as obtained from dAMUSE is depicted in the bottom of Fig. 5.5. As can be seen the water signal got removed and the protein peaks were hardly affected. Using the same procedure as in local and Kernel PCA to estimate the remaining noise in the spectrum a value of  $SNR_{noise} = 20$  dB was computed. Furthermore, the changes in the protein peaks were quantized by means of the performance measure  $SNR_{signal}$  whereas the noisy signal  $\hat{x}_{dAMUSE}^{noise}$  was taken as reference. This lead to a  $SNR_{signal}(\hat{x}_{dAMUSE}^{noise}, \hat{x}_{dAMUSE}^{denoised})$ -value of 21 dB whereas  $\hat{x}_{dAMUSE}^{denoised}$  denotes the denoised spectrum shown in the bottom of Fig. 5.5. Note, that this  $SNR_{signal}$  value cannot be compared with the corresponding values determined in local and Kernel PCA as these were based on another reference spectrum (i.e. the noisy spectrum obtained from MP-BSS). A visual comparison between the spectrum obtained by MP-BSS in combination with local PCA and the spectrum  $\hat{x}_{dAMUSE}^{noise}$  reveals that in both cases the water signal could be removed while the increase in noise in the spectra was moderate. However, in the middle of  $\hat{\mathbf{X}}_{dAMUSE}^{noise}$  some interferences remained at the location where the water signal resided before.

Finally, it must be pointed out that Autoassign was run 20 times in order to analyze the sources estimated by dAMUSE. In each of these runs the same sources were assigned to the water signal indicating that the GA with which the target function  $F$  was minimized had in fact converged to a global minimum.

## 5.4 Conclusions

In this chapter dAMUSE was presented with which both BSS and denoising can be carried out simultaneously. In this algorithm the observations are embedded into a higher dimensional feature space by means of delayed coordinates such that a comparatively large number of estimated sources is obtained. In order to automatically detect the water related signals among those sources the algorithm Autoassign was presented. This algorithm is based on a SSA estimation of the time domain water signal and on a GA for

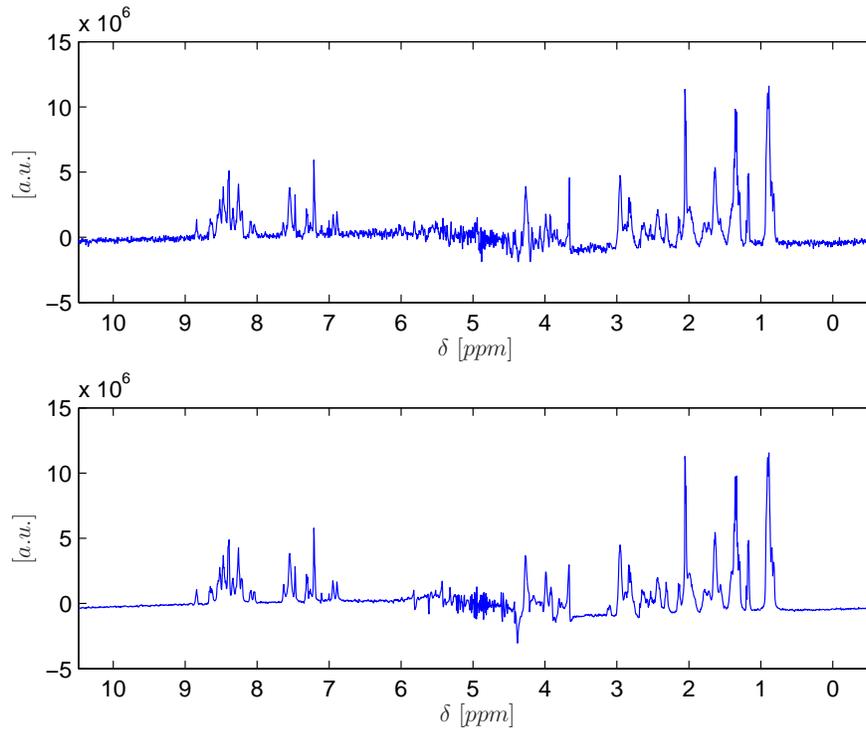


Figure 5.5: dAMUSE analysis of P11 whereas the dimension of the problem was reduced after the first EVD. Top: the noisy spectrum  $\hat{x}_{dAMUSE}^{noisy}$  which was obtained by dAMUSE whereas  $D = 2$ ,  $\Delta t = 1$ ,  $\sigma = 3$  were used and all eigenvectors were considered after the first EVD. Water related sources were identified and removed by Autoassign. Bottom: the water free and denoised P11 spectrum. The same values of  $D$ ,  $\Delta t$  and  $\sigma$  were used as above, however, only 111 dominant eigenvectors were taken into account after the first EVD. The noise got removed considerably and the peaks were well-preserved.

the assignment task. Usually, one drawback of GAs is the large number of parameters which need to be tuned, though, a fixed parameter set could be found for Autoassign.

Based on the SSA water estimation the optimal values for the parameters  $D$ ,  $\Delta t$  and  $\sigma$  could be found. Furthermore, the number of dominant eigenvectors to be considered after the first EVD in dAMUSE could be estimated at least approximately. For these optimal parameters dAMUSE removed the water signal satisfactorily while the increase in noise in the resulting spectra was rather marginal.



# Chapter 6

## Water Removal by SSA

So far linear BSS was applied to remove the water signal from NMR spectra. The motivation for this approach was on the one hand that the data recorded in 2D-NOESY experiments fit theoretically into the linear mixture model. On the other hand at least MP-BSS leads to results which are unique up to minor scaling and permutation indeterminacies. Hence, MP-BSS seems to be preferable over PCA based approaches with which e.g. the mixing matrix  $\mathbf{A}$  can only be recovered up to a unitary matrix in addition to the indeterminacies inherent in the BSS model.

However, it was also shown in Sec. 3.4.1 that MP-BSS fails to separate protein and water signals which overlap significantly as in such cases the uncorrelatedness assumption of the underlying sources is violated. Furthermore, the water-free spectrum obtained by subtracting the SSA estimate of the water resonance (see Fig. 5.2) only showed negligible remnants of the water signal and did not suffer from increased noise. Hence, the removal of the water signal by SSA will be investigated in this section and the obtained results will be compared with those achieved by MP-BSS combined with local and Kernel PCA, respectively, and by dAMUSE.

### 6.1 SSA Compared to MP-BSS and dAMUSE

In this section the results obtained by MP-BSS in combination with local and Kernel PCA, respectively, dAMUSE and SSA are compared considering as example the P11 data set. As was already shown in previous sections all of these approaches were capable of removing the water signal while keeping most of the protein peaks unaltered.

A closer inspection of the water free spectra revealed, however, that SSA best preserved the protein peaks residing in the immediate vicinity of the

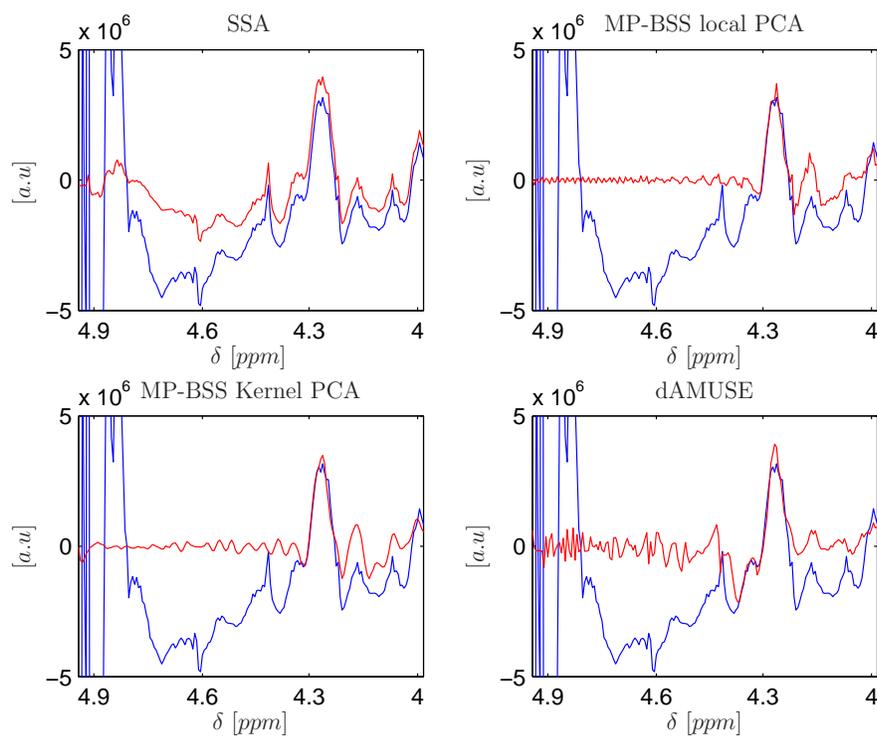


Figure 6.1: Close-up of the first P11 spectrum after the water signal was removed by SSA (upper left), MP-BSS + local PCA (upper right), MP-BSS + Kernel PCA (lower left), dAMUSE (lower right). Shown is the region right of the water resonance. The red lines represent the spectrum after the removal of the water signal while the blue lines show the original spectrum for comparison. Obviously, the peaks in the immediate vicinity of the water signal were best preserved by SSA. In particular, all of the other methods totally removed the protein peak at about 4.45 ppm.

water resonance. This region of the water free spectra as obtained by the various approaches is shown in Fig. 6.1. As can be seen SSA reproduced the protein peaks well for  $\delta < 4.6$  ppm but lead to a flatter baseline in comparison with the original spectrum. Thus the peaks in the region  $4$  ppm  $< \delta < 4.6$  ppm should be easier to analyze after the water signal estimated by SSA (cf. Alg. 10, Sec. 5.2.1.1) has been subtracted.

In comparison, MP-BSS combined with local PCA denoising could not recover any protein signals left of the relatively large protein peak at 4.3 ppm. Furthermore, also the smaller protein peak at about 4.05 ppm got distorted significantly.

This peak was better recovered if the spectra obtained by MP-BSS were denoised by Kernel PCA. However, also in this case no protein peaks were left in the region  $4.3$  ppm  $< \delta < 4.6$  ppm. dAMUSE lead to a similar result, though, the peak at about 4.05 ppm got distorted again.

In summary, these observations indicate that SSA is better suited to remove the water signal from NMR spectra than the BSS based approaches. This may seem surprising as SSA is based on a single EVD while MP-BSS and dAMUSE jointly diagonalize two correlation matrices. However, MP-BSS and dAMUSE are based on the assumption that the underlying sources are uncorrelated, a constraint which is violated if the water and the protein peaks overlap.

Apparently, the BSS based approaches are more sensitive to such overlapping signals than the simulations presented in Sec. 3.4.1 suggested. In particular it cannot be expected that signals which are totally hidden by the water resonance may be recovered by means of BSS in real life experiments albeit this was possible when the water signal was removed from the artificially created TmCSP spectra (see Sec. 3.5.1). This difference between real life and artificial data is attributed to additional violations of the linear mixture model (see Sec. 3.4.2) which exacerbate a clear separation of water and protein peaks.

Moreover, SSA has the advantage that the spectra need not be denoised after the water signal has been removed. Accordingly, only one parameter, namely the embedding dimension  $D_{H_2O}$ , needs to be tuned in order to obtain optimal results by SSA as the delay  $\Delta t_{H_2O} = 1$  and the number  $U_{H_2O} = 1$  of dominant eigenvectors considered can be kept fixed in Alg. 10. Note that in the following the subscript “ $H_2O$ ” will be dropped for simplicity.

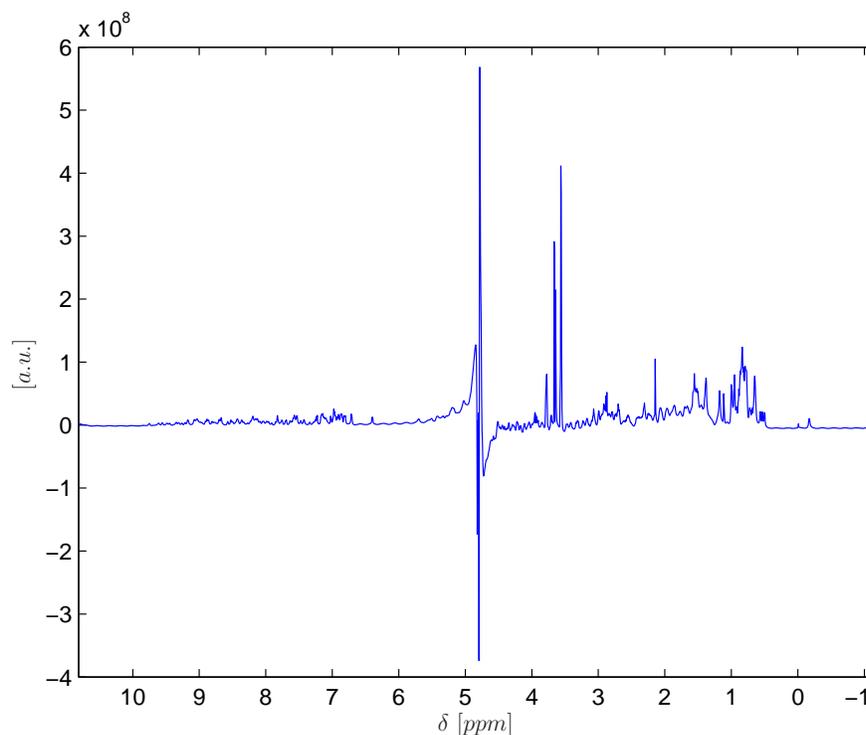


Figure 6.2: The first spectrum of the RAS-binding domain of the protein RalGEF.

## 6.2 SSA applied to RalGEF data

In order to investigate its robustness the SSA approach was applied to remove the water signal from further NMR data sets. To start with the spectrum of the RAS-binding domain of the protein RalGEF was analyzed. This domain consists of 87 amino acids and is located at the C-terminus of RalGEF. For further information about RalGEF and the function of its RAS-binding domain the interested reader is referred to [18].

As can be seen in Fig. 6.2 also the spectrum of the RAS-binding domain was corrupted by a dominant water signal which distorted the baseline significantly. This water signal was approximated in time domain by SSA whereas an embedding dimension of  $D = 30$  was used. Subtracting the estimated signal from the recorded FID followed by a  $t_2$ -Fourier transform lead to the water free spectrum shown in the bottom of Fig. 6.3. As can be seen the protein peaks were not affected by the SSA procedure while the water signal vanished almost completely.

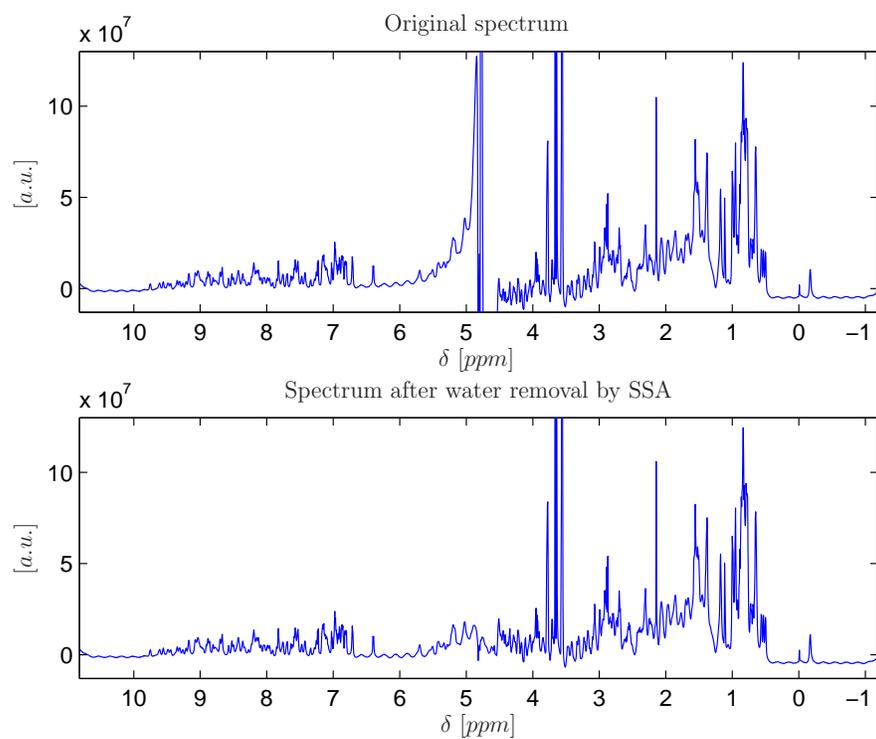


Figure 6.3: Comparison of the spectrum of the RAS-binding domain before and after the water signal was removed. Top: close-up of the original spectrum. Bottom: the same spectrum after the water signal was removed.

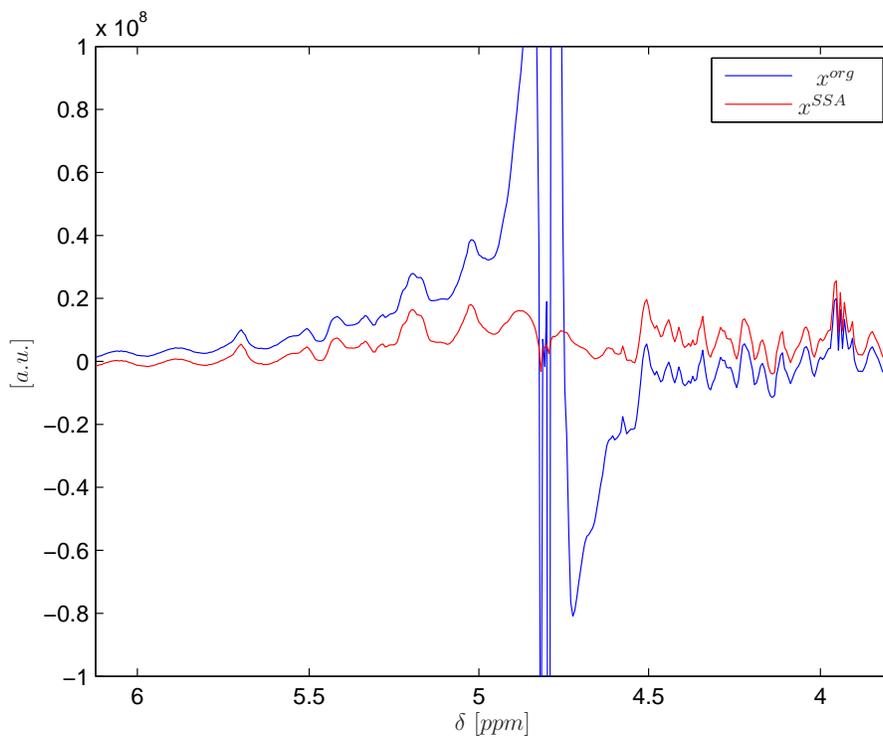


Figure 6.4: Close-up of the RAS-binding domain spectrum before ( $x^{org}$ ) and after ( $x^{SSA}$ ) the water signal was removed. Even peaks overlapping significantly with the water signal (e.g. at about  $\delta = 4.6$  ppm or  $\delta = 5$  ppm) remained in the spectrum.

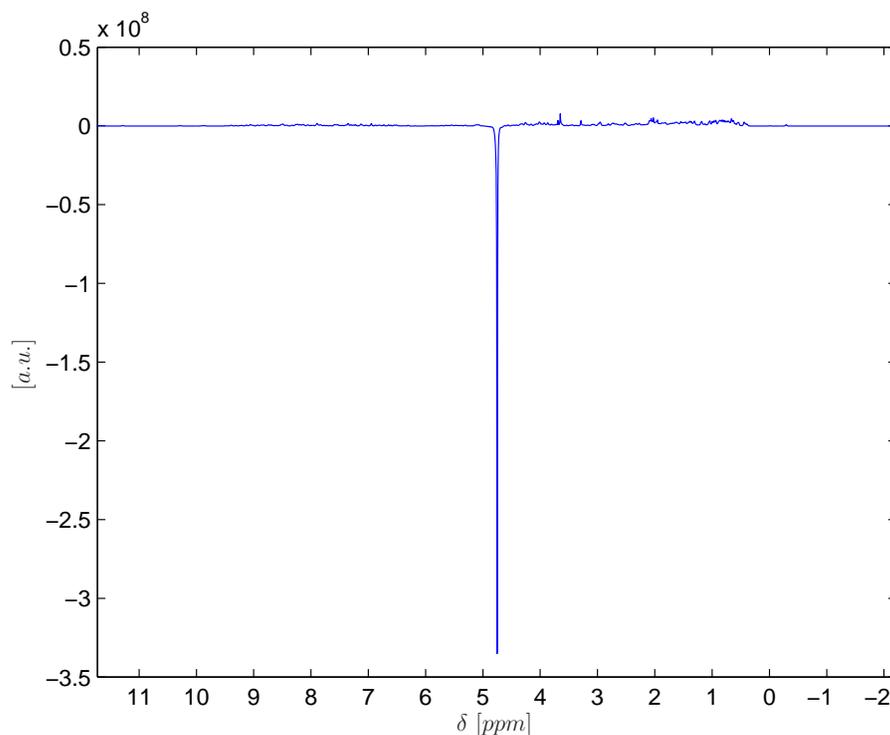


Figure 6.5: The spectrum first spectrum of HPr.

Furthermore, a closer inspection of the region formerly occupied by the water signal showed that even protein peaks which overlapped noticeably with the water resonance before remained in the spectrum (see Fig. 6.4). Thus, it may be concluded that SSA successfully removed the water signal from the RAS-binding domain spectrum.

### 6.3 SSA applied to HPr data

Similarly good results as in the case of P11 and the RAS-binding domain were achieved when the water signal was removed by SSA from spectra recorded during an investigation of the *Histidine containing phospho-carrier protein* (HPr). As can be seen in Fig. 6.5 the water signal was hardly suppressed experimentally such that the corresponding peak was at least 35 times larger than the largest protein peak. Still, the water signal could be estimated well in time domain whereas an embedding dimension of  $D = 60$  was used. The obtained estimate was subtracted from the original FID and the resulting

time domain signal was  $t_2$ -Fourier transformed.

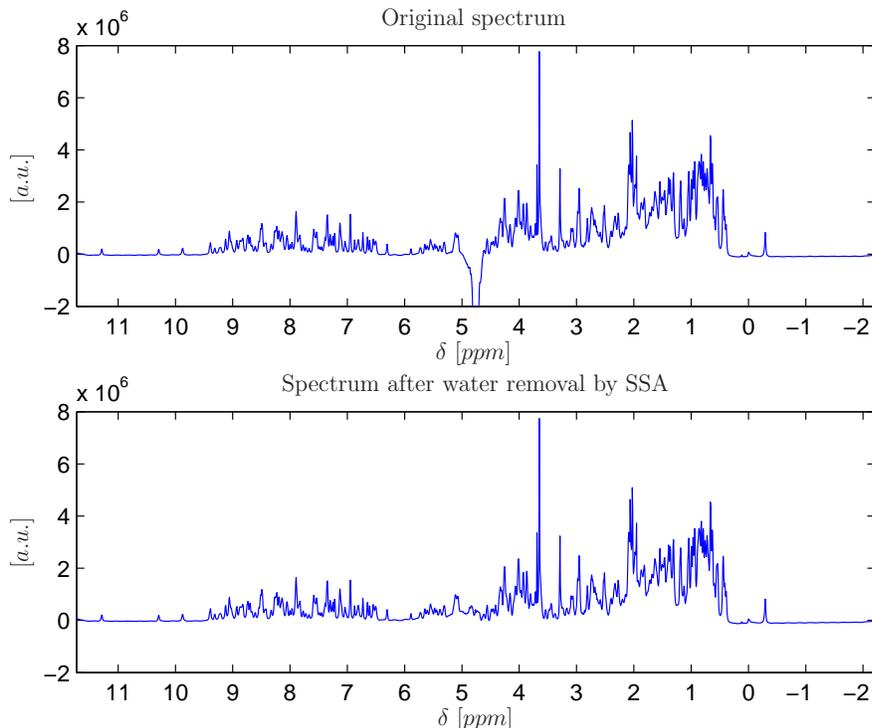


Figure 6.6: Comparison of the HPr spectrum before (top) and after (bottom) the water signal was removed by SSA. The water resonance disappeared almost completely while the protein peaks did not get distorted.

As can be seen in Fig. 6.6 and Fig. 6.7 the water resonance could be removed again while even protein peaks overlapping with it remained in the spectrum. Thus, SSA water removal approach was also applied successfully to HPr data.

## 6.4 SSA applied to TmCSP data

Eventually, SSA was also applied to real life spectra of the protein TmCSP. Note that the back-computed spectra of this protein were already used in Sec. 3.5.1 to evaluate the capacities of MP-BSS. Fig. 6.8 shows the recorded spectrum which was corrupted again by a dominant water peak. As before SSA was used to remove this peak by approximating the water signal in time domain whereas the embedding dimension  $D$  was set to 30. The obtained

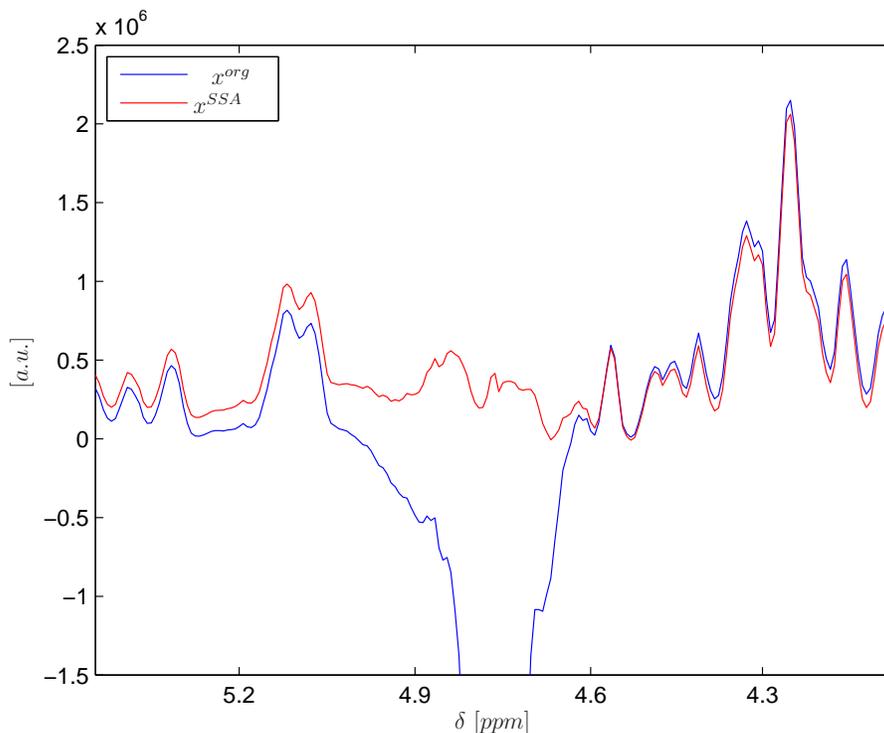


Figure 6.7: Close-up of the segment of the HPr-spectrum in which the water resonance resided.  $x^{org}$  denotes the recorded spectrum whereas  $x^{SSA}$  denotes the spectrum after the water signal was removed with the aid of SSA. Even protein peaks overlapping with the water signal (e.g. at about  $\delta = 4.6$  ppm) remained in the spectrum.

water estimate was subtracted from the recorded FID and the resulting signal was  $t_2$ -Fourier transformed. Fig. 6.9 and Fig. 6.10 show the result.

Again, the water signal was widely removed and protein peaks overlapping the water signal were well preserved.

Eventually, also the water free 2D-NOESY spectrum of TmCSP was computed. For this purpose the water peak got removed from each of the 512 spectra recorded in the experiment. This means that the water signal was estimated by SSA for each of the 512 corresponding FIDs separately whereas always the same embedding dimension  $D = 30$  was used. Subtracting the water estimates from the FIDs followed by a full  $t_2$ - and  $t_1$ -Fourier transforms lead to the 2D-NOESY spectrum shown in the right of Fig. 6.11.

In the left of Fig. 6.11 the original 2D-spectrum is depicted for comparison. In it a dominant cross like structure with bars at  $\omega_1, \omega_2 \approx 4.8$  ppm

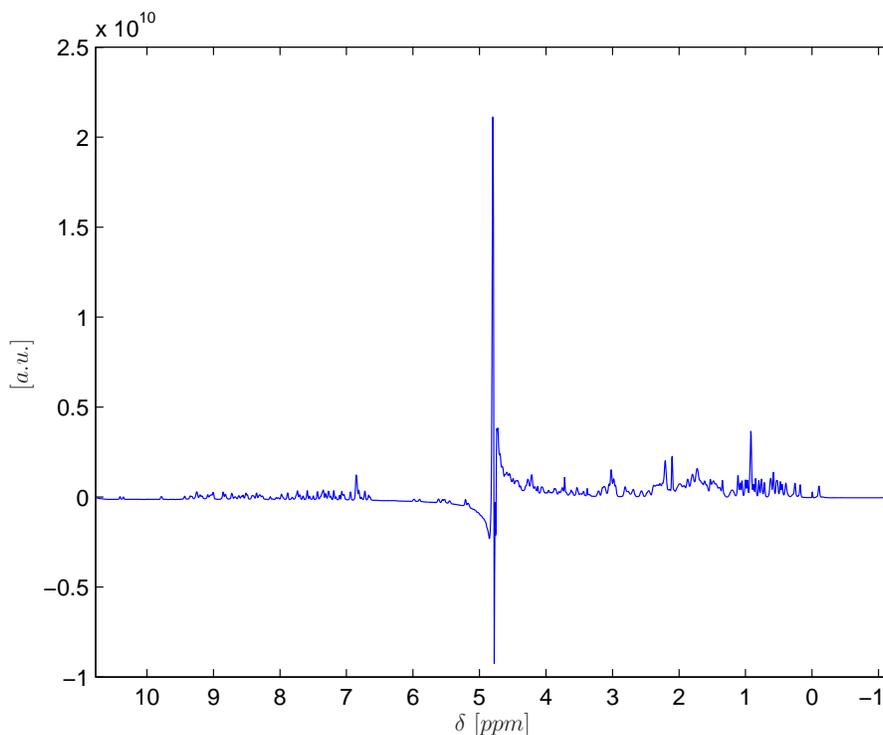


Figure 6.8: The measured TmCSP spectrum.

appeared which originated from the water protons. After the water signal was removed (see right of Fig. 6.11) the vertical bar disappeared entirely and only small remnants of the horizontal bar remained.

The close-up of the region  $8.00 \text{ ppm} \leq \omega_1 \leq 9.50 \text{ ppm}$ ,  $3.85 \text{ ppm} \leq \omega_2 \leq 5.00 \text{ ppm}$  which is depicted in Fig. 6.12 further illustrates the full capacity of the SSA approach. In the left of Fig. 6.12 the close-up of the original spectrum is shown in which some of the protein peaks (indicated by the black arrows) can hardly be identified because of the water signal. These peaks became clearly visible after the water signal was removed (see right of Fig. 6.12) by SSA.

## 6.5 Conclusions

This section dealt with the removal of the water resonance from NMR spectra by means of SSA. Compared with the methods presented in previous chapters it turned out that SSA lead to better results as it preserved more of the

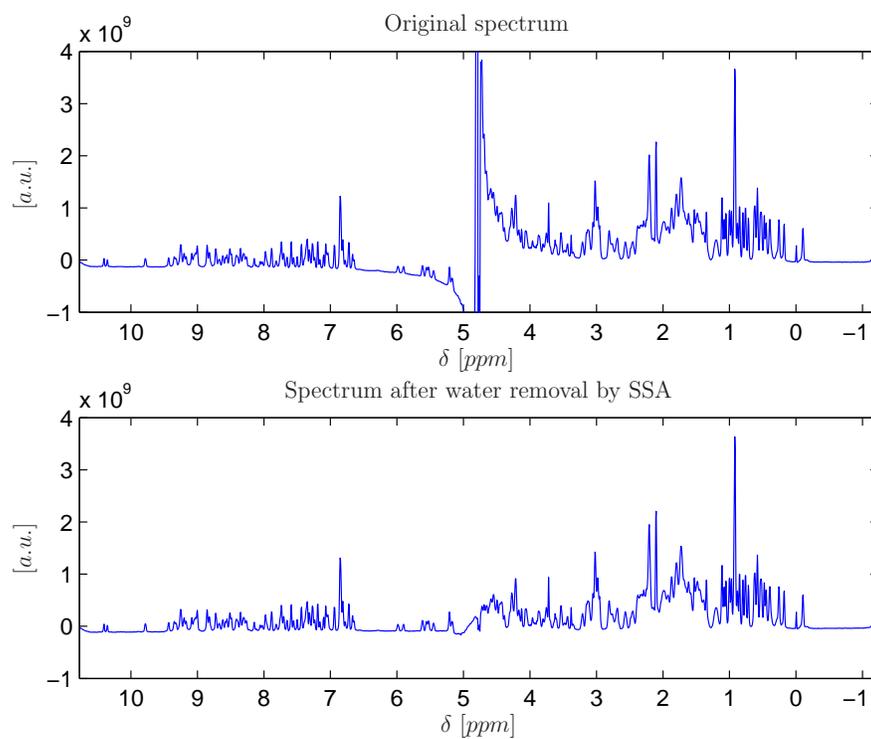


Figure 6.9: Comparison of the TmCSP spectrum before (top) and after (bottom) the water signal was removed by SSA. The water resonance disappeared almost completely while the protein peaks were not affected.

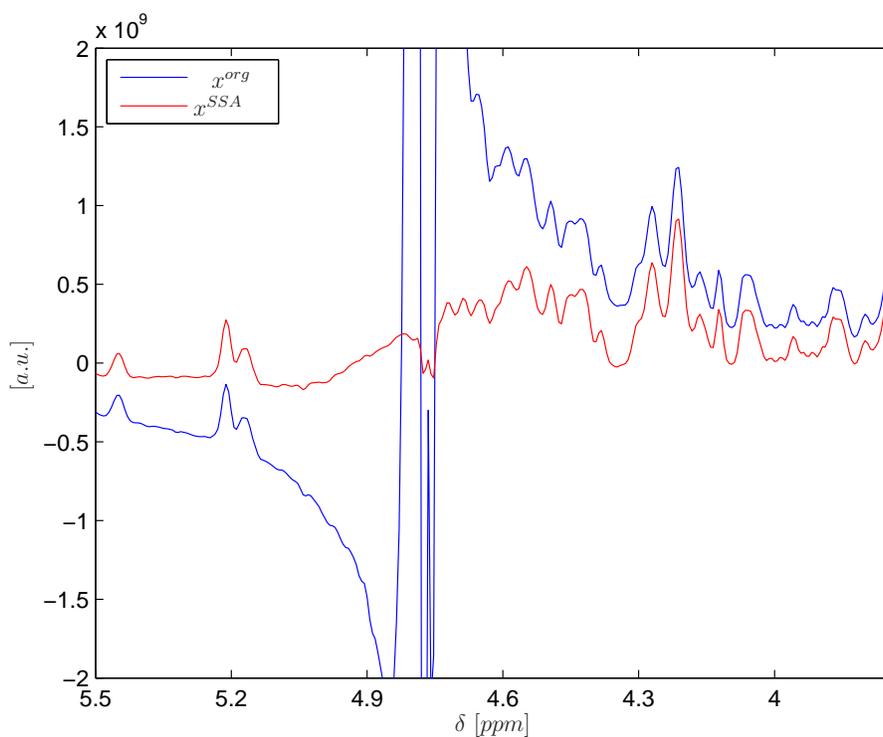


Figure 6.10: Close-up of the segment of the TmCSP-spectrum in which the water resonance resided.  $x^{org}$  denotes the recorded spectrum whereas  $x^{SSA}$  denotes the spectrum after the water signal was removed with the aid of SSA. Even protein peaks overlapping with the water signal remained in the spectrum. This becomes particularly obvious in the area right of the water signal.

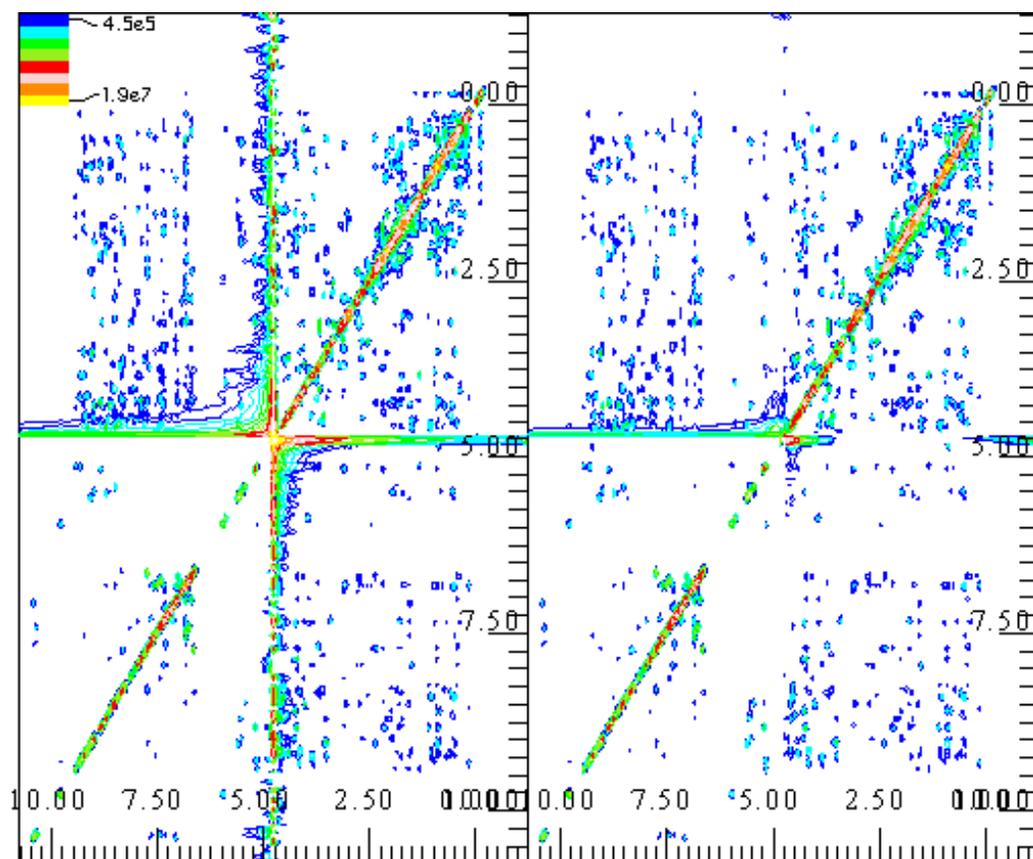


Figure 6.11: The 2D-NOESY-spectrum before (left) and after (right) the water signal was removed by means of SSA. Horizontal axis:  $\omega_2$ . Vertical axis:  $\omega_1$ . Both axes are in units of *ppm*.

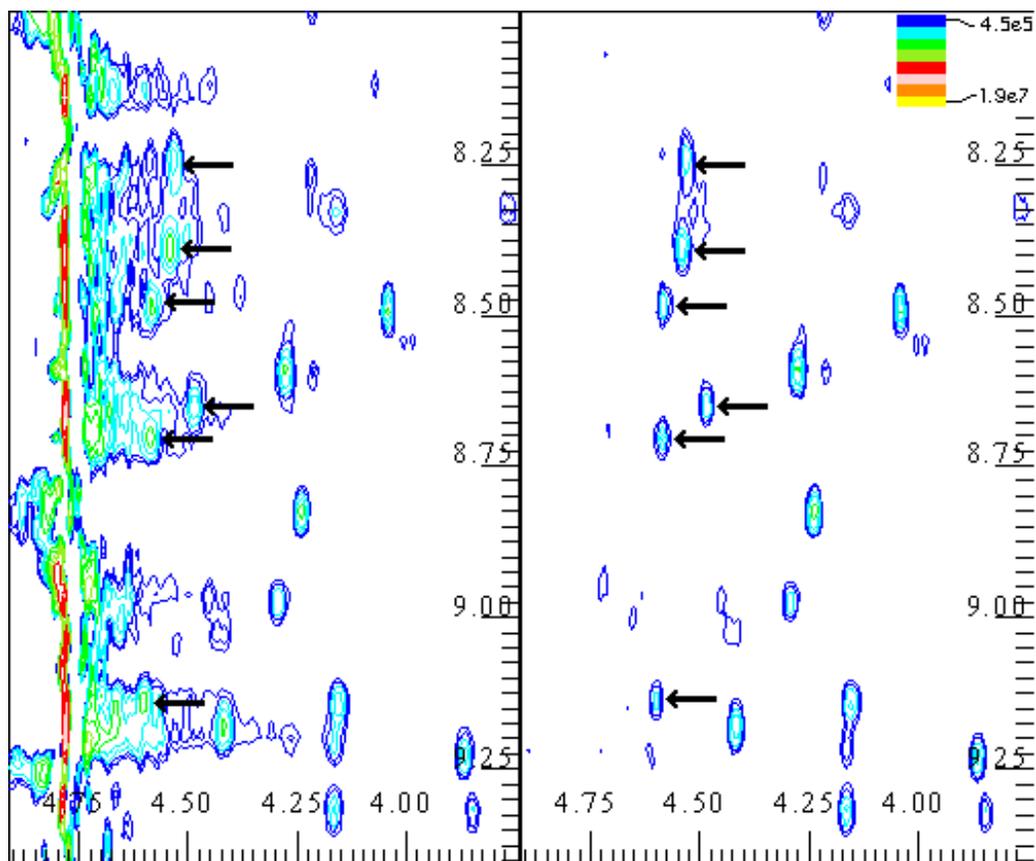


Figure 6.12: Close-up of the 2D-NOESY spectrum of TmCSP. Left: the original spectrum. Some of the peaks are hard to identify because they are partly overlapped by the water signal. Right: the spectrum after the removal of the water signal. The peaks formerly hidden by the water signal are now clearly visible.

protein peaks located in the immediate vicinity of the water resonance and as it did not lead to increased noise in the water free spectra. Seen from a practical point of view SSA had the advantage that only one parameter, namely the embedding dimension, needed to be tuned while the delay  $\Delta t$  and the number  $U$  of dominant eigenvectors considered in the PCA step could be kept fixed regardless of the particular spectrum under investigation.

The robustness of the SSA approach was proved by applying it to the spectra of four different proteins. In all these cases the water signal could be removed almost perfectly while the protein peaks remained virtually unaltered in the spectra. For the protein TmCSP a complete water free 2D spectrum was computed in which protein peaks formerly hidden by the water signal became clearly visible.

In summary SSA is an efficient, robust, and easy to use method to remove the water signal from NMR spectra which outperforms the other approaches presented in this thesis.



# Chapter 7

## Conclusions

In this thesis various approaches have been presented with which the water resonance could be removed from protein  $^1\text{H}$ -NMR spectra. In the beginning the basics of 2D-NMR spectroscopy were reviewed and the origin as well as the main features of the water resonance were discussed. Based on this introductory chapter it was shown that the data recorded in 2D-NOESY experiments could be described by the linear mixture model such that blind source separation methods appeared to be suitable to remove the water signal from NMR spectra.

For the BSS task the algorithm MP-BSS was chosen as an earlier study [53] indicated that it outperforms other BSS algorithms in the context of NMR data. Investigations using artificially generated data revealed that MP-BSS lead to better results when it was applied to  $t_2$ -Fourier transformed data than when it was used in time domain. Furthermore, the robustness of MP-BSS was analyzed with respect to deviations from the linear mixture model and to violations of the uncorrelatedness assumptions on which the algorithm is based. Eventually, MP-BSS was applied to real world P11 spectra whereas the water signal could be removed only at the expense of additional noise in the resulting spectra.

In order to remove this noise two PCA based denoising strategies, namely local and Kernel PCA, were investigated and compared. In these studies performance measures were introduced with which the optimal sets of parameters appearing in the two algorithms could be determined. It turned out that local PCA lead to better results than Kernel PCA and was principally easier to apply and more time-efficient.

Moreover, the algorithm dAMUSE was used to remove the water signal from NMR data. dAMUSE had the advantage that it comprised both BSS and denoising in a single algorithm. Again, performance measures were developed with which the optimal set of parameters could be determined.

Compared with MP-BSS combined with local PCA dAMUSE lead to similar results, i.e. it also succeeded in removing the water signal while the level of additional noise in the resulting spectra was moderate.

Together with dAMUSE the algorithm Autoassign was proposed. The latter represented a convenient tool with which the sources obtained by dAMUSE could be grouped automatically into water and protein related clusters. In Autoassign a GA was used to minimize the suggested target function. For this GA a fixed set of parameters could be found which guaranteed convergence to the global minimum within reasonable computational time.

Eventually, a SSA based approach was used to remove the water signal from NMR spectra. This method outperformed MP-BSS as well as dAMUSE in several aspects. First, even protein peaks residing in the immediate vicinity of the water resonance remained in the spectrum after SSA was applied. Second, the water-free spectra obtained by SSA did not suffer from increased noise and always the same component (i.e. the one related with the most dominant eigenvector) was rejected in the PCA step. Hence, no additional denoising or assignment procedure was necessary. And finally, only one parameter needed to be tuned in SSA which made it especially easy to apply.

The robustness of SSA was investigated considering as example the spectra of the proteins P11, TmCSP, HPr as well as the RAS binding domain of the protein RalGEF. In all cases the water signal could be removed almost perfectly while even protein signals close to the water resonance remained in the spectrum. For TmCSP the complete water-free 2D-NOESY spectrum was computed additionally whereas peaks which were formerly hidden by the water signal became clearly visible.

Hence, it is concluded that SSA is the method of choice for the removal of the water resonance from 2D-NOESY data sets.

# Appendix A

## Supplements for MP-BSS

### A.1 Diagonality of the Fourier-transformed source covariance matrix

In this section it will be shown that diagonality of the covariance matrix  $\mathbf{C}_{\mathbf{S}}$  also infers diagonality of the corresponding covariance matrix  $\mathbf{C}_{\hat{\mathbf{S}}}$  computed in frequency domain. For simplicity, it will be assumed that the underlying signals have zero mean such that the derivations can be made in terms of the correlation matrices  $\mathbf{R}_{\mathbf{S}}$  and  $\mathbf{R}_{\hat{\mathbf{S}}}$ , respectively.

**Lemma 1** (Diagonality of  $\mathbf{R}_{\hat{\mathbf{S}}}$ ). *Let  $\mathbf{R}_{\mathbf{S}} = E\{\mathbf{s}(t)^H \mathbf{s}(t)\}$  be the correlation matrix of the time domain random vector  $\mathbf{s}(t) = [s_1(t), s_2(t), \dots]$  of finite dimension. Furthermore, let  $\hat{\mathbf{s}}(\omega) = [\hat{s}_1(\omega), \hat{s}_2(\omega), \dots]$  be the Fourier transform of  $\mathbf{s}(t)$  with correlation matrix  $\mathbf{R}_{\hat{\mathbf{S}}} = E\{\hat{\mathbf{s}}^H(\omega) \hat{\mathbf{s}}(\omega)\}$ . Then diagonality of  $\mathbf{R}_{\mathbf{S}}$  infers diagonality of  $\mathbf{R}_{\hat{\mathbf{S}}}$ .*

*Proof.* Consider an arbitrary element  $r_{kl}$  of the matrix  $\mathbf{R}_{\mathbf{S}}$ :

$$r_{kl} = E\{s_k(t)s_l^*(t)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_k(t)s_l^*(t) ds_k ds_l = \delta_{kl} \quad (\text{A.1})$$

where  $\delta_{kl}$  denotes the Kronecker delta. Here, it has been assumed for simplicity that  $s_k(t)$  and  $s_l(t)$  have unit variance. The Fourier transforms of  $s_k(t)$  and  $s_l(t)$  are of the form

$$\begin{aligned} \hat{s}_k(\omega) &= \int_{-\infty}^{\infty} s_k(t)e^{-i\omega t} dt \\ \hat{s}_l^*(\omega) &= \int_{-\infty}^{\infty} s_l^*(t')e^{-i\omega t'} dt'. \end{aligned} \quad (\text{A.2})$$

The differentiation between  $t$  and  $t'$  is only made for technical reasons.

Denoting by  $p_{\cdot, \cdot}(\cdot, \cdot)$  the joint pdf the element  $\hat{r}_{kl}$  of the corresponding frequency domain matrix  $\mathbf{R}_{\mathfrak{S}}$  can now be written as

$$\begin{aligned}
\hat{r}_{kl} &= E\{\hat{s}_k(\omega)\hat{s}_l^*(\omega)\} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{s}_k(\omega)\hat{s}_l^*(\omega)p_{\hat{s}_k, \hat{s}_l}(\hat{s}_k, \hat{s}_l)d\hat{s}_k d\hat{s}_l \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_k(t)e^{-i\omega t} dt \int_{-\infty}^{\infty} s_l^*(t')e^{i\omega t'} dt' p_{s_k, s_l}(s_k, s_l) ds_k ds_l \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_k(t)e^{-i\omega t} s_l^*(t')e^{i\omega t'} \delta(t' - t) dt dt' p_{s_k, s_l}(s_k, s_l) ds_k ds_l \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_k(t) s_l^*(t') e^{i\omega(t' - t)} \delta(t' - t) dt dt' p_{s_k, s_l}(s_k, s_l) ds_k ds_l \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_k(t) s_l^*(t) p_{s_k, s_l}(s_k, s_l) ds_k ds_l}_{=\delta_{kl}} dt dt. \tag{A.3}
\end{aligned}$$

□

## A.2 Diagonality of the Eigenvector Matrix

In the following it will be shown that the eigenvector matrix of a pencil consisting of diagonal matrices is diagonal.

**Lemma 2** (Diagonality of  $\mathbf{E}$ ). *Let  $(\mathbf{A}, \mathbf{B})$  be a positive definite matrix pencil with non-degenerated eigenvalues. If  $\mathbf{A}$  and  $\mathbf{B}$  are diagonal then the eigenvector matrix  $\mathbf{E}$  of  $(\mathbf{A}, \mathbf{B})$  is diagonal up to permutations.*

*Proof.* Let  $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{M \times M}$  with  $\mathbf{A} = \text{diag}[a_{11}, a_{22}, \dots, a_{MM}]$  and  $\mathbf{B} = \text{diag}[b_{11}, b_{22}, \dots, b_{MM}]$ . Being positive definite  $(\mathbf{A}, \mathbf{B})$  has an GEVD of the form

$$\mathbf{A}\mathbf{E} = \mathbf{B}\mathbf{E}\mathbf{\Lambda}, \tag{A.4}$$

where  $\mathbf{\Lambda} \in \mathbb{R}^{M \times M}$ ,  $\mathbf{\Lambda} = \text{diag}[\lambda_{11}, \lambda_{22}, \dots, \lambda_{MM}]$ , is the eigenvalue matrix while  $\mathbf{E}$  denotes the eigenvector matrix. Considering the element  $ij$  ( $i, j = 1, 2, \dots, M$ ) of the matrices  $\mathbf{A}\mathbf{E}$  and  $\mathbf{B}\mathbf{E}\mathbf{\Lambda}$ , respectively, leads to the following componentwise equation:

$$a_{ii}e_{ij} = b_{ii}\lambda_{ii}e_{ij}. \tag{A.5}$$

Assuming  $e_{ij} \neq 0$  the above equation simplifies to

$$a_{ii} = b_{ii}\lambda_{jj}. \tag{A.6}$$

An equivalent equation is obtained for the element in the  $k$ -th column ( $k = 1, 2, \dots, M, k \neq j$ ) of the  $i$ -th row, i.e.

$$a_{ii}e_{ik} = b_{ii}\lambda_{kk}e_{ik} \quad (\text{A.7})$$

which simplifies for  $e_{ik} \neq 0$  again to

$$a_{ii} = b_{ii}\lambda_{kk}. \quad (\text{A.8})$$

Obviously, Eq. (A.5) and Eq. (A.8) can only be satisfied simultaneously if  $\lambda_{jj} = \lambda_{kk}$ . This is, however, in contradiction to the assumption that  $(\mathbf{A}, \mathbf{B})$  has non-degenerated eigenvalues. Hence,  $\mathbf{E}$  may only have one nonzero element (e.g. element  $e_{ij}$ ) per row while all other elements (i.e. elements  $e_{ik}$  with  $k = 1, 2, \dots, j-1, j+1, \dots, M$ ) have to be zero. Accordingly, also only one element per column of  $\mathbf{E}$  may be zero as otherwise the rows of  $\mathbf{E}$  would become linearly dependent.  $\square$



# Appendix B

## Supplements to Pre-Image Computations

In the following, a detailed derivation of the form of  $f(\mathbf{z})$  as shown in Eq. (4.69) will be given.

First, note that  $f(\mathbf{z}) := \|\tilde{\Phi}(\mathbf{z}) - \tilde{\Phi}(\mathbf{x}_t)^{denoised}\|^2$  (Eq. (4.67)) can be expressed as

$$\begin{aligned}
 f(\mathbf{z}) &= \|\tilde{\Phi}(\mathbf{z}) - \tilde{\Phi}(\mathbf{x}_t)^{denoised}\|^2 \\
 &= \langle \tilde{\Phi}(\mathbf{z}) - \tilde{\Phi}(\mathbf{x}_t)^{denoised}, \tilde{\Phi}(\mathbf{z}) - \tilde{\Phi}(\mathbf{x}_t)^{denoised} \rangle \\
 &= \langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{z}) \rangle - 2\langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{x}_t)^{denoised} \rangle \\
 &\quad + \underbrace{\langle \tilde{\Phi}(\mathbf{x}_t)^{denoised}, \tilde{\Phi}(\mathbf{x}_t)^{denoised} \rangle}_{\text{independent of } \mathbf{z}}
 \end{aligned} \tag{B.1}$$

Making use of the expansion  $\tilde{\Phi}(\mathbf{x}_t)^{denoised} = \sum_{u=0}^{T-1} b_u \tilde{\Phi}(\mathbf{x}_u)$  (cf. Eq. (4.66)) the term  $\langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{x}_t)^{denoised} \rangle$  in the above equation can be expressed as

$$\langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{x}_t)^{denoised} \rangle = \sum_{u=0}^{T-1} b_u \langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{x}_u) \rangle \tag{B.2}$$

By noting that

$$\begin{aligned}
& \langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{x}_t) \rangle \\
&= \left\langle \Phi(\mathbf{z}) - \frac{1}{T} \sum_{u=0}^{T-1} \Phi(\mathbf{x}_u), \Phi(\mathbf{x}_t) - \frac{1}{T} \sum_{v=0}^{T-1} \Phi(\mathbf{x}_v) \right\rangle \\
&= \langle \Phi(\mathbf{z}), \Phi(\mathbf{x}_t) \rangle - \frac{1}{T} \sum_{v=0}^{T-1} \langle \Phi(\mathbf{z}), \Phi(\mathbf{x}_v) \rangle - \underbrace{\frac{1}{T} \sum_{u=0}^{T-1} \langle \Phi(\mathbf{x}_u), \Phi(\mathbf{x}_t) \rangle}_{\text{independent of } \mathbf{z}} \\
&\quad + \underbrace{\frac{1}{T^2} \sum_{a,b=0}^{T-1} \langle \Phi(\mathbf{x}_a) \Phi(\mathbf{x}_b) \rangle}_{\text{independent of } \mathbf{z}} \tag{B.3}
\end{aligned}$$

Eq. (B.2) can be written as

$$\langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{x}_t)^{\text{denoised}} \rangle = \sum_{u=0}^{T-1} b_u \left\{ k(\mathbf{z}, \mathbf{x}_u) - \frac{1}{T} \sum_{v=0}^{T-1} k(\mathbf{z}, \mathbf{x}_v) \right\} \tag{B.4}$$

whereas terms independent of  $\mathbf{z}$  were neglected and feature space dot products  $\langle \Phi(\mathbf{z}), \Phi(\mathbf{x}_t) \rangle$  were replaced by kernel functions  $k(\mathbf{z}, \mathbf{x}_t)$ .

Furthermore, also the summand  $\langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{z}) \rangle$  in Eq. (B.1) can be expressed in terms of kernel functions as follows

$$\begin{aligned}
& \langle \tilde{\Phi}(\mathbf{z}), \tilde{\Phi}(\mathbf{z}) \rangle \\
&= \left\langle \Phi(\mathbf{z}) - \frac{1}{T} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t), \Phi(\mathbf{z}) - \frac{1}{T} \sum_{u=0}^{T-1} \Phi(\mathbf{x}_u) \right\rangle \\
&= \langle \Phi(\mathbf{z}), \Phi(\mathbf{z}) \rangle - \frac{2}{T} \sum_{t=0}^{T-1} \langle \Phi(\mathbf{z}), \Phi(\mathbf{x}_t) \rangle + \underbrace{\frac{1}{T^2} \sum_{u,v=0}^{T-1} \langle \Phi(\mathbf{x}_u), \Phi(\mathbf{x}_v) \rangle}_{\text{independent of } \mathbf{z}} \\
&= \underbrace{k(\mathbf{z}, \mathbf{z})}_{=1 \text{ for Gaussian kernel}} - \frac{2}{T} \sum_{t=0}^{T-1} k(\mathbf{z}, \mathbf{x}_t) \tag{B.5}
\end{aligned}$$

Hence, for the Gaussian kernel  $k_G$   $f(\mathbf{z})$  can be expressed as follows

$$f(\mathbf{z}) = -\frac{1}{T} \sum_{t=0}^{T-1} k_G(\mathbf{z}, \mathbf{x}_t) - \sum_{u=0}^{T-1} b_u \left\{ k_G(\mathbf{z}, \mathbf{x}_u) - \frac{1}{T} \sum_{w=0}^{T-1} k_G(\mathbf{z}, \mathbf{x}_w) \right\} + C \tag{B.6}$$

whereas the constant  $C$  summarizes any terms independent of  $\mathbf{z}$ .

# Appendix C

## Genetic Algorithms

Genetic Algorithms (GA) [20] [19] [8] are a search and optimisation method inspired by the evolution of biological systems. The first GAs were developed by US mathematician J. H. Holland in the 1970's and have since been applied to many real world optimisation problems. While being slower than other optimisation techniques like, e.g., gradient descent GAs have the considerable advantage of being applicable to discrete or discontinuous problems or to target functions with many local minima.

The core of a GA consists of a population of individuals as well as of a set of mathematical methods which are used to mimic biological reproduction. Each individual in the population represents a solution attempt to the optimisation problem under investigation. These individuals are allowed to propagate by mating, a process, in which the information stored in two of the original individuals (also called parent individuals) is combined to produce an offspring. In this process parent individuals which exhibit a better performance with respect to the optimisation problem are allowed to produce more offsprings than those which perform poorly. Hence, the majority of offsprings inherits information from already well performing individuals and is thus assumed to solve the optimisation problem even better than the parents.

After mating, small fragments of the information stored in the offsprings are altered randomly in a process called mutation. This is done in order to enable offsprings to explore areas of the search space which were not covered by any parent individuals yet.

Finally, the offsprings are inserted into the overall population and the cycle of selection, mating, mutation and reinsertion is repeated until a suitable solution is found, all the individuals represent the same solution or the search is abandoned (see Fig. C.1).

In the following, the individual steps carried out during a genetic algo-

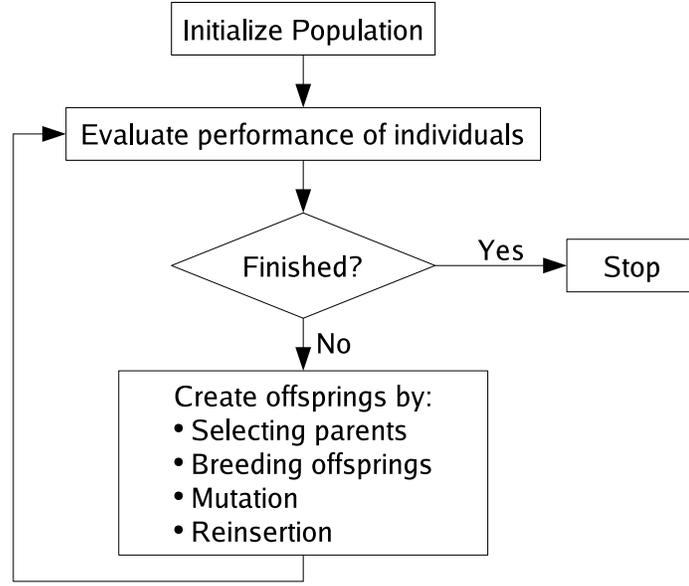


Figure C.1: Flowchart of a Genetic Algorithm. The GA has finished if either a suitable solution is found, all the individuals represent the same solution or the search is abandoned.

rithm will be elucidated considering as example the the minimization of the target function Eq. (5.33).

## C.1 Population Representation and Initialization

In the context of Autoassign the parameter set  $\mathbf{d}^{min} = [d_1^{min}, d_2^{min}, \dots, d_U^{min}]^\top$  is sought for which the target function

$$F(\mathbf{d}) := \|\hat{\mathbf{x}}_{H_2O} - \hat{\mathbf{x}}_{approx}\|^2 = \|\hat{\mathbf{x}}_{H_2O} - \mathbf{d}^\top \hat{\mathbf{S}}^{1st}\|^2 \quad (\text{C.1})$$

(cf. Eq. (5.33)) reaches its minimum. For this purpose, a population consisting of  $N_{ind}$  individuals is needed which explores the search space and which is assumed to eventually converge to the global minimum of  $F$ . Each of the individuals contains a certain realization of the parameters of  $F$ , i.e. the  $i$ -th individual,  $i = 1, 2, \dots, N_{ind}$ , consists of a binary vector  $\mathbf{d}^{i\} = [d_1^{i\}, d_2^{i\}, \dots, d_U^{i\}]^\top$  of size  $U$ . At the beginning of the GA algorithm these individuals are initialized at random in order to spread them unbiased over the search space.

## C.2 Fitness Function and Selection Probability

During each iteration, also called generation, of a GA a certain number  $N_{off}$  of offsprings has to be produced by mating individuals of the current population. At how many mating processes an individual of the current population may participate depends on its fitness with respect to the optimization problem. This means, that individuals which lead to small target function values have a higher probability to participate in mating processes than those which lead to larger values.

In order to determine these probabilities quantitatively the target function value  $F(w_1^{\{i\}}, w_2^{\{i\}}, \dots, w_N^{\{i\}})$ ,  $i = 1, \dots, N_{ind}$ , for each of the  $N_{ind}$  individuals has to be computed first. Principally, the mating probability  $p^{\{i\}}$  of the  $i$ -th individual could simply be set to

$$p^{\{i\}} = \frac{F(\mathbf{d}^{\{i\}})}{\sum_{j=1}^{N_{ind}} F(\mathbf{d}^{\{j\}})}, \quad (\text{C.2})$$

however, such an approach often leads to a premature convergence of the GA. The root cause for this problem is that individuals which are significantly fitter than the remaining ones in the beginning of the GA often dominate the population excessively and thus keep the population from exploring the search space.

This problem can be circumvented if the reproductive range of each individual is limited such that it cannot generate an exaggerated number of offsprings. One approach to achieve such a limitation is to use performance ranks in lieu of raw target function values in order to quantize the fitness the individuals for reproduction [4]. For this purpose the target function value of each individual is computed and the individuals are ranked in the descending order of these values. These ranks are then inserted into the following fitness function

$$\mathfrak{F}_{rank}^{\{i\}} = 2 - \mu + 2(\mu - 1) \frac{r^{\{i\}} - 1}{N_{ind} - 1} \quad (\text{C.3})$$

where  $\mu \in [1, 2]$  and  $r^{\{i\}}$  is the rank of the  $i$ -th individual in the population. The parameter  $\mu$  is called selective pressure and represents the probability of the best individual being selected compared to the average probability of selection of all individuals.

The mating probability of the  $i$ -th individual is then given by

$$p^{\{i\}} = \frac{\mathfrak{F}_{rank}^{\{i\}}}{\sum_{j=1}^{N_{ind}} \mathfrak{F}_{rank}^{\{j\}}}. \quad (\text{C.4})$$

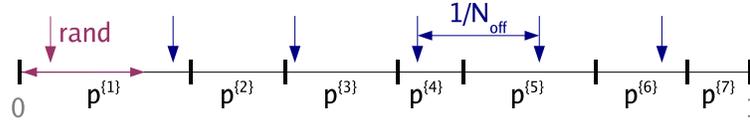


Figure C.2: Stochastic universal sampling considering as example a population of seven individuals. Six equally spaced pointers are used to select six individuals for mating. The location of the first pointer (bordeaux arrow) is chosen at random in the interval  $[0, 1/N_{off}]$  (indicated by a bordeaux double flash). The number of offsprings the  $i$ -th individual may produce corresponds to the number of arrows found in its corresponding interval  $p^{\{i\}}$ .

### C.3 Selection

During selection  $N_{off}$  individuals are chosen from the current population which will be used to produce the desired  $N_{off}$  offsprings by mating. In this process individuals with a higher selection probability  $p^{\{i\}}$  have to be chosen more often than those with lower values.

An often used scheme to select the  $N_{off}$  parent individuals is stochastic universal sampling (SUS) [4]. In this procedure the mating probabilities  $p^{\{i\}}$  of the individuals are arranged along a line segment of length one. Furthermore,  $N_{off}$  equally spaced pointers with distances of  $1/N_{off}$  are placed over this line whereas the position of the first pointer is given by a random number in the range  $[0, 1/N_{off}]$  (see Fig. C.2). The number of times an individual is selected for mating then corresponds to the number of pointers lying in its corresponding line segment.

In SUS the maximal number  $n_{max}$  of times an individual may be selected for reproduction is limited by

$$n_{max} = p_{max} N_{off} \quad (\text{C.5})$$

with  $p_{max} = \max\{p^{\{i\}}, i = 1, 2, \dots, N_{ind}\}$ , such that no individual can dominate the entire population in the early stages of the GA.

### C.4 Mating

During the selection step of the GA it is determined how many offsprings an individual may produce. Based on this information a mating population is generated in which each individual is represented as many times as it may produce offsprings [19]. From this population always two individuals, called parents, are chosen randomly and are then used to produce offsprings.

Afterwards, the parents are removed from the mating population and the process is repeated until the mating population is empty.

In the context of Autoassign the mating between two individuals  $\mathbf{d}^{\{i\}} = [d_1^{\{i\}}, d_2^{\{i\}}, \dots, d_U^{\{i\}}]$  and  $\mathbf{d}^{\{j\}} = [d_1^{\{j\}}, d_2^{\{j\}}, \dots, d_U^{\{j\}}]$  is performed as follows: first, a binary vector  $\mathbf{v} = [v_1, v_2, \dots, v_U]$  of size  $1 \times U$  is generated at random. Based on this vector two offsprings are created whereas the  $u$ -th entry of the first offspring consists of  $d_u^{\{i\}}$  and the corresponding entry of the second offspring consists of  $d_u^{\{j\}}$  for  $v_u = 1$ . The same procedure is carried out if  $v_u = 0$ , however, then  $d_u^{\{i\}}$  constitutes the  $u$ -th element of the second offspring and  $d_u^{\{j\}}$  becomes the  $u$ -th element of the first offspring.

## C.5 Mutation

During the mutation step [19] of a GA a small fraction of the offsprings obtained by mating are altered slightly. These alterations occur at random but only with a low probability such that the major information stored in the affected offsprings is preserved. The motivation for including mutations in GAs is to ensure a minimum level of diversity in the population even if the algorithm has already converged.

Binary individuals are mutated by flipping a tiny fraction of their elements. In this process the probability that a particular gene will be affected by mutation is given by a user provided parameter  $p_{mut}$  which usually ranges between 0.001 and 0.01.

## C.6 Reinsertion

Finally, the  $N_{off}$  offsprings obtained by mating and mutation have to be inserted into the population [19]. As the size of the population should not increase from generation to generation individuals of the current population have to be replaced by newly created offsprings.

In Autoassign the  $N_{off}$  least fit individuals of the current population are replaced by the offsprings. In this context the fractional difference  $g$  between the population size  $N_{ind}$  and the number of offsprings to be inserted is called generation gap. Furthermore, the  $n_{elit}$  fittest individuals of the current population are protected from replacement in Autoassign, i.e. these individuals are guaranteed to propagate to the next generation.

After the offsprings have been inserted into the population their target function values are computed before a new generation of the GA begins.

## C.7 Subpopulations

Apart from the ranking based fitness assignment also dividing the overall population into smaller subpopulations may help to keep the GA from converging prematurely [25]. These subpopulations evolve independently of each other for  $T_{mig}$  generations whereupon they exchange a certain fraction  $\mu_{mig}$  of their individuals. In Autoassign, the so-called complete net structure migration scheme is used in which each subpopulation exchanges a certain number  $N_{migrate}$  of individuals with each other subpopulation. In this process the fittest individuals of each subpopulation are used as emigrants and replace the least fit individuals in their new host population.

The purpose of migration is the following: after  $T_{mig}$  generations the individual subpopulations are assumed to have partly converged such that the diversity in their individuals is already low. This diversity is increased if individuals of other populations, which may have converged to another point in search space, immigrate. Hence, after migration the individual subpopulations will explore larger areas of search space and may hence be able to leave local minima.

## C.8 Termination

Eventually a suitable stopping criterion is needed which determines when a GA should terminate. In Autoassign the GA algorithm is stopped if the fittest individual of the population does not change within  $T_{mig}$  generations. For the large  $T_{mig}$  value used in Autoassign this also means that the entire population has converged to a single minimum.

# Bibliography

- [1] <http://www.kyb.mpg.de/bs/people/spider/>.
- [2] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821 – 837. 1964.
- [3] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48. 2002.
- [4] James E. Baker. Reducing bias and inefficiency in the selection algorithm. *Proceedings of the Second International Conference on Genetic Algorithms*, pages 14–21. 1987.
- [5] G.H. Bakir, J. Weston, and B. Schölkopf. Learning to find pre-images. *Advances in Neural Information Processing Systems*, 16:449–456. 2004.
- [6] A. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159. 1995.
- [7] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444. 1997.
- [8] Erick Cantú-Paz. Efficient and Accurate Parallel Genetic Algorithms, volume 1 of *Genetic Algorithms and Evolutionary Computation*. Kluwer Academic Publishers Group, Norwell, Massachusetts, USA, 2 edition. ISBN 0792372212. 2000.
- [9] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114. 1997.
- [10] A.J. Chipperfield, P.J. Fleming, and C.M. Fonseca. Genetic algorithm tools for control systems engineering. In Proc. Adaptive Computing in

- Engineering Design and Control, pages 128–133. Plymouth Engineering Design Center. 1994.
- [11] S. Choi and A. Cichocki. Blind separation of nonstationary and temporally correlated sources from noisy mixtures. *IEEE Workshop on Neural Networks for Signal Processing, NNSP*, pages 405–414. December 2000.
- [12] P. Comon, C. Jutten, and J. Héroult. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10. 1991.
- [13] P. Comon, C. Jutten, and J. Héroult. Blind separation of sources, part II: Problems statement. *Signal Processing*, 24:11–20. 1991.
- [14] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1. Interscience Publishers Inc., New York. 1953.
- [15] T. Cox and M. Cox. *Multidimensional Scaling*. Number 88 in *Mono-graphs on Statistics and Applied Probability*. Chapman & Hall, second edition edition. 2001.
- [16] N. Dunford and J.T. Schwartz. *Linear operators Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space*. John Wiley & Sons, New York. 1963.
- [17] P. Georgiev, F. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Trans. on Neural Networks*, 16(4):992–996. July 2005.
- [18] M. Geyes, Ch. Herrmann, S. Wohlgemuth, A. Wittinghofer, and H.-R. Kalbitzer. Structure of the RAS-binding domain of RasGEF and implications for RAS binding and signalling. *Natural Structural Biology*, 4:694–699. 1997.
- [19] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, Boston, Massachusetts, USA. ISBN 0201157675. 1989.
- [20] David E. Goldberg. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*, volume 7 of *Genetic Algorithms and Evolutionary Computation*. Kluwer Academic Publishers Group, Norwell, Massachusetts, USA, 2 edition. ISBN 1402070985. 2002.
- [21] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, USA. 1989.

- [22] A. Görler and H. R. Kalbitzer. RELAX, a flexible program for the back calculation of noesy spectra based on a complete relaxation matrix formalism. *Journal of Magnetic Resonance*, 124:177–188. 1997.
- [23] J. Gower. Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55:582–585. 1968.
- [24] T. Graepel and K. Obermayer. Fuzzy topographic kernel clustering. In Proceedings of the 5th GI Workshop Fuzzy Neuro Systems '98, pages 90–97. 1998.
- [25] P. Grosso. Computer simulations of genetic adaptation: Parallel Sub-component Interaction in a Multilocus Model. Ph.D. thesis, University of Michigan. 1985.
- [26] P. Gruber, K. Stadlthanner, M. Böhm, F. J. Theis, and E. W. Lang. Denoising using local projective subspace methods. *Neurocomputing*, 69:1485–1501. 2006.
- [27] S. Harmeling, A. Ziehe, M. Kawanabe, and K.R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15:1089–1124. 2003.
- [28] K. H. Hausser and H. R. Kalbitzer. NMR in Medicine and Biology. ISBN 3-540-50281-5. Springer-Verlag, Berlin, Heidelberg, New York. 1989.
- [29] J. Héroult and J. Jutten. Space or time adaptive signal processing by neural network models. In J. Denker (editor), *Neural Networks for Computing*. Proceedings of the AIP Conference, pages 206–211. American Institute of Physics, New York. 1986.
- [30] J. P. Hornak. The basics of NMR. September 2006, <http://www.cis.rit.edu/htbooks/nmr/>.
- [31] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634. 1999.
- [32] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent Component Analysis. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons, Inc., New York, Chichester, Weinheim, Brisbane, Singapore, Toronto. 2001.

- [33] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice Hall. 1988.
- [34] John Keeler. Understanding NMR spectroscopy. September 2006, <http://www.spectroscopynow.com/coi/cda/detail.cda?id=10171&type=EducationFeature&chId=5&page=1>.
- [35] J.B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129. 1964.
- [36] J.T.-Y. Kwok and I.W.-H. Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15(6):1517–1525. 2004.
- [37] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.R. Müller. Fisher discriminant analysis with kernels. In Neural Networks for Signal Processing IX. Proceedings of the 1999 IEEE Signal Processing Society Workshop, pages 41–48. 1999.
- [38] Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel PCA and de-noising in feature spaces. In Proceedings of the 1998 conference on Advances in neural information processing systems II, pages 536 – 542. MIT Press, Cambridge, MA, USA. 1999.
- [39] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer’s theorem, feature maps and smoothing. In Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006). Springer. to be published.
- [40] L. Molgedey and G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637. 1994.
- [41] D. Neuhaus, M. P. Williamson, and A. Marchand. The nuclear overhauser effect in structural and conformational analysis. Wiley-VCH, Weinheim, Germany. 1992.
- [42] A. Papoulis. Probability, Random Variables, and Stochastic Processes. McGraw-Hill, 3rd edition. 1991.
- [43] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572. 1901.
- [44] M. Plumbley. Conditions for nonnegative independent component analysis. *Signal Processing Letters, IEEE*, 9:177–180. 2002.

- [45] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471. 1978.
- [46] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326. 2000.
- [47] S. Saitoh. Theory of Reproducing Kernels and its Applications. Longman Scientific & Technical, Harlow, England. 1988.
- [48] A. Samarov and A. Tsybakov. Nonparametric independent component analysis. *Bernoulli*, 10:565–682. 2004.
- [49] I. W. Sandberg and L. Xu. Uniform approximation of multidimensional myoptic maps. *Transactions on Circuits and Systems*, 44:477–485. 1997.
- [50] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Müller, Gunnar Rätsch, and A. J. Smola. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017. 1999.
- [51] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319. 1998.
- [52] K. Stadlthanner, F.J. Theis, E.W. Lang, A.M. Tomé, C.G. Puntonet, P. Gómez Vilda, T. Langmann, and G. Schmitz. Sparse nonnegative matrix factorization applied to microarray data sets. In J. Rosca, D. Erdogamus, J.C. Principe, and S. Haykins (editors), Independent Component Analysis and Blind Source Separation, 6th International Conference, ICA 2006, Charleston, SC, USA, volume 3889 of *Lecture Notes in Computer Science*, pages 254–261. Springer-Verlag, Berlin Heidelberg. 2006.
- [53] Kurt Stadlthanner. Adaptive statistische Verfahren zur Analyse mehrdimensionaler NMR-Spektren. Diploma thesis, University of Regensburg. March 2003.
- [54] F. Takens. On the numerical determination of the dimension of an attractor. *Dynamical Systems and Turbulence, Annual Notes in Mathematics*, 898:366–381. 1981.
- [55] A. M. Tomé. An iterative eigendecomposition approach to blind source separation. In Proceedings of the 3rd International Conference on Independent Component Analysis and Signal Separation, pages 424–428. San Diego, USA. 2001.

- [56] A.M. Tomé, A.R. Teixeira, E.W. Lang, K. Stadlthanner, A.P. Rocha, and R. Almeida. dAMUSE - a new tool for denoising and blind source separation. *Digital Signal Processing*, 15:400–421. 2005.
- [57] L. Tong, R.W. Liu, V. Soon, and Y.F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38:499–509. 1991.
- [58] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York. 1995.
- [59] Rolf Vetter. Extraction of efficient and characteristic features of multidimensional time series. Ph.D. thesis, EPFL, Lausanne, Switzerland. 1999.
- [60] Rolf Vetter, J.M. Vesin, Patrick Celka, Philippe Renevey, and Jens Krauss. Automatic nonlinear noise reduction using local principal component analysis and mdl parameter selection. In *Proceedings of the IASTED International Conference on Signal Processing Pattern Recognition and Applications (SPPRA02)*, Crete, pages 290–294. 2002.
- [61] G. Wider. Structure determination of biological macromolecules in solution using NMR spectroscopy. *BioTechniques*, 29:1278–1294. 2000.
- [62] G. Wider, S. Macura, A. Kumar, R. R. Ernst, and K. Wüthrich. Homonuclear two-dimensional  $^1\text{H}$  NMR of proteins. experimental procedures. *Journal of Magnetic Resonance*, 56:207–234. 1994.
- [63] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition. *Neural Computations*, 13(4). 2001.
- [64] A. Ziehe and K.-R. Müller. Tdsep - an efficient algorithm for blind separation using time structure. *ICANN'98*, pages 675–680. 1998.

# Acknowledgments

I want to thank all the people who have helped me writing this thesis. Special thanks go to

- Prof. Dr. Elmar Lang, my supervisor. He kindly helped me whenever I had questions or got stuck during my thesis. I also want to thank him for all the support he gave me during my time in Granada and in Portugal.
- Prof. Dr. Ana Maria Tomé for the many interesting discussions especially concerning matrix pencil based BSS and dAMUSE. Furthermore, I want to thank her for her hospitality during my time in Portugal.
- Prof. Dr. Carlos Puntonet for the many fruitful discussions during my time in Granada.
- Dr. Dr. Fabian Theis for helping me with mathematical questions.
- my former diploma student Matthias Böhm for the experiments he had carried out during the development of Autoassign.
- Peter Gruber for the good cooperation during the development of the denoising techniques.
- all the members of the AG Lang for the fruitful discussions.
- the Graduiertenkolleg “Nonlinearity and Nonequilibrium in condensed matter” and to the DFG for financially supporting me during the last years.
- Natacha, my beloved wife for all her support, her encouragement and all the patience she had with me during the last years. Merci, mon amour.
- my parents and my sister. It is great privilege to have a family you can fully rely on.