Blind Source Separation based on Joint Diagonalization of Matrices with Applications in Biomedical Signal Processing

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium
– Dr. rer. nat. –

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Andreas Ziehe

Potsdam, im April 2005

Contents

Abstract

This thesis is concerned with the solution of the blind source separation problem (BSS). The BSS problem occurs frequently in various scientific and technical applications. In essence, it consists in separating meaningful underlying components out of a mixture of a multitude of superimposed signals.

In the recent research literature there are two related approaches to the BSS problem: The first is known as Independent Component Analysis (ICA), where the goal is to transform the data such that the components become as independent as possible. The second is based on the notion of diagonality of certain characteristic matrices derived from the data. Here the goal is to transform the matrices such that they become as diagonal as possible. In this thesis we study the latter method of approximate joint diagonalization (AJD) to achieve a solution of the BSS problem. After an introduction to the general setting, the thesis provides an overview on particular choices for the set of target matrices that can be used for BSS by joint diagonalization.

As the main contribution of the thesis, new algorithms for approximate joint diagonalization of several matrices with non-orthogonal transformations are developed.

These newly developed algorithms will be tested on synthetic benchmark datasets and compared to other previous diagonalization algorithms.

Applications of the BSS methods to biomedical signal processing are discussed and exemplified with real-life data sets of multi-channel biomagnetic recordings.

Zusammenfassung

Diese Arbeit befasst sich mit der Lösung des Problems der blinden Signalquellentrennung (BSS). Das BSS Problem tritt häufig in vielen wissenschaftlichen und technischen Anwendungen auf. Im Kern besteht das Problem darin, aus einem Gemisch von überlagerten Signalen die zugrundeliegenden Quellsignale zu extrahieren.

In wissenschaftlichen Publikationen zu diesem Thema werden hauptsächlich zwei Lösungsansätze verfolgt:

Ein Ansatz ist die sogenannte "Analyse der unabhängigen Komponenten", die zum Ziel hat, eine lineare Transformation $\mathbf{V}$ der Daten $\mathbf{X}$ zu finden, sodass die Komponenten U_n der transformierten Daten $\mathbf{U} = \mathbf{VX}$ (die sogenannten "independent components") so unabhängig wie möglich sind. Ein anderer Ansatz beruht auf einer simultanen Diagonalisierung mehrerer spezieller Matrizen, die aus den Daten gebildet werden. Diese Möglichkeit der Lösung des Problems der blinden Signalquellentrennung bildet den Schwerpunkt dieser Arbeit.

Als Hauptbeitrag der vorliegenden Arbeit präsentieren wir neue Algorithmen zur simultanen Diagonalisierung mehrerer Matrizen mit Hilfe einer nicht-orthogonalen Transformation.

Die neu entwickelten Algorithmen werden anhand von numerischen Simulationen getestet und mit bereits bestehenden Diagonalisierungsalgorithmen verglichen. Es zeigt sich, dass unser neues Verfahren sehr effizient und leistungsfähig ist. Schließlich werden Anwendungen der BSS Methoden auf Probleme der biomedizinischen Signalverarbeitung erläutert und anhand von realistischen biomagnetischen Messdaten wird die Nützlichkeit in der explorativen Datenanalyse unter Beweis gestellt.

Acknowledgements

Above all, I would like to thank Prof. Dr. Klaus-Robert Müller for supervising the present dissertation. Without his guidance and inexhaustible support I would have never completed this work.

I am also grateful to Prof. Dr. Erkki Oja and Prof. Dr. Klaus Pawelzik for agreeing to be *Gutachter* of my thesis.

The work for this thesis has been carried out at the Fraunhofer Institute FIRST (formerly known as GMD FIRST) in Berlin and therefore I would like to thank Prof. Dr. Stefan Jähnichen as the head of this Institute for continued support of my work.

In the Intelligent Data Analysis (IDA) group at FIRST, I found an open-minded research atmosphere and a constant source of profound knowledge from which I always profited a lot. It is my pleasure to thank all current and former members, including Dr. Gilles Blanchard, Dr. Benjamin Blankertz, Mikio Braun, Guido Dornhege, Dr. Stefan Harmeling, Dr. Julian Laub, Dr. Motoaki Kawanabe, Dr. Jens Kohlmorgen, Matthias Krauledat, Dr. Pavel Laskov, Steven Lemm, Frank Meinecke, Dr. Sebastian Mika, Dr. Noboru Murata, Dr. Guido Nolte, Dr. Takashi Onoda, Dr. Gunnar Rätsch, Christin Schäfer, Prof. Dr. Bernhard Schölkopf, Rolf Schulz, Dr. Anton Schwaighofer, Dr. Alex Smola, Sören Sonnenburg, Dr. Masashi Sugiyama, Dr. Koji Tsuda, Dr. Ricardo Vigário and Dr. Olaf Weiss.

Very special thanks go to Christin, Frank, Motoaki, Pavel, Sebastian and Stefan who have helped me so much with proof-reading and most valuable mental support in the final phase of the work.

In particular, I also want to mention Dr. Benjamin Blankertz, Prof. Dr. Gabriel Curio, Dr. Stefan Harmeling, Dr. Motoaki Kawanabe, Dr. Pavel Laskov, Prof. Dr. Klaus-Robert Müller, Dr. Bruno-Marcel Mackert, Frank Meinecke, Prof. Dr. Noboru Murata, Dr. Guido Nolte, Dr. Lutz Trahms, Dr. Ricardo Vigário, Dr. Gerd Wübbeler and Dr. Arie Yeredor with whom I have closely collaborated and co-authored the scientific papers that formed the basis of the present thesis. Sincere thanks are given to them all.

Furthermore, I want to express my gratitude for the possibility to use unique real-world datasets provided by Prof. Dr. Gabriel Curio's Neurophysics group in the Department of Neurology of the Charitè University Medicine Berlin and the Biomagnetism group of the Physikalisch-Technische

Bundesanstalt, headed by Prof. Dr. Hans Koch and Dr. Lutz Trahms.

I would like to thank the members of the European Project BLISS, Prof. Dr. Christian Jutten, Prof. Dr. Luis Almeida, Prof. Dr. Dinh-Tuan Pham and Prof. Dr. Erkki Oja. Their ideas and insights had a crucial influence on my own research.

Finally, I gratefully acknowledge financial support from DFG grants (JA 379/5-2, 379/7-1, DFG SFB 618-B4), from the EU project BLISS (IST-1999-14190) and from the EU PASCAL network of excellence (IST-2002-506778).

Last—but by no means least—I thank my parents.

Chapter 1

Introduction

In this introductory chapter we give an overview of the problem and dicuss why it is important. Furthermore we outline the thesis.

Lack of data is hardly the problem these days since with our modern devices we can read and measure practically everything. But how are we supposed to cope with the growing amount of signals and data?

In this thesis we follow the approach of multivariate data analysis. In particular, we consider this question as relating especially to the field of unsupervised data analysis and blind source separation (BSS).

The BSS problem occurs frequently in various scientific and technical applications. In essence, it consists in separating meaningful underlying components out of a mixture of a multitude of superimposed signals. Signals are mixed since they are transmitted over a shared medium. A popular example to illustrate this problem is the so called 'cocktail-party' effect: in a conversation, which is held in a crowded room with many people speaking at the same time, we are often remarkably well able to separate a particular voice from the background babble. In contrast, a computer program, aimed at automatic speech recognition would fail miserably under these circumstances, since the speech recognition system can not match the mixed utterance to a single word or phrase.

As in this example, efficient methods to separate superimposed signals originating from different sources without knowing about the source characteristics in detail are of great importance and practical relevance in many scientific and technical applications. Our strongest motivation to study the blind source separation problem in the first place, originates from the goal of studying the human brain by measuring the electrical or magnetical signals as they are detected outside of the body. Here the BSS approach has a great potential to reveal highly useful information about the electrophysiological processes inside the brain as discussed in the following section.

1.1 The Biomedical Signal Processing Challenge

Recent advances in biomedical signal processing allow to monitor the active brain non-invasively at high spatial and temporal resolution. In particular modern MEG and EEG hardware routinely record signals in the femto-Tesla range, at hundreds of points all over the head or body, up to 4000 times each second (Drung, 1995). This high sensitivity is needed to enable physicians to get precise information about ongoing electro-physiological processes in the brain. Hence the new measurement techniques provide a valuable tool for clinical applications and for the longterm research goal to better understand the mechanism of information processing in the brain. However, the increased sensitivity poses an enormous challenge for signal processing and data analysis since signals from a multitude of different biological processes and noise sources obfuscate the signal of interest. Thus it is of utmost importance in this undertaking to improve the signal-to-noise ratio, especially when the ongoing activity of the brain is to be studied on a single-trial basis.

For example in MEG sophisticated active and passive shielding is used to reduce unwanted signals, like the omni-present power-line interferences, which otherwise contaminate the measurements.

In this thesis we are interested in efficient and robust mathematical algorithms to reduce disturbances originating from technical or body-intern noise sources. Here we make use of the fact that many of these processes vary in intensity *independently* of each other.

We focus on the application of a recently developed unsupervised data analysis technique known as blind source separation to process multi-channel recordings of biomedical signals. In this setting one has only access to measurements of mixed, i.e. superimposed signals and the question is how to construct suitable algorithms that allow to demix and thus find the underlying (unmixed) signals of interest. Blind source separation techniques aim exactly to reveal unknown underlying sources of an observed mixture $\mathbf{x}(t)$ using two ingredients (I) a model about the mixing process (typically a linear superposition as in equation 1.1) and (II) the assumption of statistical independence. As opposed to other signal processing techniques like beamforming or spectral analysis, BSS does not rely on precise information about the geometry of the sensor array or the knowledge of the frequency content of the underlying sources. Therefore this source separation method is called "blind".

Furthermore, in the context of EEG and MEG data such a "blind" decomposition approach reveals important information about the analyzed brain signals in the sense of a spatio-temporal model:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1.1}$$

The columns of $\mathbf{A}$ represent the coupling of a source with each sensor. This

information gives rise to a spatial pattern. The sources $\mathbf{s}(t)$ describe the dynamics, i.e. the time courses, of the components. This decoupling of spatial and temporal information offers a valuable tool for exploratory data analysis and hence the BSS approach can be used to extract meaningful features from large-scale multi-channel data.

1.2 Algorithmical Solutions

There are two related approaches to the BSS problem: The most popular is known as Independent Component Analysis, where the goal is to transform the data such that the components become as independent as possible. An alternative approach is based on the notion of diagonality of certain characteristic matrices derived from the data. The solution for the BSS problem is obtained by estimating the generalized eigenvectors of suitably defined matrix-valued statistics of the observed data. Thus the BSS problem can be solved by solving an analogous problem of joint diagonalization (JD). The goal of joint diagonalization consists in the following problem: Given a set of K $N \times N$ "target matrices" $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K$, find a $N \times N$ non-singular matrix $\mathbf{V}$ such that the transformed matrices $\mathbf{V}\mathbf{C}_k\mathbf{V}^T$ become diagonal (or as diagonal as possible) for all k.

What makes this problem a difficult one is the fact that it is a non-linear constrained optimization problem.

In this thesis we present a general strategy how to cope with such *constrained* optimization problems by exploiting the special structure of our problem.

As the main contribution we will derive new algorithms operating on the manifold of invertible matrices with unit determinant aimed at optimizing a joint diagonalization cost function. The key point is that by ensuring those structural constraints we naturally circumvent the trivial (zero) minimizer and at the same time have the possibility to *simplify* the optimization problem.

Thus we are in a position to strongly advocate algebraic methods for BSS, which estimate a solution for the BSS problem by estimating generalized eigenvectors that simultaneously diagonalize certain, suitably defined matrix-valued statistics of the observations. This approach allows us to efficiently use the time-structure of signals as a criterion to separate the observed mixtures in real-world biomedical applications.

1.3 Outline of the Thesis

In chaper 2 we introduce the notion of blind source separation and review a statistical (maximum likelihood) approach for its solution. Furthermore we find evidence that BSS can be equally well formulated as an approximate joint diagonalization (AJD) problem, which allows to conveniently use the time structure of the signals for the separation.

In chapter 3 we present new algorithms for AJD using multiplicative exponential updates for non-linear constrained optimization. In particular we derive two novel algorithmical solutions: a gradient method, called DOMUNG and a Newton-like method, called FFDiag.

In chapter 4 we test and compare the newly developed algorithms by numerical simulations.

In chapter 5 we come back to our original problem: biomedical signal processing in real-world environments. We present results towards clinical applications.

In chapter 6 a discussion is given, a conclusion is drawn and recommendations for future research are made.

Chapter 2

Blind Source Separation

In this chapter we establish a link between two problems: First we give an introduction to the problem of blind source separation (BSS). The nature of the problem and typical approaches to its solution are briefly reviewed. Then we find evidence that many of these approaches can be formulated as a related problem of approximate joint diagonalization (AJD). Diagonalization techniques provide a unifying framework to design efficient numerical algorithms for BSS. Thus we review some of the joint diagonalization criteria available in the rich BSS literature.

The concepts of Blind Source Separation (BSS) and Independent Component Analysis (ICA) are actively researched since the early 1980s. They are truly interdisciplinary and attracted the attention from researchers in signal processing, statistics and machine learning mainly in the context of artificial neural networks. A wealth of novel successful algorithms have emerged and so BSS has now become a well-established method in unsupervised learning and statistical signal processing. In the following we introduce only some of the basic ideas. For a more broad overview we refer to the excellent books of Hyvärinen et al. (2001) or Haykin (2000). Also the proceedings of the regularly held ICA workshops provide a wealth of related material (Cardoso et al., 1999; Pajunen and Karhunen, 2000; Lee et al., 2001; Amari et al., 2003; Puntonet and Prieto, 2004).

2.1 Problem Statement

BSS is an highly relevant problem of great interest in many scientific and technical applications. It constitutes a classical goal of science to separate an observed mixture (of signals) into several basic components.

Let us define the BSS problem. In the simplest case we consider a linear

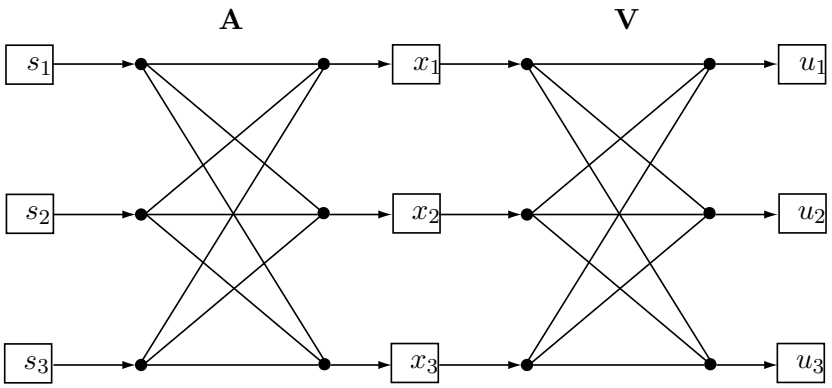

Figure 2.1: Graphical model of the blind separation setting for three sources s_1, s_2, s_3.

and instantaneous superposition of independent signals. More formally, we assume we are given some linear mixtures $x_i(t)$ of a number of statistically independent source signals $s_j(t)$, where t is a (time-)index, obeying the equation

$$x_i(t) = \sum_{j=1}^{m} A_{ij} s_j(t), \;\; (i = 1, \ldots, N, \, j = 1, \ldots, M). \tag{2.1}$$

For convenience, the mixing model of equation (2.1) can also be written in matrix notation:

$$\mathbf{X} = \mathbf{AS}, \tag{2.2}$$

where the entries of the data matrix $\mathbf{X}$ are samples of the $x_i(t)$ in equation 2.1 giving rise to column vectors $\mathbf{x}[t] = [x_1[t], ..., x_N[t]]^T$, the $N \times M$ matrix $\mathbf{A}$ has elements A_{ij} and the (source signals) matrix $\mathbf{S}$, analogous to the construction of $\mathbf{X}$, has column vectors $\mathbf{s}[t] = [s_1[t], ..., s_M[t]]^T$.

The goal of BSS consists of recovering the set of source signals $\mathbf{S}$ solely from the observed (instantaneous and linear) mixtures $\mathbf{X}$, by estimating either the mixing matrix $\mathbf{A}$ or its inverse $\mathbf{V} = \mathbf{A}^{-1}$ (silently assuming that $\mathbf{A}$ is invertible).

Restated in the matrix formulation, the BSS problem consists in *factoring* the observed signals data matrix $\mathbf{X}$ into the mixing matrix $\mathbf{A}$ and the source signals matrix $\mathbf{S}$.

Though without further constraints this factorization problem is not uniquely determined, i.e. many solutions exist that fulfill equation (2.2), one can think of a variety of constraints utilizing prior knowledge about the source characteristics or the mixing model (depending on the application) that will allow to resolve almost all of these the indeterminacies.

In any case, however, since a scalar factor can always be exchanged between each row of $\mathbf{S}$ and the corresponding column of $\mathbf{A}$ without changing the product, the amplitudes and signs of the source signals s_j are not

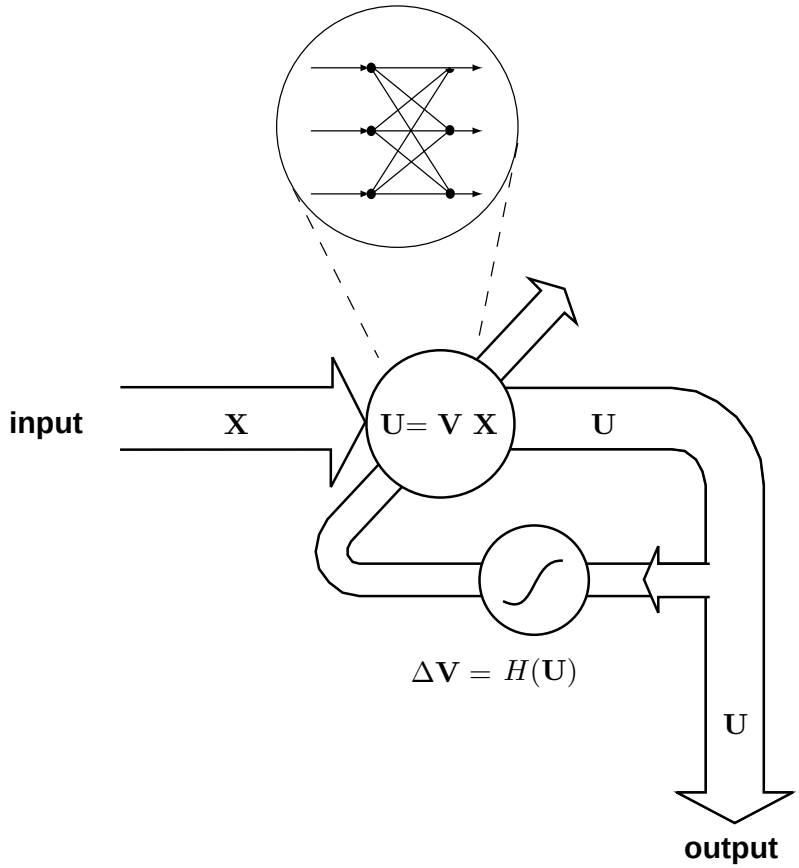

Figure 2.2: General architecture of an BSS algorithm. One tries to adapt the unmixing matrix $\mathbf{V}$ such that the output components $\mathbf{U}$ fulfill a criterion $H(\mathbf{U})$.

uniquely defined. In the same way the ordering of the sources is ambiguous. For this reason the scale and the order of the source signals (and the corresponding columns of $\mathbf{A}$) is meaningless and will at best be determined by a suitable notational convention. Hence, any demixing procedure can recover the original set of source signals except for two (minor) deviations: amplitude scaling and permutation, i.e. the estimated signals $\mathbf{U} = \mathbf{VX}$ will resemble the sought-after original source signals $\mathbf{S}$ up to left-multiplication by a diagonal matrix $\mathbf{D}$ and a permutation matrix $\mathbf{P}$:

$$\mathbf{U} = \mathbf{VX} = \mathbf{PDS}.$$

Another interpretation of these indeterminacies, first noticed in (Cardoso, 1998b), is that due to the inherent indeterminacies the blind source separation problem actually consists in *identifying* an unordered set of one-dimensional *source signal subspaces*. Recovering the sources s_j corresponds to projecting $\mathbf{X}$ to one-dimensional subspaces defined by the columns of $\mathbf{A}$.

2.2 ICA Approach

The key concept that allows for a solution of the BSS problem is the notion of *statistical independence* (Jutten and Herault, 1991; Comon, 1994). As stated above, it is assumed that the source signals s_j which form the rows of $\mathbf{S}$, are statistically independent.

Intuitively, this property is important for blind source separation because the mixing process introduces dependencies, hence maximizing the independence is equivalent to separation. ICA tries to find the most independent components of the observed data.

Popular methods for ICA are based on maximization of the output entropy (Bell and Sejnowski, 1995) or minimization of mutual information between the outputs (Amari et al., 1996) which is in fact the minimization of the Kullback-Leibler divergence between the joint and the product of the marginal distributions of the outputs.

Since both approaches have been shown to be mathematical equivalent (Cardoso, 1997) to the statistical principle of maximum-likelihood estimation, we briefly mention the main concepts in the following subsections and present the maximum-likelihood approach in detail in subsection 2.2.4.

2.2.1 Statistical Independence

Statistical independence is stated mathematically in terms of the probability density function (pdf):

$$p(s_1, \ldots, s_M) = \prod_j^M p(s_j) \tag{2.3}$$

where $p(s_1, \ldots, s_M)$ denotes the joint pdf and the $p(s_j)$ denote the marginal pdf's of the sources. In other words, for independent random variables the joint probability distribution has a very simple form: it is just the product of the marginal distributions.

2.2.2 Characteristic Function

A related concept to assess the independence of variables is based on the so called characteristic function. The characteristic function of an n-dimensional random vector $\mathbf{x}$ is defined as

$$\mathbf{\Phi}_{\mathbf{x}}(\boldsymbol{\omega}) = \int_{\Re^p} e^{i\boldsymbol{\omega}^T\mathbf{x}} dF(\mathbf{x}) = \mathrm{E}_{\mathbf{x}}\{e^{i\boldsymbol{\omega}^T\mathbf{x}}\} \tag{2.4}$$

where $i = \sqrt{-1}$ and $\boldsymbol{\omega}$ is the transformed variable corresponding to $\mathbf{x}$.

2.2.3 Mutual Information

A well-known measure of independence of random variables is their *mutual information* (MI)[1]. For two random variables X and Y MI is defined as

$$MI(X,Y) = \int\int dXdY \log \frac{p(X,Y)}{p_X(X)p_Y(Y)}.$$

This is the relative entropy or Kullback Leibler divergence between the joint pdf and the product of the marginal distributions $p_X(X), p_Y(Y)$. MI is zero if and only if the random variables are independent (Cover and Thomas, 1991). Another name for mutual information is redundancy since the mutual information $MI(X,Y)$ can be understood as the reduction in the uncertainty about X given the knowledge of Y. If there is no more redundancy we have reached independence and knowing one variable does not provide any additional information about the other.

2.2.4 Maximum-Likelihood Estimation

Among the approaches to solve the ICA/BSS problem, we briefly restate the method of maximum-likelihood estimation (Pham and Garrat, 1997; Cardoso, 1997; Hyvärinen et al., 2001), because it is a fundamental method of statistical estimation and a unique framework for a variety of algorithms. Loosely speaking, in maximum-likelihood estimation we answer the question: Given a certain probability distribution model, what is the most likely set of parameters that would have generated the observed data?

In the language of the ICA problem, the ML principle is the following: Given the observation vector $\mathbf{x}$, maximize the (log-)likelihood function of the mixing matrix $\mathbf{A}$ or, equivalently, the demixing matrix $\mathbf{V} = \mathbf{A}^{-1}$. As a first step we need to derive the likelihood function of the demixing matrix in the ICA model. The pdf of the observations $\mathbf{x}$ is:

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= |\det \mathbf{V}|\ p_{\mathbf{s}}(\mathbf{s}) \\ &= |\det \mathbf{V}| \prod_{i=1}^{N} p_{s_i}(s_i) \\ &= |\det \mathbf{V}| \prod_{i=1}^{N} p_{s_i}(\mathbf{v}_i \mathbf{x}) \end{aligned} \tag{2.5}$$

where we used the statistical independence of the marginal components s_i and the fact that $\frac{1}{\det \mathbf{A}} = \det \mathbf{A}^{-1} = \det \mathbf{V}$.

We consider a fixed sample set $\mathbf{X}$ with T independent samples to obtain the *likelihood* function:

[1] MI is an important concept of information theory and has many useful properties (Cover and Thomas, 1991).

$$\ell(\mathbf{V}) = \prod_{t=1}^{T} |\det \mathbf{V}| \prod_{i=1}^{N} p_{s_i}(\mathbf{v}_i \mathbf{x}) \tag{2.6}$$

$$= (|\det \mathbf{V}|)^T \prod_{t=1}^{T} \prod_{i=1}^{N} p_{s_i}(\mathbf{v}_i \mathbf{x}), \tag{2.7}$$

where $\mathbf{v}_i$ is the i-th row of $\mathbf{V}$.

For practical optimization is preferable to use the normalized minus-log-likelihood function:

$$\mathcal{L}(\mathbf{V}) = -\log \ell(\mathbf{V}) = -\log |\det \mathbf{V}| + \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} -\log p_{s_i}(\mathbf{v}_i \mathbf{x}) \tag{2.8}$$

Unfortunately, we can not use equation (2.8) immediately, because we do not know the pdf of the sources. Therefore one resorts to a *quasi maximum-likelihood* approach by choosing a specific "contrast" function h which approximates the negative logarithm of the unknown source densities $p_{s_i}(s_i)$.

Depending on the choice of $h(\cdot)$ one may only yield approximate statistical independence of the variables but it can be shown that for a broad class of functions one yields sufficient conditions to solve the BSS problem. For example, it has been shown by Zibulevsky (2003) that this approach is highly efficient, if the sources are sparse or sparsely representable. In this case the absolute value function is a good choice for $h(\cdot)$.

Finally, we have to solve the following nonlinear optimization problem:

$$\min_{\mathbf{V}} \mathcal{L}(\mathbf{V}; \mathbf{x}, h) \tag{2.9}$$

Minimizing this function w.r.t. $\mathbf{V}$ by a suitable numerical method means to solve the BSS problem (see Fig. 2.2).

2.3 Joint Diagonalization Approach

As pioneered by the works of Comon (1994); Molgedey and Schuster (1994); de Lathauwer (1997); Belouchrani et al. (1997); Cardoso (1999), we aim to make use of the notion of approximate joint diagonalization (AJD) to solve the BSS problem in a unifying framework. As the leitmotiv we want to replace the measure of independence by a measure of diagonality of a set of matrices. This will also provide us a flexible framework for generic BSS algorithms where the numerical optimization part can be treated efficiently in a separate step.

2.3.1 From BSS to AJD

It is straightforward to see that under the assumption of the linear, instantaneous BSS model (2.1) there exists a certain set of (unknown) "target matrices" which, in theory, gives rise to an joint diagonalization problem. For example, as proposed in Molgedey and Schuster (1994), we may consider (spatial) covariance matrices $\mathbf{C}_\tau(\mathbf{x})$ of time-lagged mixed signals $\mathbf{x}(t)$,

$$\mathbf{C}_\tau(\mathbf{x}) \stackrel{\text{def}}{=} \mathrm{E}\{\mathbf{x}(t)\mathbf{x}(t+\tau)^T\}$$

where the expectation is taken over t and τ is a time-shift parameter, we see that the covariance of $\mathbf{x}$ is related to the covariance of $\mathbf{s}$ according to

$$\begin{aligned} \mathbf{C}_\tau(\mathbf{x}) &= \mathrm{E}\{(\mathbf{A}\mathbf{s}(t))(\mathbf{A}\mathbf{s}(t+\tau))^T\} \\ &= \mathbf{A}\ \mathrm{E}\{\mathbf{s}(t)\mathbf{s}(t+\tau)^T\}\,\mathbf{A}^T \\ &= \mathbf{A}\,\mathbf{C}_\tau(\mathbf{s})\,\mathbf{A}^T \end{aligned} \tag{2.10}$$

due to the linearity of the expectation operator and the mixing model.

The key observation is that all cross-correlation terms which are the off-diagonal elements of $\mathbf{C}_\tau(\mathbf{s})$ are zero for independent signals and thus $\mathbf{C}_\tau(\mathbf{s})$ is a diagonal matrix. Hence the mixing matrix $\mathbf{A}$ can be identified as the solution of a matrix diagonalization problem in equation (2.10). If $\mathbf{A}$ is invertible, this can also be written as

$$\mathbf{V}\,\mathbf{C}_\tau(\mathbf{x})\,\mathbf{V}^T = \mathbf{C}_\tau(\mathbf{s}) = \mathbf{D}_\tau, \tag{2.11}$$

where the matrix $\mathbf{V} = \mathbf{A}^{-1}$ is diagonalizing all $\mathbf{C}_\tau(\mathbf{x})$ simultaneously. In practice, the target matrices $\mathbf{C}_\tau(\mathbf{x})$ have always to be estimated from the available data with a finite sample size T. Typically the expectation is computed as a sample average

$$\hat{\mathbf{C}}_\tau(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T-\tau}\left(\mathbf{x}(t)\mathbf{x}(t+\tau)^T\right). \tag{2.12}$$

It is clear that this procedure inevitably gives rise to estimation errors, however, the pragmatic approach is to assume that these errors can be neglected for sufficiently large T. Thus for large T we conclude analogously to (2.10) that

$$\hat{\mathbf{C}}_\tau(\mathbf{x}) = \mathbf{A}\hat{\mathbf{C}}_\tau(\mathbf{s})\mathbf{A}^T.$$

This means that the BSS problem has been translated into an equivalent problem of approximate joint diagonalization.

2.3.2 Symmetrizing

We note that the matrices $\mathbf{C}_\tau(x)$ are not symmetric by construction, however it is appropriate to symmetrize them, because under the ICA model the anti-symmetric part is assumed to be zero and thus the diagonalization problem can be fully based on the symmetric part of $\mathbf{C}_\tau(\mathbf{x})$:

$$(\mathbf{C}_\tau(\mathbf{x}) + \mathbf{C}_\tau(\mathbf{x})^T) \quad = \quad \mathbf{A}\,\underbrace{(\mathbf{C}_\tau(\mathbf{s}) + \mathbf{C}_\tau(\mathbf{s})^T)}_{\downarrow}\,\mathbf{A}^T \tag{2.13}$$

$$\mathbf{V}\ \ (\mathbf{C}_\tau(\mathbf{x}) + \mathbf{C}_\tau(\mathbf{x})^T)\ \ \mathbf{V}^T = \qquad\qquad \mathbf{D}_\tau \tag{2.14}$$

2.3.3 A Two-Stage Algorithm

From (2.14) we conclude that finding a transformation matrix $\mathbf{V}$ which diagonalizes the estimated, symmetrized target set "as good as possible" provides us an estimate of the demixing matrix.

A typical algorithm proceeds in two stages, "sphering" and "rotation" (see Figure 2.3).

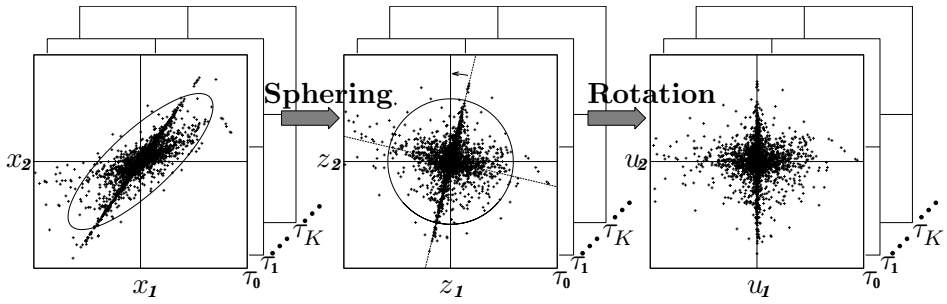

Figure 2.3: Decorrelation method using temporal structure of the signals

Sphering (or whitening) is aimed at orthogonalizing the observed signals in a new coordinate system, i.e. the goal is to transform the data such that they have unit covariance (Fukunaga, 1990). By transforming the observation vector with $\mathbf{Q} = \sqrt{\mathbf{C}_0^{-1}(\mathbf{x})}$ we obtain

$$\mathbf{z}(t) = \mathbf{Q}\mathbf{x}(t) = \mathbf{Q}\mathbf{A}\mathbf{s}(t),$$

and

$$\mathbf{C}_0(\mathbf{z}) = (\mathbf{Q}\mathbf{A})\mathbf{C}_0(\mathbf{s})(\mathbf{Q}\mathbf{A})^\top = \mathbf{I}.$$

Since $\mathbf{C}_0(\mathbf{s}) = \mathbf{I}$, the product $(\mathbf{Q}\mathbf{A})$ is an orthogonal matrix:

$$(\mathbf{Q}\mathbf{A})(\mathbf{Q}\mathbf{A})^\top = \mathbf{I}.$$

It is easily seen that rotating the signals $\mathbf{z}(t)$ with any orthogonal matrix $\mathbf{B}$ will not change the covariance matrix:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(\mathbf{Bz}(t))(\mathbf{Bz}(t))^\top &= \mathbf{B}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}(t)\mathbf{z}(t)^\top\right)\mathbf{B}^\top \\ &= \mathbf{BIB}^\top \\ &= \mathbf{I}.\end{aligned}$$

Thus after applying the sphering transform there remains an ambiguity of rotation. The correct rotation can be determined by minimizing the off-diagonal elements of several time-delayed correlation matrices with a joint-diagonalization algorithm (Belouchrani et al., 1997; Ziehe and Müller, 1998).

2.3.4 Generic Algorithm for BSS

In this thesis we propose to extend this approach and use the joint diagonalization technique as an "engine" of a generic BSS method. We will show in section 2.3.5 that there are many more possibilities to define target matrices that have the same property as the covariance matrices above, i.e. matrices which are diagonal for the source signals and 'similar to diagonal' for the observed mixtures.

In algorithm 1 we outline our generic procedure for BSS based on approximate joint diagonalization of a set of matrices:

Algorithm 1 The AJD4BSS algorithm

INPUT: $\mathbf{x}(t)$
$\hat{\mathbf{C}}_k = \ldots$ {Estimate a number of matrices $\mathbf{C}_k(\mathbf{x})$}
$\mathbf{V} = \mathrm{AJD}(\hat{\mathbf{C}}_k)$ {Apply joint diagonalization method}
$\mathbf{u}(t) = \mathbf{Vx}(t)$ {unmix signals}
OUTPUT: $\mathbf{u}(t)$, $\mathbf{V}$

In order to implement this method we need two things:

- an estimation procedure for suitable matrices $\mathbf{C}_k$
- a joint-diagonalization algorithm

In chapter 3 we will present in detail the different strategies to solve the AJD problem and in the remainder of the next section we review established choices for $\mathbf{C}_k$ available in the rich BSS literature.

2.3.5 Possible Target Matrices

In this section we catalog different choices for the set of target matrices. The main purpose is to give a "cookbook- like" overview how particular properties of the involved signals are used to construct such matrices from the data and to show the connection to existing BSS algorithms.

Historically, the suitability of joint diagonalization criteria as BSS cost functions have been realized by several authors. Pioneered in the early works of Tong et al. (1991); Comon (1994); Molgedey and Schuster (1994); Matsuoka et al. (1995); Laheld and Cardoso (1996); Belouchrani et al. (1997) related methods emerged in articles of Wu and Principe (1999); Hori (1999); Pham and Cardoso (2001); Yeredor (2002). Interestingly, all of the "three easy routes" to ICA pointed out in Cardoso (2001) have analogous specific definitions of the set of target matrices $\{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K\}$. The basic idea is to exploit certain 'non–properties' of the signals. The three most often used properties of this kind are:

- non-Gaussianity
- non-Stationarity
- spectral non-Flatness

Non-Gaussianity

If we assume that the source processes $s_i(t)$ are independent identically distributed, the BSS method has to rely on the non-Gaussianity of the sources. Non-Gaussianity means that the source density is sufficiently different from a Gaussian density. The key point is that by mixing the signals become more and more Gaussian, since the distribution of the sum of many independent random variables tends to be Gaussian (Hyvärinen et al., 2001).

- **Cumulant-based method**

  Non-Gaussianity can be exploited with higher-order statistics. In order to obtain suitable target matrices we resort to cumulants. Theoretically, cumulants are defined as the coefficients of the Taylor expansion of the logarithm of the joint characteristic function $\mathbf{\Phi}(\omega_1, \ldots, \omega_N)$ at the origin $\boldsymbol{\omega} = \mathbf{0}$. In practice, however, cumulants are computed from higher-order moments which are estimated from the data (Comon, 1994; Cardoso, 1999). The fourth-order cumulant tensor is a four-way array:

$$\begin{aligned} \operatorname{cum}(x_i, x_j, x_k, x_l) &= \mathrm{E}\{x_i x_j x_k x_l\} - \mathrm{E}\{x_i x_j\}\,\mathrm{E}\{x_k x_l\} \\ &\quad - \mathrm{E}\{x_i x_k\}\,\mathrm{E}\{x_j x_l\} - \mathrm{E}\{x_i x_l\}\,\mathrm{E}\{x_j x_k\} \end{aligned} \tag{2.15}$$

The matrices to be diagonalized can be obtained from 'parallel slices' of the fourth-order cumulant tensor:

$$\mathbf{C}_{(ij)}(\mathbf{M}) = \sum_{kl} \mathbf{M}_{(kl)} \operatorname{cum}(x_i, x_j, x_k, x_l), \tag{2.16}$$

where $\mathbf{M}$ is an arbitrary matrix (see also Hyvärinen et al. (2001)).

The popular JADE[2] algorithm (Cardoso and Souloumiac, 1993) belongs to this class. After whitening the data, JADE performs an approximate diagonalization of the set of eigen-matrices of the cumulant tensor with an orthogonal transformation composed of a sequence of plane rotations (Cardoso and Souloumiac, 1993; Comon, 1994).

The plane rotation $\mathbf{R}(\theta; i, j)$ is defined as the identity matrix where the (i, i) and (j, j) entries are replaced by $\cos(\theta)$ and the (i, j) entry is replaced by $-\sin(\theta)$ and the (j, i) entry is replaced by $\sin(\theta)$. Then for each pair (i, j) one computes the optimal angle θ which mininizes the cost function.

Due to the high computational load for storing and processing the fourth-order order cumulants the application of this method often requires a dimension reduction.

- **CHESS**

  An interesting alternative for exploiting non-Gaussianity has been proposed in Yeredor (2000). There he realized that using the coefficients of an Taylor expansion of the logarithm of the joint characteristic function $\mathbf{\Phi}(\omega_1, \ldots, \omega_N)$ not at the origin $\boldsymbol{\omega} = \mathbf{0}$ but for some *off the origin* processing points $\boldsymbol{\omega}_k$ gives rise to a set of K target matrices for joint diagonalization.

  In the CHESS[3] algorithm these are defined as specially weighted empirical covariances (second-order statistics).

$$\mathbf{C}(\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{\sum_{t=1}^{T} \lambda_t} \sum_{t=1}^{T} \lambda_t [\mathbf{x}(t) - \bar{\mathbf{x}}][\mathbf{x}(t) - \bar{\mathbf{x}}]^T \tag{2.17}$$

  where $\lambda_t = e^{\boldsymbol{\omega}^T \mathbf{x}(t)}$ and where $\bar{\mathbf{x}} = \sum \lambda_t \mathbf{x}(t) / \sum \lambda_t$.

  The set of target matrices $\{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K\}$ is constructed by choosing different processing points $\boldsymbol{\omega}_k \in I\!R^N$

$$\mathbf{C}_k = \mathbf{C}(\mathbf{x}, \boldsymbol{\omega}_k).$$

[2] JADE stands for Joint Approximate Diagonalization of Eigen-matrices.
[3] CHESS stands for CHaracteristic function Enabled Source Separation

The advantage of this method is that the induced computational load and the statistical robustness are more favorable than for the above cumulant method.

Non-Stationarity

The i.i.d.assumption is often too restrictive and it is useful to exploit possible temporal structure. Then we can even recover Gaussian sources. A very simple form of temporal structure is related to non-stationarity. Here we assume that the variance σ^2 of the sources is not constant over time, but varies according to some "amplitude profile" $\sigma(t)$. Furthermore the variation is assumed to be relatively slow. Thus we rely on the following properties:

- signals are supposed to be stationary within a short time-scale,
- and signals are intrinsically non-stationary over the long run.

In this case the set of target matrices is constructed from the empirical covariance matrix in different segments of the data (Matsuoka et al., 1995; Pham and Garrat, 1997; Parra and Spence, 2000; Pham and Cardoso, 2000, 2001; Choi et al., 2001).

Non-Flatness

The second case of non-i.i.d.sources originates from time dependencies. These data exhibit broad band power spectra that are not constant over the frequencies (non-flat spectra).

It is interesting to note that most 'natural' signals, like speech signals or neurophysiological signals (EEG, MEG, etc.) have a rich dynamical time structure. Hence we want to directly exploit their diversity in the time-frequency domain for blind source separation.

Examples for possible target matrices in this class are time-lagged covariances (cf. Tong et al., 1991; Molgedey and Schuster, 1994; Belouchrani et al., 1997; Ziehe and Müller, 1998), where the respective auto– and cross–correlation functions $\phi_{x_i,x_j}(\tau) = \mathrm{E}\{x_i(t)x_j(t-\tau)\}$ in are arranged in matrix form:

$$\mathbf{C}_\tau(\mathbf{x}) = \begin{bmatrix} \phi_{x_1,x_1}(\tau) & \cdots & \phi_{x_1,x_N}(\tau) \\ \phi_{x_2,x_1}(\tau) & \cdots & \phi_{x_2,x_N}(\tau) \\ \vdots & \ddots & \vdots \\ \phi_{x_N,x_1}(\tau) & \cdots & \phi_{x_n,x_N}(\tau) \end{bmatrix}. \tag{2.18}$$

In (Ziehe et al., 2000b), the $\mathbf{C}_k$ are defined as:

$$\mathbf{C}_k(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}(t)\ \left(\boldsymbol{h}_k \star \mathbf{x}(t)\right)^T, \tag{2.19}$$

where $\star$ denotes convolution and $\boldsymbol{h}_k$ are impulse responses of linear filters.

The time-lagged correlation matrices of equation (2.18) are a special case of (2.19), where one uses the linear filter: $\boldsymbol{h}_k = \delta_{t\tau_k}$; i.e. the filter is parametrized by a single parameter, the time-shift τ.

A further generalization is to use time-frequency distributions (TFD) to obtain target matrices for joint diagonalization (Belouchrani and Amin, 1998; Pham and Cardoso, 2001; Pham, 2002).

In (Pham, 2002) the following method is proposed. First, one estimates the time-domain covariance function using a sliding window w with $\sum_l w^2(l) = 1$ and computes:

$$\hat{\mathbf{R}}_X(t,\tau) \stackrel{\text{def}}{=} \sum_l [w(l-t)\mathbf{x}(l)][w(l-\tau-t)\mathbf{x}(l-\tau)]^T \tag{2.20}$$

Then, in a second step, the instantaneous spectral density matrix is estimated from the Fourier transform of the (instantaneous) covariance function $\hat{\mathbf{R}}_X(t,\tau)$, smoothed with a suitable kernel $k(\tau)$ (e.g. a Parzen window).

$$\hat{f}_{\mathbf{x}}(t,\omega) \stackrel{\text{def}}{=} \frac{1}{2\pi} \sum_\tau k(\tau)\hat{\mathbf{R}}_X(t,\tau)e^{i\omega\tau} \tag{2.21}$$

The matrices $\mathbf{C}_k$ are obtained using (2.21) in a quite flexible way by tiling the time-frequency-plane into (overlapping) blocks and computing one target matrix per block: $\mathbf{C}_k = \hat{f}_{\mathbf{x}}(t_k, \omega_k)$

2.4 Summary and Conclusion

There are very efficient BSS algorithms based on relatively weak assumptions, like vanishing spatio-temporal cross-correlations instead of full statistical independence, and those methods can be implemented in a unified framework of simultaneous diagonalization of several, appropriately defined matrices. These algebraic methods, provide both a computationally efficient and generally applicable framework for BSS.

Chapter 3

Approximate Joint Diagonalization of Matrices

In this chapter we address the problem of approximate joint diagonalization (AJD) of several real-valued, symmetric matrices. In the previous chapter (section 2.3) we have seen that AJD provides a general framework for handling generic BSS problems. In the following sections, we introduce the joint diagonalization concept and derive new, computationally efficient algorithms implementing these ideas. This chapter is mainly based on the publications (Ziehe et al., 2003c, 2004) and (Yeredor, Ziehe and Müller, 2004).

3.1 Introduction

Joint diagonalization of square matrices is an important general problem of numeric computation. Besides other applications, joint diagonalization techniques provide a generic algorithmic tool for blind source separation. In this chapter the joint diagonalization problem is formulated and state-of-the-art approaches for its solution are reviewed.

The main part of this chapter is devoted to the derivation of new algorithms to efficiently perform a joint diagonalization of several symmetric matrices. The general structure of these algorithms will be based on a multiplicative update with a matrix exponential in order to constrain the solution to a particular manifold. We will see that the use of such matrix exponential update enables the development of efficient optimization algorithms.

Before we come to the joint diagonalization problem we recall some basic facts from linear algebra.

Eigenvalues, Eigenvectors, Diagonalization

A matrix $\mathbf{D}$ is diagonal if $D_{ij} = 0$ whenever $i \neq j$.

The notion of diagonalizing a matrix $\mathbf{M}$ is closely related to solving an eigenvalue problem. In matrix form the eigenvalue problem is: Given an $N \times N$ matrix $\mathbf{M}$, find a $N \times N$ matrix $\mathbf{E}$ and a diagonal $N \times N$ matrix $\mathbf{D}$, such that

$$\mathbf{ME} = \mathbf{ED}. \tag{3.1}$$

If $\mathbf{M}$ is symmetric and real-valued, then a solution always exists where $\mathbf{E}$ is an orthogonal matrix (i.e. $\mathbf{EE}^T = \mathbf{I}$) consisting of the eigenvectors and the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{M}$.

Thus writing $\mathbf{M} = \mathbf{EDE}^T$, where $\mathbf{E}$ is orthogonal and $\mathbf{D}$ is diagonal, gives a factorization known as the eigenvalue decomposition (EVD).

Generalized Eigenvalue Problem

Also for two normal matrices, it is well known that exact joint diagonalization is possible and is referred to as the generalized eigenvalue problem: Given two $N \times N$ matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, find a $N \times N$ matrix $\mathbf{E}$ and a diagonal $N \times N$ matrix $\mathbf{D}$, such that

$$\mathbf{M}_1\mathbf{E} = \mathbf{M}_2\mathbf{ED}. \tag{3.2}$$

If $\mathbf{M}_2$ is non-singular, the problem (3.2) can be reduced to problem (3.1) by multiplying (3.2) from the left with $\mathbf{M}_2^{-1}$. Extensive literature exists on this topic (e.g. Noble and Daniel, 1977; Golub and van Loan, 1989; Bunse-Gerstner et al., 1993; Vorst and Golub, 1997, and references therein).

Approximate Joint Diagonalization

In general, it is not possible to diagonalize more than two matrices with one single transformation. However, exact diagonalization of more than two matrices is possible if the matrices possess a certain common structure, as it is the case for the blind source separation application (see e.g. equation (2.11)) . If the ideal model of equation (2.11) holds, exact joint diagonalization is possible, otherwise, one can only speak of approximate joint diagonalization (AJD). In the remainder of the chapter we will use the terms "approximate joint diagonalization" and "joint diagonalization" interchangeably for diagonalization of *more than two* matrices with a single transformation. The approximation is understood in the sense of minimizing a suitable diagonality criterion.

Many algorithms for joint diagonalization have been previously proposed (e.g. Flury and Gautschi, 1986; Cardoso and Souloumiac, 1993, 1996; Hori, 1999; Pham, 2001; van der Veen, 2001; Yeredor, 2002; Joho and Rahbar, 2002).

In order to better understand the challenges of AJD and to explore possible directions to improve the existing algorithms, we first study some existing approaches to solve the joint diagonalization problem.

3.2 Solving the Joint Diagonalization Problem

3.2.1 Three Cost Functions

In this section we consider the approximate joint diagonalization of a of real-valued symmetric matrices of size $N \times N$.[1] The goal of a joint diagonalization algorithm is to find a matrix $\mathbf{V}$ that simultaneously transforms $\mathbf{C}_1, \ldots, \mathbf{C}_K$ as good as possible to diagonal form. The notion of closeness to diagonality and the corresponding formal statement of the joint diagonalization problem can be defined in different ways:

1. *Subspace fitting formulation.*

   Approximate joint diagonalization consists of the following optimization problem (van der Veen, 2001; Yeredor, 2002): Given a set of K $N \times N$ "target matrices" $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K$, find a $N \times N$ matrix $\mathbf{A}$ and K diagonal matrices $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K$ such that the following quantity is minimized:

   $$J_1(\mathbf{A}, \mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K) = \sum_{k=1}^{K} ||\mathbf{C}_k - \mathbf{A}\mathbf{D}_k\mathbf{A}^T||_F^2 \tag{3.3}$$

   where $||\cdot||$ denotes the squared Frobenius norm.

2. *Frobenius norm formulation.*

   This formulation of the joint diagonalization problem has been used most frequently in the literature, e.g. in Bunse-Gerstner et al. (1993); Cardoso and Souloumiac (1993, 1996); Hori (1999); Joho and Rahbar (2002); Joho and Mathis (2002). Here, the goal is to find the inverse $\mathbf{V} = \mathbf{A}^{-1}$ of the matrix $\mathbf{A}$, by minimizing the diagonality criterion:

   $$J_2(\mathbf{V}) = \sum_{k=1}^{K} \text{off}(\mathbf{V}\mathbf{C}_k\mathbf{V}^T) \tag{3.4}$$

   where the off$(\cdot)$ is the Frobenius norm of the off-diagonal elements,

   $$\text{off}(\mathbf{M}) \stackrel{\text{def}}{=} ||\mathbf{M} - \text{diag}(\mathbf{M})||_F^2 = \sum_{i \neq j} (M_{ij})^2. \tag{3.5}$$

[1] The formulations and the proposed algorithms will be presented for real-valued, symmetric matrices only. We note however, that extensions to the complex-valued case could be obtained in a similar manner.

While the diagonality measure (3.5) appears very natural and intuitive, the cost function (3.4) has the disadvantage that the Frobenius norm is obviously minimized by the trivial solution $\mathbf{V} = \mathbf{0}$. Therefore an optimization method with an additional constraint to exclude the zero solution is required.

3. *Positive definite formulation.* Often it is reasonable to assume that in the initial problem all the matrices $\mathbf{C}_k$ are positive-definite. This assumption is motivated by the fact that in many applications matrices $\mathbf{C}_k$ are covariance matrices of some random variables. In this case, as proposed in Matsuoka et al. (1995); Pham (2001) the criterion

$$J_3(\mathbf{V}) \stackrel{\text{def}}{=} \log \det(\operatorname{diag}(\mathbf{V}\mathbf{C}_k\mathbf{V}^T)) - \log \det(\mathbf{V}\mathbf{C}_k\mathbf{V}^T) \tag{3.6}$$

can be used instead of the cost function (3.5). Here the operator $\operatorname{diag}(\mathbf{M})$ returns a diagonal matrix containing only the diagonal entries of $\mathbf{M}$.

This measure can be traced back to information theory as the Kullback-Leibler distance between a Gaussian process with covariance matrix $\mathbf{C}$ and its diagonal part $\operatorname{diag}(\mathbf{C})$. The additional advantage of this criterion is its scale invariance (Pham and Cardoso, 2001).

However, in certain applications, the matrices are not always guaranteed to be positive-definite. For example in blind source separation based on time-delayed decorrelation (Belouchrani et al., 1997; Ziehe and Müller, 1998), correlations can be positive or negative and in this case the criterion J_3 can not be used.

Compared to the approaches 2. and 3., the algorithms based on subspace fitting have two advantages: they do not require orthogonality, positive-definiteness or any other normalizing assumptions on the matrices $\mathbf{A}$ and $\mathbf{C}_k$, and they are able to handle non-square mixture matrices. These advantages, however, come at the price of a high computational cost: the algorithm of van der Veen (2001) has quadratic convergence in the vicinity of the minimum, but its running time per iteration is $\mathcal{O}(KN^6)$; the AC-DC algorithm of Yeredor (2002) converges linearly with a running time per iteration of order $\mathcal{O}(KN^3)$.

The algorithms relying on the positive-definiteness assumption are also efficient thanks to the favorable invariance properties, but they fail for non-positive-definite matrices. Least-squares subspace fitting algorithms, which do not require such strong a-priori assumptions, are computationally much more demanding. Our work is motivated by the question: could we develop a method that combines all the good features and at the same time avoids the shortcomings of the previous joint diagonalization algorithms?

3.2.2 Our Approach

In our approach we want to employ the Frobenius off-diagonal norm formulation and minimize the cost function J_2. For this we have to solve a constrained non-linear optimization problem, i.e. essentially a quadratic least-squares problem with an constraint that prevents the algorithm from converging to the trivial solution.

One of the standard algorithms for solving nonlinear least-squares problems is for example the Levenberg-Marquardt (LM) algorithm (Levenberg, 1944; Marquardt, 1963). However, the LM algorithm cannot be directly applied to our problem, because the classical LM algorithm does not provide means for incorporation of additional constraints, such as orthogonality or invertibility of the diagonalizer $\mathbf{V}$. In what follows we present a different strategy how to cope with the *constrained* optimization problem which naturally incorporates the additional structure of our problem into the algorithm.

The key point is to make use of matrix exponential updates and to exploit the fact that our parameter space consists of the group of orthogonal, or more general, invertible matrices.

3.2.3 General Structure of Our Algorithm

Our goal is to solve the following constrained non-linear optimization problem:

$$\min_{\mathbf{V}} \sum_{k=1}^{K} \sum_{i \neq j} ((\mathbf{V}\mathbf{C}_k\mathbf{V}^T)_{ij})^2. \tag{3.7}$$

Due to the non-linearity of the problem, we can not obtain a solution for $\mathbf{V}$ in closed form. Instead we have to use an iterative scheme to successively improve an initial solution. The main problem with this approach is however to avoid the trivial solution $\mathbf{V} = \mathbf{0}$ which poses a constraint to (3.7).

In the traditional approach one would introduce a penalty term which has a minimum when an additional normalization constraint is satisfied. For example, Joho and Rahbar (2002) proposed to use:

$$J_4 = ||\mathbf{V}\mathbf{V}^T - \mathbf{I}||_F \tag{3.8}$$

$$J_5 = ||\operatorname{diag}(\mathbf{V} - \mathbf{I})||_F \tag{3.9}$$

In contrast, we prefer to use the group structure of the search space as a hard constraint preventing convergence of the minimizer of the cost function in Equation (3.7) to the zero solution.

We propose the following iterative process. We start with a matrix $\mathbf{V}^{(0)}$ that belongs to the group and carry out multiplicative updates:

$$\mathbf{V}^{(m+1)} \leftarrow \exp(\mathbf{W}^{(m+1)})\, \mathbf{V}^{(m)}, \tag{3.10}$$

where $\exp(\cdot)$ denotes the matrix exponential and $\mathbf{V}^{(m+1)}$ the estimated diagonalizing (demixing) matrix after the $(m+1)$-th iteration (see Figure 3.1). The new parameter $\mathbf{W}^{(m+1)}$ of the update multiplier is to be determined so as to minimize the cost function (3.7).

For the update it is also important that the matrix $\mathbf{V}^{(m+1)}$ remains always within the group manifold. This can be ensured by certain conditions on $\mathbf{W}^{(m+1)}$ and will be discussed in subsection 3.2.4.

Pseudo-code summarizing the matrix exponential update method for performing approximate joint diagonalization is outlined in Algorithm 2.

Algorithm 2 Matrix Exponential Updates for AJD.

INPUT: $\mathbf{C}_k^{(0)}$ { Matrices to be diagonalized}
$\mathbf{W}^{(0)} \leftarrow 0$, $\mathbf{V}^{(0)} \leftarrow \mathbf{I}$, $m \leftarrow 0$
repeat
  compute $\mathbf{W}^{(m+1)}$ from $\mathbf{C}_k^{(m)}$ according to Equation (3.18) or (3.26) or (3.27)
  if $||\mathbf{W}^{(m+1)}||_F > \theta$ **then**
    $\mathbf{W}^{(m+1)} \leftarrow \frac{\theta}{||\mathbf{W}^{(m+1)}||_F}\mathbf{W}^{(m+1)}$
  end if
  $\mathbf{V}^{(m+1)} \leftarrow \exp(\mathbf{W}^{(m+1)})\,\mathbf{V}^{(m)}$
  $\mathbf{C}_k^{(m+1)} \leftarrow \mathbf{V}^{(m+1)}\,\mathbf{C}_k^{(0)}\,(\mathbf{V}^{(m+1)})^T$
  $m \leftarrow m+1$
until convergence
OUTPUT: $\mathbf{V}^{(m+1)}, \mathbf{C}_k^{(m+1)}$

Such an multiplicative update scheme is rarely used in classical optimization algorithms; however, it is common for many successful BSS algorithms, such as relative-gradient (Laheld and Cardoso, 1996; Amari et al., 2000), relative Newton (Akuzawa and Murata, 2001; Zibulevsky, 2003), as well as for some previous joint diagonalization methods (Cardoso and Souloumiac, 1996; Pham, 2001). A further important feature of this approach is the use of the matrix exponential which, to the best of our knowledge, has only been used in Akuzawa and Murata (2001).

3.2.4 Structure Preserving Updates

In order to derive algorithms that take the group structure of our parameter space into account, we consider repeated multiplicative updates (see Figures 3.3 and 3.1). The fundamental concept is the matrix exponential function.

The Exponential Map

The exponential of a real valued square matrix $\mathbf{M}$, denoted by $e^{\mathbf{M}}$ or $\exp(\mathbf{M})$, is defined as

$$\begin{aligned} e^{\mathbf{M}} = \exp(\mathbf{M}) &= \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{M}^k \\ &= \mathbf{I} + \mathbf{M} + \frac{\mathbf{M}^2}{2!} + \dots \end{aligned} \tag{3.11}$$

The matrix exponential satisfies the following properties:

1. For the $N \times N$ zero matrix O, $e^0 = \mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix.
2. If $\mathbf{M} = \mathbf{Q} \begin{pmatrix} D_1 & & \\ & \ddots & \\ & & D_N \end{pmatrix} \mathbf{Q}^{-1}$ for an invertible $N \times N$ matrix $\mathbf{Q}$, then $e^{\mathbf{M}} = \mathbf{Q} \begin{pmatrix} e^{D_1} & & \\ & \ddots & \\ & & e^{D_N} \end{pmatrix} \mathbf{Q}^{-1}$.
3. If $\mathbf{M}'$ is a matrix of the same type as $\mathbf{M}$, and $\mathbf{M}$ and $\mathbf{M}'$ commute, then $e^{\mathbf{M}+\mathbf{M}'} = e^{\mathbf{M}} e^{\mathbf{M}'}$.
4. The trace of $\mathbf{M}$ and the determinant of $e^{\mathbf{M}}$ are related by the formula $\det e^{\mathbf{M}} = e^{\operatorname{tr} \mathbf{M}}$. Therefore $e^{\mathbf{M}}$ is always invertible and the inverse is $(e^{\mathbf{M}})^{-1} = e^{-\mathbf{M}}$.

In applications of approximate joint diagonalization to the BSS problem we are mainly interested in two special cases where the transformation $\mathbf{V}$ is assumed to be:

- orthogonal, i.e. $\mathbf{V} \in \mathrm{O(N)}$ or
- invertible, i.e. $\mathbf{V} \in \mathrm{GL(N)}$.

The multiplicative matrix exponential update preserves these important features, because the product of orthogonal (invertible) matrices is an orthogonal (invertible) matrix.

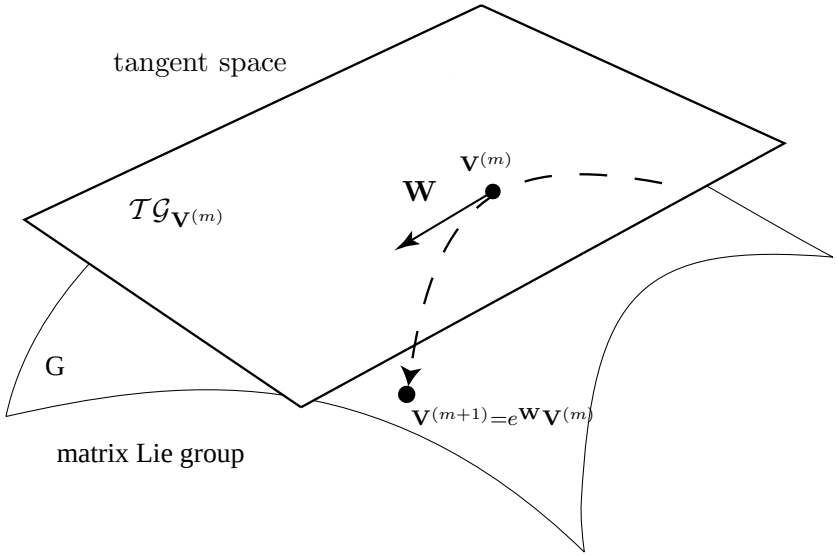

Figure 3.1: Illustration of multiplicative updates using the matrix exponential map: a local step in the tangent space is mapped to a distinct point on the manifold.

Orthogonal case

The most popular structural assumption is orthogonality of $\mathbf{V}$. In fact, such an assumption seems natural if the joint diagonalization problem is seen as an extension of the eigenvalue problem and one asks for the common orthogonal basis of several matrices.

A further reason for the importance of this special case is the fact that we can use a sphering step as a pre-processing (see also section 2.3.3). The sphering can be done as follows: First, we pick one positive-definite symmetric matrix $\mathbf{C}$ from the set $\mathcal{M}$. Then, the sphering transform $\mathbf{Q}$ is defined by the matrix that is obtained as the inverse square root of $\mathbf{C}$.

This matrix can be computed via the eigenvalue decomposition of $\mathbf{C}$, $\mathbf{C} = \mathbf{E}\mathbf{D}\mathbf{E}^T$, which implies:

$$\mathbf{Q} \stackrel{\text{def}}{=} \mathbf{C}^{-\frac{1}{2}} = (\mathbf{E}\,\mathbf{D}\,\mathbf{E}^T)^{-\frac{1}{2}} = \mathbf{E}\,\mathbf{D}^{-\frac{1}{2}}\,\mathbf{E}^T.$$

To see that this is indeed a sphering transformation, we apply $\mathbf{Q} = \mathbf{C}^{-\frac{1}{2}}$ to $\mathbf{C} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^T$, where $\mathbf{E}$ is orthogonal. This yields

$$\mathbf{Q}\mathbf{C}\mathbf{Q}^T = \mathbf{E}\,\mathbf{D}^{-\frac{1}{2}}\,\mathbf{E}^T\mathbf{E}\,\mathbf{D}\,\mathbf{E}^T\mathbf{E}\,\mathbf{D}^{-\frac{1}{2}}\,\mathbf{E}^T = \mathbf{I}.$$

This transformation is illustrated in Fig. 3.2 for a 2×2 positive-definite matrix.

In general, the simultaneously diagonalizable matrices $\mathbf{C}_k$ can be written in the form $\mathbf{A}\mathbf{D}_k\mathbf{A}^T$, where $\mathbf{A}$ is a non-orthogonal invertible matrix and $\mathbf{D}_k$ are diagonal matrices. Applying the sphering transform to $\mathbf{C}_k = \mathbf{A}\,\mathbf{D}_k\,\mathbf{A}^T$

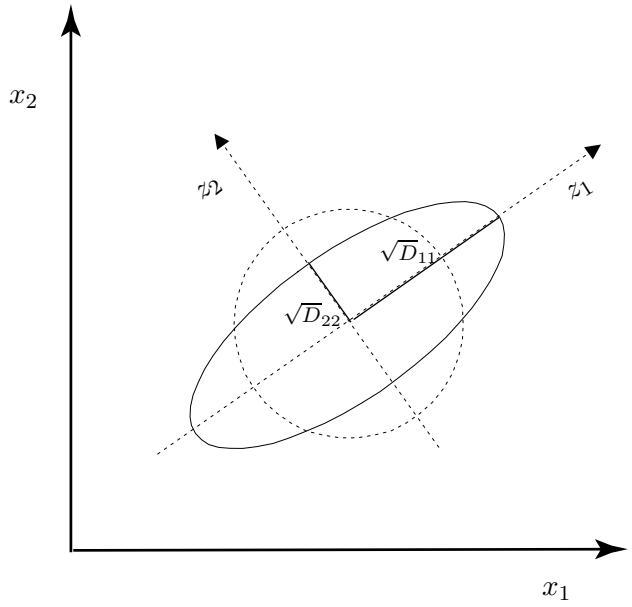

Figure 3.2: Sphering.

yields

$$\mathbf{Q}\mathbf{A}\,\mathbf{D}_k\,\mathbf{A}^T\mathbf{Q}^T = (\mathbf{Q}\mathbf{A})\,\mathbf{D}_k\,(\mathbf{Q}\mathbf{A})^T.$$

Since eigenvalue decompositions for symmetric matrices are unique, we know that the product $\mathbf{QA}$ must be an orthogonal matrix. Thus the diagonalization problem has been reduced to the orthogonal case.

Joint diagonalization with orthogonal matrices can be performed e.g. with the extended Jacobi method (Cardoso and Souloumiac, 1996). The Jacobi method implicitely restricts the solution of the optimization problem to the group of orthogonal matrices by multiplying a sequence of elementary (plane) rotations (Jacobi, 1846).

In order to preserve the orthogonality of $\mathbf{V}$ in the iterative Algorithm 2, we use the properties of the matrix exponential. We initialize with an orthogonal matrix $\mathbf{V}^{(0)}$ and perform the matrix exponential update

$$\mathbf{V}^{(m+1)} \leftarrow \exp(\mathbf{W}^{(m+1)})\mathbf{V}^{(m)},$$

where $\mathbf{W}^{(m+1)}$ is constrained to be skew-symmetric, i.e. $\mathbf{W} = -\mathbf{W}^T$. This ensures that $\mathbf{V}^{(m+1)}$ remains always orthogonal.

To see this, consider the transpose of $\mathbf{V} = \exp(\mathbf{W})$:

$$\mathbf{V}^T = \exp(\mathbf{W})^T = \exp(\mathbf{W}^T) = \exp(-\mathbf{W}) = \mathbf{V}^{-1}.$$

This implies that $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, i.e. $\mathbf{V}$ is indeed orthogonal.

Non-Orthogonal case

The algorithm 2 can also be used in the non-orthogonal case. Here we want to employ the constraint that $\mathbf{V}$ has to be an invertible (non-singular) matrix. Mathematically this condition means $\det \mathbf{V} \neq 0$. Due to the properties of the matrix exponential this is always guaranteed, since $\det(e^{\mathbf{W}}) = e^{\operatorname{tr} \mathbf{W}}$ and the exponential function is always non-zero.

Additionally, we may enforce $\mathbf{V}$ to be volume-preserving, i.e. $\det(\mathbf{V}) = 1$. Based on the following fact, which is also a consequence of the relation $\det(e^{\mathbf{W}}) = e^{\operatorname{tr} \mathbf{W}}$ for the matrix exponential, this can be done by setting the trace of $\mathbf{W}$ to zero:

Theorem (volume preservation). *If* $\operatorname{tr} \mathbf{W} = 0$*, then* $\det(e^{\mathbf{W}}) = 1$.

This means that the mutltiplicative update $\mathbf{V}^{(m+1)} \leftarrow \exp(\mathbf{W}^{(m+1)})\mathbf{V}^{(m)}$, with $\operatorname{tr} \mathbf{W} = 0$, preserves the determinant. If we start with a matrix $\mathbf{V}_0$, where $\det \mathbf{V}_0 = 1$, this ensures $\det(\mathbf{V}^{(m)}) = 1, \forall m$.

However the exact matrix exponential is relativly expensive to compute ($\mathcal{O}(N^3)$) and thus one may want to use a computationally cheaper first-order approximation $e^{\mathbf{W}} \approx \mathbf{I} + \mathbf{W}$ (see Figure 3.3). In order to maintain invertibility of $\mathbf{V}$ when using a such a first-order approximation, it suffices to ensure invertibility of $\mathbf{I} + \mathbf{W}$. For this purpose we can resort to the following results of matrix analysis (Horn and Johnson, 1985).

Definition. *An* $N \times N$ *matrix* $\mathbf{M}$ *is said to be strictly diagonally dominant if*

$$|M_{ii}| > \sum_{j \neq i} |M_{ij}|, \quad \text{for all } i = 1, \ldots, N.$$

Theorem (Levi-Desplanques). *If an* $N \times N$ *matrix* $\mathbf{M}$ *is strictly diagonally-dominant, then it is invertible.*

With $\mathbf{M} = \mathbf{I} + \mathbf{W}$ the Levi-Desplanques theorem helps us to control the invertibility of $\mathbf{I} + \mathbf{W}$. We notice that the diagonal entries in $\mathbf{I} + \mathbf{W}$ are all equal to 1; therefore, it suffices to ensure that

$$1 > \max_i \sum_{j \neq i} |\mathbf{W}_{ij}| = ||\mathbf{W}||_\infty.$$

This can be done by scaling $\mathbf{W}$ by its infinity norm $||\mathbf{W}||_\infty$ whenever the latter exceeds some fixed threshold $\theta < 1$. An even stricter condition can be imposed by using a Frobenius norm $||\mathbf{W}||_F$ in the same way:

$$\mathbf{W} \leftarrow \frac{\theta}{||\mathbf{W}||_F} \mathbf{W}. \tag{3.12}$$

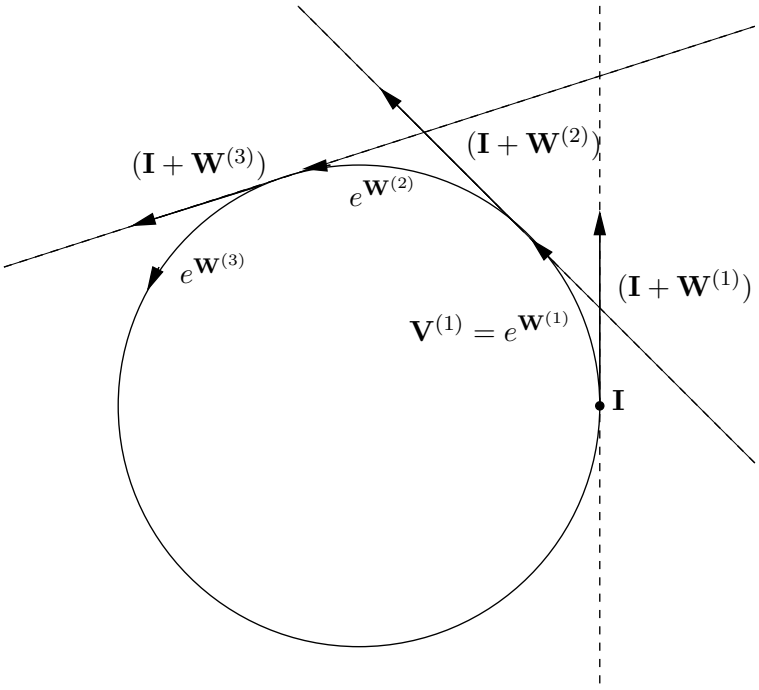

Figure 3.3: The (matrix) exponential can be approximated by $\mathbf{I}+\mathbf{W}$ for small $\mathbf{W}$.

3.2.5 Discussion

In principle, a restriction of $\mathbf{V}$ to the group of orthogonal matrices can be used if at least one matrix of the set of target matrices happens to be positive-definite. In this case, the pre-sphering step can be applied. We note however that such a two step method may degrade the overall performance, because one (arbitrary) matrix of the set is diagonalized exactly at the expense of a worse diagonalization of the remaining matrices. This can be especially problematic in the context of blind source separation (Cardoso, 1994; Yeredor, 2002; Akuzawa and Murata, 2001). Thus we want to relax the orthogonality assumption and consider the case of approximate joint diagonalization with *non-orthogonal* matrices. In the non-orthogonal case however, we additionally need enforce some constraint to prevent trivial solutions. Here we make use of the invertibility of $\mathbf{V}$ which is implicitly guaranteed when using the multiplicative matrix exponential update. Invertibility is a inherent necessity in many applications of diagonalization algorithms, especially in blind source separation, therefore making use of such a constraint is very natural and does not limit the usefulness from the practical point of view.

The cost of computing the matrix exponential are $\mathcal{O}(N^3)$, which is relatively high, but since for typical BSS problems the dimensionality of the

matrices is of the order $N \approx 100$ it is by no means computational prohibitive. A computationally cheaper alternative to the exact matrix exponential update is to use a first-order approximation $\exp(\mathbf{W}) \approx \mathbf{I} + \mathbf{W}$ for sufficiently small $\mathbf{W}$.

3.3 Computation of the Update Matrix

In this section we are going to derive the update rules for the matrix $\mathbf{W}^{(m+1)}$ such that we actually minimize our joint diagonality criterion. We will consider two approaches: a gradient method, called DOMUNG (Yeredor, Ziehe and Müller, 2004) and a Newton-like method, called FFDiag (Ziehe et al., 2003c, 2004).

3.3.1 Relative Gradient Algorithm: DOMUNG

Throughout the following derivations the operation of setting the diagonal of a matrix to zero is frequently used. Thus we denote this operation by putting an upper bar above the respective expression. More specifically, for any square matrix $\mathbf{M}$ we define the notation $\overline{\mathbf{M}}$ as

$$\overline{\mathbf{M}} \stackrel{\text{def}}{=} \mathbf{M} - \text{diag}(\mathbf{M}) \tag{3.13}$$

Note that the off$(\cdot)$ operator defined in (3.5) can then be expressed based on the trace of a matrix:

$$\text{off}(\mathbf{M}) = ||\overline{\mathbf{M}}||_F^2 = \text{tr}\{\overline{\mathbf{M}}^T\overline{\mathbf{M}}\} = \text{tr}\{\mathbf{M}^T\overline{\mathbf{M}}\}. \tag{3.14}$$

To determine the updates $\mathbf{W}$ at each iteration, first-order optimality constraints for the objective (3.7) are used. We may therefore define, for each iteration m,

$$\tilde{J}_2^{(m)}(\mathbf{W}) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \text{off}((\mathbf{I}+\mathbf{W})\mathbf{C}_k^{(m)}(\mathbf{I}+\mathbf{W})^T), \tag{3.15}$$

as the cost function which we seek to minimize w.r.t. $\mathbf{W}$. To this end, we now seek the gradient $\partial\tilde{J}_2^{(m)}(\mathbf{W})/\partial\mathbf{W}$, which is a matrix whose (i,j)-th element is the derivative of $\tilde{J}_2^{(m)}(\mathbf{W})$ w.r.t. W_{ij} (W_{ij} denoting the (i,j)-th element of $\mathbf{W}$). To find this gradient matrix, we first compute the gradient of each summand in (3.15). We do so by expressing the off$(\cdot)$ function in (3.15) in the vicinity of $\mathbf{W} = \mathbf{0}$ up to first-order terms in $\mathbf{W}$, i.e. we assume that $\mathbf{W}$ is a sufficiently small matrix (for shorthand we omit the indices in

the following expressions, i.e. we use $\mathbf{C}$ instead of $\mathbf{C}_k^{(m)}$):

$$\begin{aligned}
\mathrm{off}((\mathbf{I}+\mathbf{W})\mathbf{C}(\mathbf{I}+\mathbf{W})^T) &= \mathrm{tr}\{[(\mathbf{I}+\mathbf{W})\mathbf{C}(\mathbf{I}+\mathbf{W})^T]^T\overline{(\mathbf{I}+\mathbf{W})\mathbf{C}(\mathbf{I}+\mathbf{W})^T} \\
&= \mathrm{tr}\{(\mathbf{I}+\mathbf{W})\mathbf{C}(\mathbf{I}+\mathbf{W})^T\overline{(\mathbf{I}+\mathbf{W})\mathbf{C}(\mathbf{I}+\mathbf{W})^T}\} \\
&\approx \mathrm{tr}\{(\mathbf{C}+\mathbf{W}\mathbf{C}+\mathbf{C}\mathbf{W}^T)\overline{(\mathbf{C}+\mathbf{W}\mathbf{C}+\mathbf{C}\mathbf{W}^T)}\} \\
&\approx \mathrm{tr}\{\mathbf{C}\overline{\mathbf{C}}+\mathbf{C}\overline{\mathbf{W}\mathbf{C}}+\mathbf{C}\overline{\mathbf{C}\mathbf{W}^T}+\mathbf{W}\mathbf{C}\overline{\mathbf{C}}+\mathbf{C}\mathbf{W}^T\overline{\mathbf{C}}\} \\
\tilde{\tilde{J}}_2(\mathbf{W}) &= \mathrm{tr}\{\mathbf{C}\overline{\mathbf{C}}+\mathbf{C}\overline{\mathbf{C}}\mathbf{W}+\mathbf{C}\overline{\mathbf{C}}\mathbf{W}+\mathbf{C}\overline{\mathbf{C}}\mathbf{W}+\mathbf{C}\overline{\mathbf{C}}\mathbf{W}\} \\
&= \mathrm{tr}\{\mathbf{C}\overline{\mathbf{C}}\}+4\,\mathrm{tr}\{\mathbf{C}\overline{\mathbf{C}}\mathbf{W}\}. \quad (3.16)
\end{aligned}$$

We used (3.14) in the first line, and the identities $\mathrm{tr}\{\mathbf{M}\} = \mathrm{tr}\{\mathbf{M}^T\}$, $\mathrm{tr}\{\mathbf{M}\mathbf{Q}\} = \mathrm{tr}\{\mathbf{Q}\mathbf{M}\}$ and $\mathrm{tr}\{\mathbf{M}\overline{\mathbf{Q}}\} = \mathrm{tr}\{\overline{\mathbf{M}}\mathbf{Q}\}$ in the transition from the fourth line to the fifth. The $\approx$ symbol on the third and fourth lines indicates the elimination of terms of second or higher order in $\mathbf{W}$ in the respective transitions.

Noting that $\partial\,\mathrm{tr}\{\mathbf{M}\mathbf{W}\}/\partial\mathbf{W} = \mathbf{M}^T$, we obtain that the gradient of the off$(\cdot)$ function w.r.t. $\mathbf{W}$ is $4(\overline{\mathbf{C}}\mathbf{C})$.

Reactivating the full notation we obtain the gradient of $\tilde{\tilde{J}}_2^{(m)}$ w.r.t. $\mathbf{W}$ at the m-th iteration:

$$\frac{\partial\tilde{\tilde{J}}_2^{(m)}(\mathbf{W})}{\partial\mathbf{W}} = 4\sum_{k=1}^{K}\overline{\mathbf{C}_k^{(m)}}\mathbf{C}_k^{(m)}. \quad (3.17)$$

Since the goal is to decrease the value of $\tilde{\tilde{J}}_2^{(m)}$ in each iteration, we take a "steepest descent" step, by setting

$$\mathbf{W}^{(m+1)} = -\mu\,\frac{\partial\tilde{\tilde{J}}_2^{(m)}(\mathbf{W})}{\partial\mathbf{W}}, \quad (3.18)$$

where μ is some positive constant.

The stepsize is either set heuristically to some small fixed value (e.g. $\mu = 0.01$) or adaptively controlled using a strategy as in Murata et al. (2002). In Yeredor et al. (2004), it has been shown that even the optimal value for μ can be found by calculating the roots of a polynomial of degree 3.

3.3.2 Relative Newton-like Algorithm: FFDiag

A further approximation of the objective function can be used to compute $\mathbf{W}^{(m+1)}$ even more efficiently. To this end we now split the target matrices $\mathbf{C}_k^{(m)}$ in two parts:

the diagonal $\mathbf{D}_k^{(m)} \stackrel{\mathrm{def}}{=} \mathrm{diag}(\mathbf{C}_k^{(m)})$ and off-diagonal $\mathbf{E}_k^{(m)} \stackrel{\mathrm{def}}{=} \overline{\mathbf{C}_k^{(m)}}$ part.

We note that in the vicinity of the solution the norm of $\mathbf{E}_k^{(m)}$ is small. In order to simplify the cost function further we exploit this fact by ignoring those terms which are a product of two small factors.

$$\begin{aligned}\tilde{\tilde{J}}_2^{(m)}(\mathbf{W}) &= \sum_{k=1}^{K} \mathrm{off}((\mathbf{I}+\mathbf{W})(\mathbf{D}_k^{(m)}+\mathbf{E}_k^{(m)})(\mathbf{I}+\mathbf{W})^T) \\ &\approx \sum_{k=1}^{K} \mathrm{off}(\mathbf{D}_k^{(m)} + \mathbf{W}\mathbf{D}_k^{(m)} + \mathbf{D}_k^{(m)}\mathbf{W}^T + \mathbf{E}_k^{(m)}). \end{aligned} \tag{3.19}$$

The first term in (3.19) is already diagonal and thus irrelevant for the minimization. Dropping those irrelevant terms, we obtain the threefold approximated cost function:

$$\tilde{\tilde{J}}_2^{(m)}(\mathbf{W}) = \sum_{k=1}^{K} \mathrm{off}(\mathbf{W}\mathbf{D}_k^{(m)} + \mathbf{D}_k^{(m)}\mathbf{W}^T + \mathbf{E}_k^{(m)}). \tag{3.20}$$

The linearity of (3.20) in terms of $\mathbf{W}$ simplifies the problem enormously and will allow us to explicitly compute the optimal update matrix $\mathbf{W}^{(m+1)}$ by minimizing the criterion $\tilde{\tilde{J}}_2(\mathbf{W})$ with a Newton step.

In contrast to solving a full Newton system (which would involve the inversion of a large $N(N-1)\times N(N-1)$ matrix, the key to the computational efficiency of the FFDiag algorithm lies in exploiting the sparseness introduced by the approximation (3.20). Due to this sparse structure the inverse of the (approximated) Hessian matrix can be computed in closed form. In order to see this favorable special structure of the problem, we restate the problem in a matrix-vector notation presented next.

Let the $N(N-1)$ off-diagonal entries of the matrix $\mathbf{W}$ be stacked in a large vector $\mathbf{w}$ as

$$\mathbf{w} = [W_{12}, W_{21}, \ldots, W_{ij}, W_{ji}, \ldots]^T. \tag{3.21}$$

Notice that this is *not* the usual vectorization operation vec $\mathbf{W}$, as the order of elements in $\mathbf{w}$ reflects the pairwise relationship of the elements in $\mathbf{W}$. In a similar way the $KN(N-1)$ off-diagonal entries of the matrices $\mathbf{E}_k$ are arranged as

$$\mathbf{e} = [(E_1)_{12}, (E_1)_{21}, \ldots, (E_1)_{ij}, (E_1)_{ji}, \ldots, (E_k)_{ij}, (E_k)_{ji}, \ldots]^T. \tag{3.22}$$

Finally, a large but very sparse, $KN(N-1)\times N(N-1)$ matrix $\mathbf{J}$ is built, in the form:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_K \end{pmatrix} \text{ with } \mathbf{J}_k = \begin{pmatrix} (B_k)_{12} & & & \\ & \ddots & & \\ & & (B_k)_{ij} & \\ & & & \ddots \end{pmatrix},$$

where each $\mathbf{J}_k$ is block-diagonal, containing $N(N-1)/2$ matrices of dimension 2×2

$$(B_k)_{ij} = \begin{pmatrix} (D_k)_{jj} & (D_k)_{ii} \\ (D_k)_{jj} & (D_k)_{ii} \end{pmatrix}, \quad i,j = 1,\ldots,N,\ i \neq j,$$

where $(D_k)_{ii}$ is the (i,i)-th entry of a diagonal matrix $\mathbf{D}_k$.

Using this notation the threefold approximated cost function can be rewritten in the familiar form of a linear least-squares problem (Press et al., 1992).

$$\tilde{\tilde{\tilde{J}}}_2(\mathbf{W}) = (\mathbf{J}\mathbf{w} + \mathbf{e})^T(\mathbf{J}\mathbf{w} + \mathbf{e}). \tag{3.23}$$

The vector $\mathbf{w}$ that minimizes (3.23) can be obtained in closed form as:

$$\mathbf{w} = -(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T\mathbf{e}. \tag{3.24}$$

We can now make use of the sparseness of $\mathbf{J}$ to enable the direct computation of the elements of $\mathbf{w}$ in (3.24). Writing out the matrix product $\mathbf{J}^T\mathbf{J}$ yields a block-diagonal matrix

$$\mathbf{J}^T\mathbf{J} = \begin{pmatrix} \sum_k ((B_k)_{12})^T (B_k)_{12} & & & \\ & \ddots & & \\ & & \sum_k (B_k)_{ij})^T (B_k)_{ij} & \\ & & & \ddots \end{pmatrix}$$

whose blocks are 2×2 matrices. Thus the system (3.24) actually consists of decoupled equations

$$\begin{pmatrix} W_{ij} \\ W_{ji} \end{pmatrix} = -\begin{pmatrix} z_{jj} & z_{ij} \\ z_{ij} & z_{ii} \end{pmatrix}^{-1} \begin{pmatrix} y_{ij} \\ y_{ji} \end{pmatrix}, \quad i,j = 1,\ldots,N,\ i \neq j, \tag{3.25}$$

where

$$z_{ij} = \sum_k (D_k)_{ii}(D_k)_{jj}$$

$$y_{ij} = \sum_k (D_k)_{jj} \frac{(E_k)_{ij} + (E_k)_{ji}}{2} = \sum_k (D_k)_{jj}(E_k)_{ij}.$$

The matrix inverse in equation (3.25) can be computed in closed form, leading us to the following expressions for the update of the entries of $\mathbf{W}$:

$$\begin{aligned} W_{ij} &= \frac{z_{ij}y_{ji} - z_{ii}y_{ij}}{z_{jj}z_{ii} - z_{ij}^2}, \\ W_{ji} &= \frac{z_{ij}y_{ij} - z_{jj}y_{ji}}{z_{jj}z_{ii} - z_{ij}^2}. \end{aligned} \tag{3.26}$$

Here, only the off-diagonal elements ($i \neq j$) need to be computed and the diagonal terms of $\mathbf{W}$ are set to zero.

In the orthogonal case, due to the skew-symmetry of $\mathbf{W}$, only one of each pair of the entries (3.26) needs to be computed since the other entry is set to $W_{ji} = -W_{ij}$. This yields the simpler expression for the elements of $\mathbf{W}$:

$$W_{ij} = \frac{\sum_k (E_k)_{ij} \left((D_k)_{ii} - (D_k)_{jj}\right)}{\sum_k ((D_k)_{ii} - (D_k)_{jj})^2}, \quad i, j = 1, \ldots, N,\ i \neq j, \tag{3.27}$$

and reduces the computational cost by a factor of two.

Discussion

The simplifying assumptions used in (3.19) require some further discussion. Our motivation for assuming that $\mathbf{W}$ and $\mathbf{E}_k$ are small is based on the observation that in the neighborhood of the solution, the matrices $\mathbf{C}_k$ are almost diagonal and thus the steps $\mathbf{W}$ towards the optimum are small.

Hence, in the neighborhood of the optimal solution the algorithm is expected to behave similarly to Newton's method and can converge quadratically.

We note however, that the assumption of small $\mathbf{E}_k$ is potentially problematic, especially in the case where exact diagonalization is impossible. In this case it is preferable to carried out the gradient descent step in equation (3.18), where $\mathbf{E}_k$ is fully taken into account.

In any case the assumption of $\mathbf{W}$ being small is crucial for the convergence of the algorithm and needs to be carefully controlled. The latter is done by the normalization (3.12).

Some general remarks on convergence properties of the proposed algorithmic scheme are due at this point. Newton-like algorithms are known to converge only in the neighborhood of the optimal solution; however, when they converge, the rate of convergence is quadratic (e.g. Kantorovich, 1949). Since the essential components of our algorithm—the second-order approximation of the objective function and the computation of optimal steps by solving the linear system arising from first-order optimality conditions—are inherited from Newton's method, the same convergence behavior can be expected.

The Newton direction in FFDiag is computed based on the Hessian of the threefold approximated cost function.[2] Taking advantage of the resulting special structure, this computation can be carried out very efficiently:

Instead of performing inversion and multiplication of large matrices, which would have brought us to the same $\mathcal{O}(KN^6)$ complexity as in the algorithm of van der Veen (2001), computation of the optimal $\mathbf{W}^{(m+1)}$ leads to a simple formula (3.26) which has to be evaluated for all $N(N-1)$ entries of $\mathbf{W}$. Since the computation of z_{ij} and y_{ij} also involves a loop over K, the overall complexity of the update step is $\mathcal{O}(KN^2)$.

[2] Furthermore, the Hessian is approximated by the product of the Jacobian matrices.

3.4 Summary

We proposed new algorithms for simultaneous diagonalization of a set of symmetric matrices using multiplicative updates based on the matrix exponential as a structural constraint to prevent trivial solutions.

The close relations between the blind source separation problem and the approximate joint diagonalization problem led us to specially adapted parameterizations and approximations of a nonlinear least-squares cost function.

The efficiency of the derived algorithm comes from the special second-order approximation of the cost function, which yields a block-diagonal Hessian and thus allows for highly efficient computation of a (quasi-) Newton update step. The main result is a closed form solution for the update of a pair of matrix elements.

The multiplicative update has the advantage of preserving the group structure of the problem.

For simplicity of the derivations we considered the case that the target matrices $\mathbf{C}_k$ are all real-valued and symmetric. An extension to the more general, complex-valued case is possible, but has to be left for future work.

Chapter 4

Numerical Simulations

In the following, a series of numerical experiments aimed at comparing the performance of the algorithms on synthetic approximate joint diagonalization tasks is provided.

4.1 Introduction

The experiments in this chapter are intended to demonstrate the performance of the newly developed AJD algorithms in general and in particular to compare the FFDiag algorithm with other state-of-the-art algorithms for approximate joint diagonalization of synthetic benchmark data and of typical target matrices occurring in BSS applications.

To facilitate a comparison, we first introduce suitable measures of performance. Then we present the results of five progressively more complex experiments. As a starting point we perform a "sanity check" experiment, i.e. to diagonalize a set of perfectly diagonalizable matrices which is a relatively easy task. This experiment is intended to emphasize that for small-size diagonalizable matrices the algorithm's performance matches the expected quadratic convergence. In the second experiment we compare the FFDiag algorithm with the extended Jacobi method as used in the JADE algorithm of Cardoso and Souloumiac (1993) (orthogonal Frobenius norm formulation), Pham's algorithm for positive-definite matrices (Pham, 2001) and Yeredor's AC-DC algorithm (Yeredor, 2002) (non-orthogonal, subspace fitting formulation). In the third experiment we investigate the scaling behavior of our algorithm as compared to AC-DC. Furthermore, the performance of the FFDiag algorithm is tested and compared with the AC-DC algorithm on noisy, non-diagonalizable matrices. Finally, the application of our algorithm to BSS is illustrated.

4.2 Performance Measures

Evaluating the algorithms for approximate joint diagonalization requires definite performance measures. The most straightforward measure of performance is to monitor the evolution of the objective function.

In synthetic experiments with artificial data the distance from the true solution is a good evaluation criterion. To be meaningful, this distance has to be invariant w.r.t. the irrelevant scaling and permutation ambiguities. For this reason, we choose a performance index that is commonly used in the context of ICA/BSS where the same invariances exist (see e.g. in Amari and Cichocki, 1998; Cardoso, 1999). Following the formulation of Moreau (2001) a suitable performance index is defined on the normalized "global" matrix $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{VA}$ according to

$$\text{score}(\mathbf{M}) = \frac{1}{2}\left[\sum_i \left(\sum_j \frac{|M_{ij}|^2}{\max_l |M_{il}|^2} - 1\right) + \sum_j \left(\sum_i \frac{|M_{ij}|^2}{\max_l |M_{lj}|^2} - 1\right)\right] \quad (4.1)$$

The first sum is small when each column of $\mathbf{M}$ has exactly one dominating element. The second sum gets small when each row of $\mathbf{M}$ has exactly one dominating element. The non-negative index (4.1) becomes zero iff $\mathbf{M}$ is a product of an invertible diagonal matrix $\mathbf{D}$ and of a permutation matrix $\mathbf{P}$, i.e., $\mathbf{M} = \mathbf{DP}$. Thus, if the algorithm was successful, score($\mathbf{M} = \mathbf{VA}$) should be close to zero.

4.3 FFDiag in Practice

4.3.1 "Sanity check" Experiment

The test data in this experiment is generated as follows. We use $K = 15$ diagonal matrices $\mathbf{D}_k$ of size 5×5 where the elements on the diagonal are drawn from a uniform distribution in the range $[-1, \ldots, 1]$ (cf. Joho and Rahbar, 2002). These matrices are 'mixed' by an orthogonal matrix A according to $\mathbf{A}\,\mathbf{D}_k\mathbf{A}^T$ to generate the set of target matrices $\{\mathbf{C}_k\}$ to be diagonalized.[1] The FFDiag algorithm is initialized with the identity matrix $\mathbf{V}^{(0)} = \mathbf{I}$, and the skew-symmetric update rule (3.27) is used.

The convergence behavior of the algorithm in 10 runs is shown in Figure 4.1. The diagonalization error is measured by the off$(\cdot)$ function. One can see that the algorithm has converged to the correct solution after less than 10 iterations in all trials. A quadratic convergence rate is observed from early iterations.

[1] The orthogonal matrix was obtained from a singular value decomposition of a random 5×5 matrix, where the entries are drawn from a standard normal distribution.

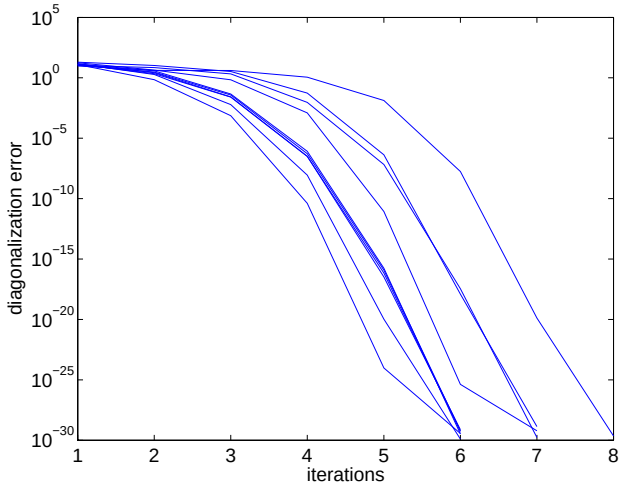

Figure 4.1: Evolution of the diagonalization error of the FFDIAG algorithm for a diagonalizable problem.

4.3.2 Diagonalizable vs Non-Diagonalizable Case

We now investigate the impact of non-diagonalizability of the set of matrices on the performance of the FFDIAG algorithm. Again, two scenarios are considered: the one of the "sanity check" experiment and the comparative analysis against the established algorithms. Non-diagonalizability is modeled by adding a random non-diagonal symmetric "noise" matrix to each of the input matrices:

$$\mathbf{C}_k = \mathbf{A}\,\mathbf{D}_k\mathbf{A}^T + \sigma^2(\mathbf{R}_k)(\mathbf{R}_k)^T,$$

where the elements of $\mathbf{R}_k$ are drawn from a standard normal distribution. The parameter σ allows one to control the impact of the non-diagonalizable component. Another example, with a more realistic noise model, will be presented in subsection 4.6.

Figure 4.2 shows the convergence plots of FFDIAG for various values of σ. The experimental setup is the same as in Section 4.3.1, apart from the additive noise. The impact of the latter can be quantified by computing the off$(\cdot)$ function on the noise terms only (averaged over all runs), which is shown by the dotted line in Figure 4.2. One can see that the algorithm converges quadratically to the level determined by the noise factor.

Similar to the second scenario in Section 4.4, the previously mentioned algorithms are tested on the problem of approximate joint diagonalization with non-orthogonal transforms. (Only the extended Jacobi algorithm had to be excluded from the comparison since it is not designed to work with non-orthogonal diagonalizers.) However, in contrast to Section 4.4, positive-definite target matrices were generated in order to facilitate a comparison with Pham's method.

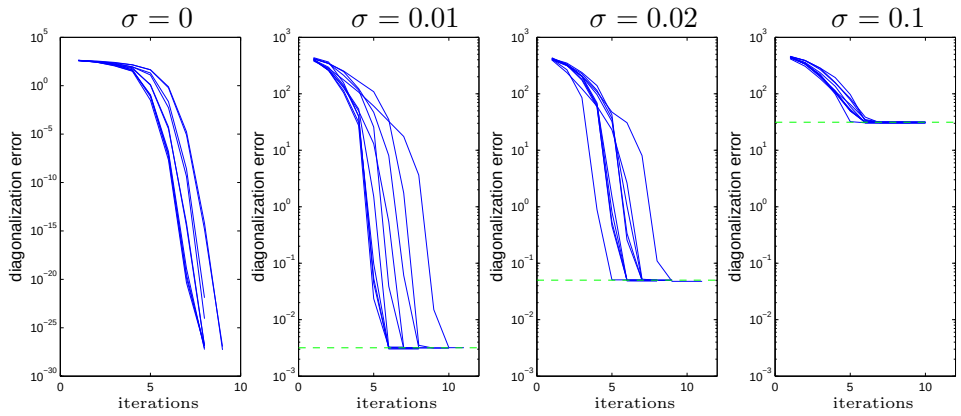

Figure 4.2: Diagonalization errors of the FFDIAG algorithm on non-diagonalizable matrices.

4.4 Comparison with other Algorithms

Two scenarios are considered for a comparison of the four selected algorithms: FFDIAG, the extended Jacobi method, Pham's algorithm and AC-DC. First, we test these algorithms on diagonalizable matrices under the conditions satisfying the assumptions of all of them. Such conditions are: positive-definiteness of the target matrices $\mathbf{C}_k$ and orthogonality of the true transformation $\mathbf{A}$ used to generate those matrices. These conditions are met by generating the target matrices $\mathbf{C}_k = \mathbf{A}\,\mathbf{D}_k\mathbf{A}^T$ where D_k are diagonal matrices with positive entries on the main diagonal. The data set consists of 100 random matrices of size 10×10 satisfying the conditions above.

A comparison of the four algorithms on orthogonal positive-definite matrices is shown in Figure 4.3. Two runs of the algorithms are presented, for the AC-DC algorithm 5 AC steps were interlaced with 1 DC step at each iteration. Although the algorithms optimize different objective functions, the off$(\cdot)$ function is still an adequate evaluation criterion provided that the arbitrary scale is properly normalized.

To achieve this, we evaluate $\sum_k \text{off}(\mathbf{V}\mathbf{C}_k\mathbf{V}^T)$ where $\mathbf{V}$ is the estimated diagonalizer. At the true solution the criterion must attain zero. One can see that the convergence of Pham's algorithm, the extended Jacobi method and FFDIAG is quadratic, whereas the AC-DC algorithm converges linearly. The average iteration complexity of the four algorithms is shown in Table 4.1. It follows from this table that the FFDIAG algorithm indeed lives up to its name: its running time per iteration is superior to both Pham's algorithm and AC-DC, and is comparable to the extended Jacobi method algorithm.[2]

In the second scenario, the comparison of the FFDIAG and the AC-DC

[2]In all experiments, MATLAB implementations of the algorithms were run on a standard PC with a 750MHz clock.

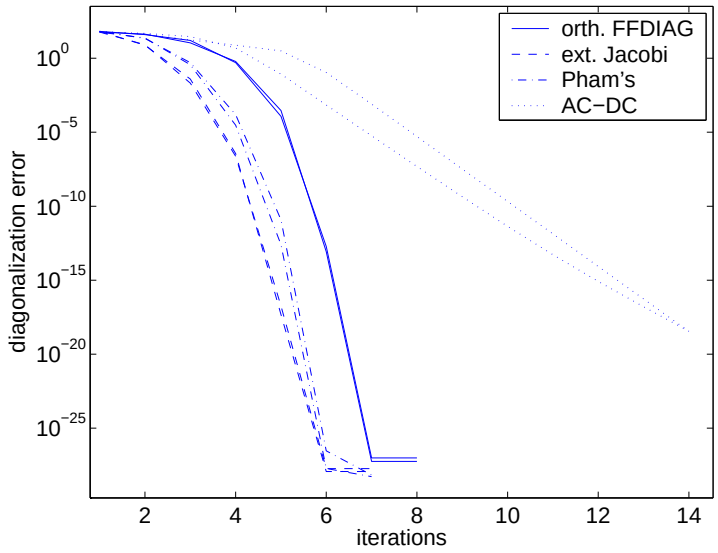

Figure 4.3: Comparison of the FFDiag, the extended Jacobi method, Pham's algorithm and AC-DC in the orthogonal, positive-definite case: diagonalization error per iteration measured by the off(·) criterion.

FFDiag	ext. Jacobi	Pham's	AC-DC
0.025	0.030	0.168	2.430

Table 4.1: Comparison of the FFDiag, ext. Jacobi, Pham's and AC-DC algorithms in the orthogonal, positive-definite case: average running time per iteration in seconds.

algorithms is repeated for non-positive-definite matrices obtained from a non-orthogonal mixing matrix. This case cannot be handled by the other two algorithms, therefore they are omitted from the comparison. The convergence plots are shown in Figure 4.4; average running time per iteration is reported in Table 4.2. Convergence behavior of the two algorithms is the same as in the orthogonal, positive-definite case; the running time per iteration of FFDiag increases due to the use of non-skew-symmetric updates.

FFDiag	AC-DC
0.034	2.64

Table 4.2: Comparison of the FFDiag and AC-DC algorithms in the non-orthogonal, non-positive-definite case: average running time per iteration in seconds.

The results of the comparison of the FFDiag, Pham's and AC-DC algorithms on a non-orthogonal positive-definite problem (5 matrices of dimen-

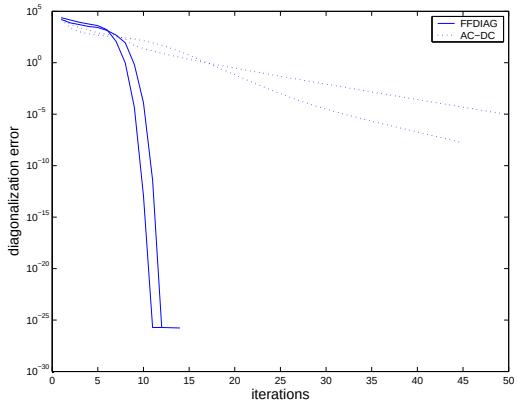

Figure 4.4: Comparison of the FFDIAG and AC-DC algorithms in the non-orthogonal, non-positive-definite case: diagonalization error per iteration measured by the off(·) criterion.

sion 5×5) at various noise levels are shown in Figure 4.5 for three typical runs. The graphs illustrate some interesting aspects of the convergence behavior of the algorithms. Both the FFDIAG and Pham's algorithm converge within a small number of iterations to approximately the same error level. The AC-DC algorithm converges linearly, and occasionally convergence can be very slow, as can be seen in each of the plots in Figure 4.5. However, when AC-DC converges, it exhibits better performance as measured by the score function; the higher the noise level, the stronger the difference.

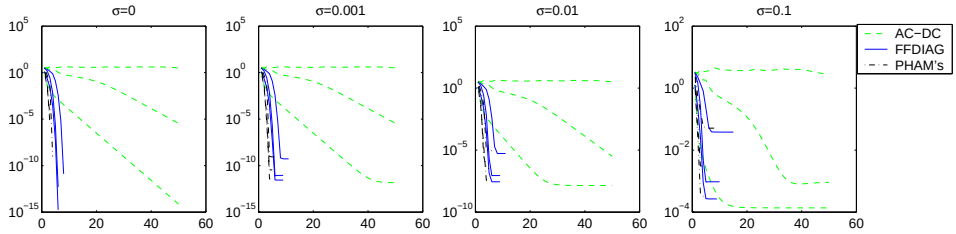

Figure 4.5: Comparison of the FFDIAG, Pham's and AC-DC algorithms in the non-diagonalizable, non-orthogonal, positive-definite case at various noise levels: performance index as measured by the score function (4.1).

4.4.1 Gradient vs Newton-like Updates

Here we compare the three algorithms DOMUNG (Yeredor et al., 2004) AC-DC (Yeredor, 2002) and FFDIAG (Ziehe et al., 2004).

The test data for this experiment is generated as follows. We use $K = 10$ diagonal matrices $\mathbf{D}_k$ of size 3×3 where the elements on the diagonal are drawn from a uniform distribution in the range $[-1 \dots 1]$. These matrices are 'mixed' using the fixed matrix $\mathbf{A} = \begin{bmatrix} 8 & 1 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 2 \end{bmatrix}$ according to the model

$\mathbf{A}\mathbf{D}_k\mathbf{A}^T$ to obtain the set of matrices $\mathbf{C}_k$ to be diagonalized.

The convergence behavior of the 3 algorithms in 10 runs is shown in Fig.4.6. The diagonalization error is measured by the off$(\cdot)$ function. The shaded area denotes the minima and maxima, while the bold line indicates the median over the 10 runs. In all cases the algorithms converged to the correct solution within the numerical computing precision. The differences in the final levels are only due to the use of slightly different stopping criteria. However, the convergence rates are clearly different. AC-DC and DOMUNG have a linear convergence rate and need quite many iterations. In contrast, FFDiag has quadratic convergence rate and needs less then 10 iterations.

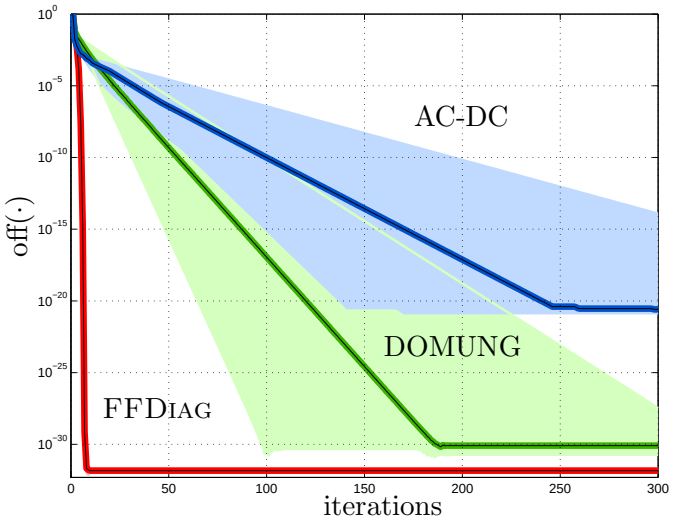

Figure 4.6: Diagonalization errors of the FFDiag, DOMUNG and AC-DC algorithm for a perfectly diagonalizable problem.

4.5 Computational Efficiency

Computational efficiency is essential for application of an algorithm to real-life problems. The most important parameter of the simultaneous diagonalization problem affecting the computational efficiency of an algorithm is the size of the matrices. Figure 4.7 shows the running time per iteration of the FFDiag and the AC-DC algorithms for problems with increasing matrix sizes, plotted at logarithmic scale. One can see that both algorithm exhibit running times of $\mathcal{O}(N^2)$; however, in absolute terms the FFDiag algorithm is almost two orders of magnitude faster.[3]

[3]This seemingly controversial result—theoretically expected scaling factor of AC-DC is $\mathcal{O}(N^3)$—is due to high constants hidden in the setup phase of AC-DC. The setup phase has $\mathcal{O}(N^2)$ complexity, but because of the constants it outweighs the main part of the algorithm in our experiment.

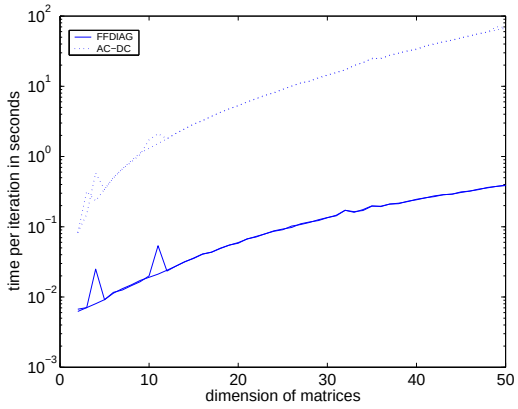

Figure 4.7: Scaling of the FFDiag and AC-DC algorithms with respect to the matrix size. Two repetitions of the experiment have been performed.

4.6 Blind Source Separation

4.6.1 Blind Separation of Audio Signals

We apply the joint diagonalization algorithms to a blind source separation task. Here the source signal matrix $\mathbf{S}$ contains seven audio signals containing 10000 points recorded at 8kHz and one Gaussian noise source of the same length (see Fig. 4.9). These signals are mixed by a 8×8 Hadamard matrix,

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}.$$

This scaled orthogonal matrix produces a complete mixture $\mathbf{X} = \mathbf{AS}$ (see Fig. 4.9, middle panel), in the sense that each observation contains a maximal contribution from each source.

In order to separate these signals, we compute 100 symmetrized, time-lagged correlation matrices according to Equation (2.12) and then apply the FFDiag algorithm with $\mathbf{V}^{(0)} = \mathbf{I}$. Figure 4.8 shows the evolution of the normalized diagonalization error. One can see that the algorithm converged after 6 iterations and that the normalized global system $\mathbf{V}^{(m)}\mathbf{A}$ converges to a permutation matrix (as shown in the left and the right panels of Fig. 4.8, respectively). The good performance can also be seen from Figure 4.9, where we observe a good match between the separated signals $\mathbf{U}$ and the true source signals $\mathbf{S}$.

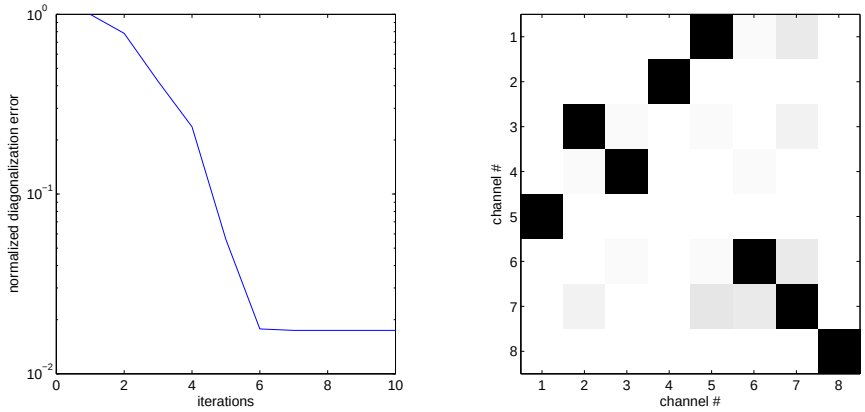

Figure 4.8: Convergence progress of the AJD4BSS algorithm on the BSS task. The right panel shows the entries of the matrix $\mathbf{V}^{(m)}\mathbf{A}$ for the final (10th) iteration and indicates successful separation, since the cross-talk is minimized and $\mathbf{V}^{(10)}\mathbf{A}$ resembles a scaled and permuted identity matrix. Here, black and white squares correspond to values 1 and 0, respectively.

4.6.2 Noisy mixtures

In order to study the behavior of the FFDiag algorithm in a more realistic noisy scenario the following experiment is conducted.

Input data is generated by mixing three stationary, time-correlated sources with the fixed matrix $A = \begin{pmatrix} 8 & 1 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 2 \end{pmatrix}$. The sources are generated by feeding an i.i.d. random noise signal into a randomly chosen, auto-regressive (AR) model of order 5 whose coefficients are drawn from a standard normal distribution and are sorted in decreasing order (to ensure stability). The generated signals have a total length of 50000 samples. To separate the sources we estimate 10 symmetrized, time-lagged correlation matrices of the mixed signals according to Equation (2.12) and perform simultaneous diagonalization of these matrices. The number T of samples used to estimate the correlation matrices determines the quality of the estimates. Thus, by varying T, we simulate different noise levels corresponding to different variances of the estimate. This procedure is more realistic than simply corrupting the target matrices with small additive i.i.d. noise.

The results of the experiment are shown in Figure 4.10. The performance of the FFDiag and the AC-DC algorithm, as measured by the score (4.1), is displayed for four different sample sizes, where small samples sizes correspond to a higher noise level. 100 repetitions are performed for each sample size, and the 25%, 50% and 75% quantiles of the log-score are shown in the plots. Two observations can be made from Figure 4.10: FFDiag converges much faster than AC-DC, and when converged, FFDiag yields a better score (on average), with the difference more pronounced for samples sizes 10000 and 30000 in our experiment.

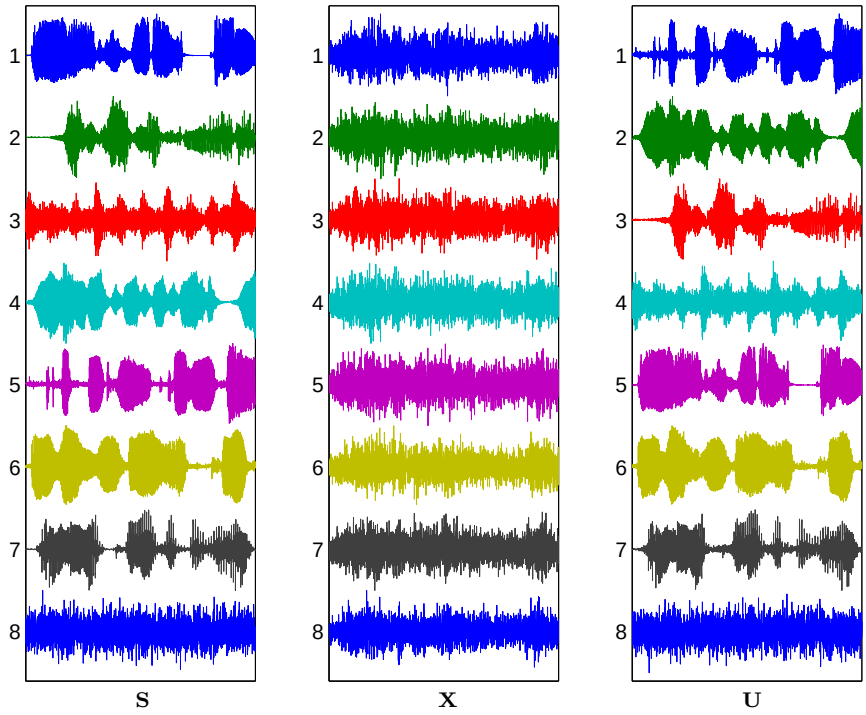

Figure 4.9: Waveforms of the original source signals **S**, mixed signals **X** and separated signals **U**.The separated signals **U** resemble the original signals **S** up to a random permutation of the order (e.g. S_1 matches U_5, S_2 matches U_3, ect.).

4.7 Summary

We have presented extensive experimental evidence indicating that the proposed algorithms work efficient and reliable. The gradient-based method DOMUNG has a linear convergence rate and for the quasi-Newton method FFDiag we observe even quadratic convergence in the neighborhood of the solution.

A series of comparisons of the FFDiag algorithm with state-of-the-art diagonalization algorithms showed that our algorithm is competitive with the best previous algorithms. In particular, FFDiag outperforms the AC-DC algorithm, which is the only competing algorithm applicable under the same general conditions. The FFDiag algorithm yields excellent results in cases where the model holds and it also performs reliably on non-diagonalizable data, for which only an approximate solution is possible. We demonstrated furthermore that the FFDiag algorithm succeeds in typical BSS scenarios.

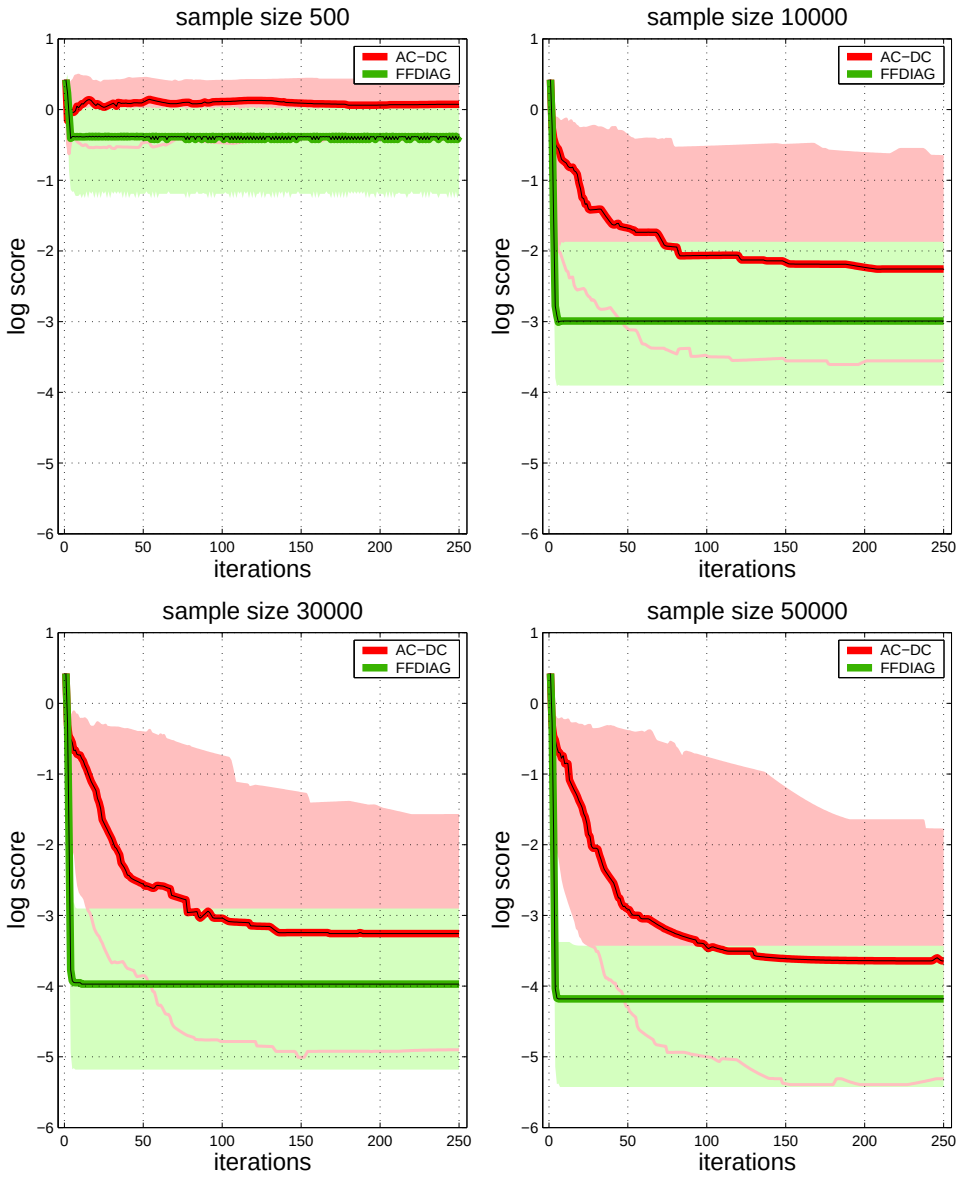

Figure 4.10: Performance of FFDiag and AC-DC measured by the log of the score (4.1) for different sample sizes and 100 trials each. 25% (lower edge of the shaded region), 50% (thick line in the middle) and 75% quantiles (upper edge of the shaded region) are shown.

Chapter 5

Applications

This chapter describes applications of blind source separation methods in biomedical signal processing. We demonstrate the usefulness of the proposed approach in the analysis of real-world neuro-physiological signals. The focus lies on using the joint diagonalization techniques in different scenarios where appropriate target matrices derived from the multi-channel measurements can be approximately diagonalized. The material presented in this chapter is partially based on the publications Ziehe et al. (2001) and Vigário et al. (2002); Wübbeler et al. (2000).

In the following we will present applications involving only the linear instantaneous BSS problem according to Equation (2.1). Even though this model is rather simple, it is particularly useful for biomedical signal processing as will be shown in the next section.

5.1 Biomedical signal processing

An example of BSS application for real-world biomedical signal processing is the analysis and pre-processing of electroencephalographic (EEG) and magnetoencephalographic (MEG) measurements. Here the routine usage of large sensor arrays (some of them consist of up to 300 SQUID-magnetometers) to record neuromagnetic fields in humans, produces data that suits the BSS approach quite well. Due to the fact that the electromagnetic waves superimpose linearly and virtually instantaneously (because of the relatively small distance from sources to sensors) the instantaneous linear model (5.1) is valid (Makeig et al., 1996; Vigário et al., 1998; Vigário et al., 2000; Wübbeler et al., 2000; Ziehe et al., 2000a).

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{5.1}$$

The columns of $\mathbf{A}$ represent the coupling of a source with each sensor. Since the array geometry is known, this information gives rise to a spatial field pattern. The the sources $\mathbf{s}(t)$ display the time courses of the components. It is precisely this decoupling of spatial and temporal information that makes these decompositions a valuable tool for exploratory data analysis.

The difficulties in the application those (blind) decompositions —as for any unsupervised data analysis—lie in the correctness and validation of the used modeling assumptions. Although our proposed methods are based on relatively "mild" assumptions they must still be verified for the particular application. In other words, it should always be checked whether or not the assumptions are fulfilled. In cases where this is not fully possible, one can still apply the method but one should be aware that a critical evaluation of the results —preferably based on medical knowledge— is needed.

Some of the limitations of the presented BSS-based approaches are the following:

These methods can extract at most as many underlying sources as the number of sensors. For example, in MEG and EEG analysis, a multitude of microscopic brain sources is present, which outnumbers by far the number of sensors used to record the magnetic or electric fields. Nevertheless, it can be argued that the total number of *macroscopically* observable sources, active at a given time-interval, is sufficiently smaller, and thus the use of ICA/BSS methods can be justified.

The use of separation criteria based on high-order statistics (e.g. in the JADE algorithm or kurtosis-based FastICA variants) comes at a price: an increased sensitivity to outliers. These often turn out to be the dominating factors in the decomposition due to their extreme non-Gaussianity.

Another issue that has to be kept in mind when applying higher-order statistics methods to large-scale sensor array data is the high computational load. For example in JADE the effort for storing and processing the 4-th order cumulants is $\mathcal{O}(N^4)$, where N is the number of sensors. Since this may be prohibitively large, the data has to be projected to a lower dimensional subspace as a preprocessing step.

For this reason we advocate methods which rely on second-order statistics only and additionally make use of the time-structure of the data. The diagonalization based procedures if employed on ensembles of suitable correlation matrices are expected to be superior in those respects.

5.1.1 Artifact Reduction by Adaptive Spatial Filtering

In the analysis of EEG and MEG data one often faces the problem that noise from biological or technical origins (like alpha rhythm activity or interference from power-lines, respectively) is corrupting the measurements.

In this section we study the capabilities of blind source separation to

construct an adaptive *spatial* filter that isolates the power-line signal while preserving evoked responses. We compare the spatial filtering by BSS to a classical notch filtering.

These frequency domain filtering techniques can also be applied to suppress power-line artifacts, this standard approach is clearly limited to narrow band signals. In contrast, BSS based spatial filtering methods are not relying on known narrow-band spectral characteristics but rather look for independent or temporally uncorrelated source signals (Vigário et al., 1998; Ziehe et al., 2000a).

Data

In this case study we analyze the effects of artifact removal in a typical experimental setting. The specific data set studied is obtained from measurements of somatosensory evoked fields (SEF, N20) and provides an attractive and rather controlled testbed since the signal of interest (N20) is relatively strong and the origin of the generator in the post-central gyrus is well-known.

Signals were measured using a low noise 63-channel DC-SQUID system (white noise level $2.7fT/\sqrt{Hz}$) operated in a first order axial electronic gradiometer mode with $70mm$ baseline; $7mm$ diameter SQUID pick up area, 49 sensors in a planar hexagonal configuration for the registration of the vertical field component, 30 mm distance between neighboring SQUID positions covering an area of 210 mm diameter (Drung, 1995).

The right median nerve was stimulated over 12.000 epochs, while the the magnetic field above the left somatic sensory cortex was measured using 49 planar gradiometer sensors. A sampling rate of 2 kHz and an inter-stimulus interval (ISI) of 333 msec were used to avoid steady state effects.

The recordings were carried out by the PTB[1] at the biomagnetism laboratory in the Department of Neurology at the campus Benjamin Franklin of the Charitè[2].

Although sophisticated magnetic shielding is used, a contamination of the measured biosignals by power-line interference can often not be completely avoided.

Artifact Reduction Procedure

The BSS based artifact reduction procedure consists of the following steps:

- apply a sphering and dimensionality reduction (based on PCA).
- decompose the transformed data into independent components by a BSS algorithm.

[1] The Physikalisch-Technische Bundesanstalt of Germany.
[2] The Charité - University Medicine Berlin.

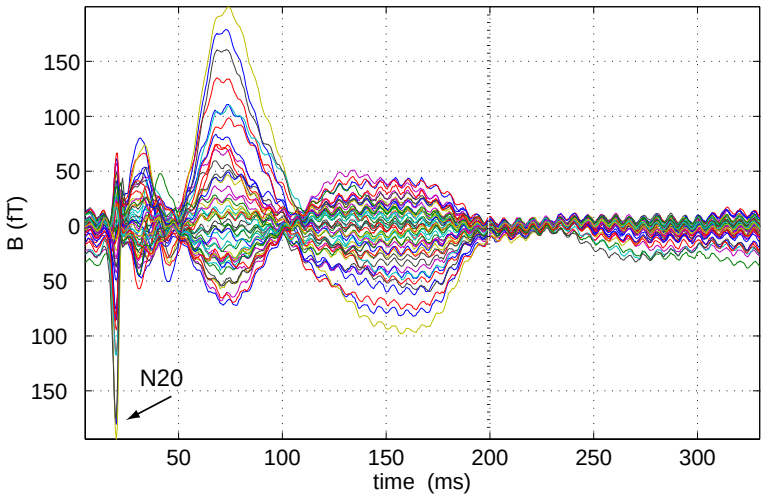

Figure 5.1: SEF averaged data of all 49 channels

- decide which components correspond to artifactual or relevant signals (e.g. by using prior medical knowledge).

- project the previously selected components of interest back to sensor space.

Applying this procedure we obtain a set of cleaned measurements.

Another possibility is to remove the artifact fields by making use of their estimated spatial structure (contained in the columns of the mixing matrix **A**) by Signal-Space Projection (SSP) (Uusitalo and Ilmoniemi, 1997). In case of multiple artifacts the whole space spanned by these artifacts, which we will refer to as "artifact space", has to be projected out. The corresponding projector is conveniently written in terms of an orthonormal basis of the artifact space. The essential requirement for applying SSP is that the unwanted fields are known up to unknown multiplicative constants which is exactly the case for the BSS model.

Performance Evaluation

The results of the BSS based procedure are compared to notch filtering (with a properly tuned notch frequency) as a "gold standard". A further indication of the success of the method can be gained by looking at the reduction of the peak at 150 Hz in the power spectrum Also the "goodness-of-fit" of an equivalent current dipole (ECD) model may provide a validation criteria, at least if based on a realistic volume conductor model.

Results

Predominantly two components with 150 Hz power were identified in the decomposition (Fig. 5.2) when using BSS.

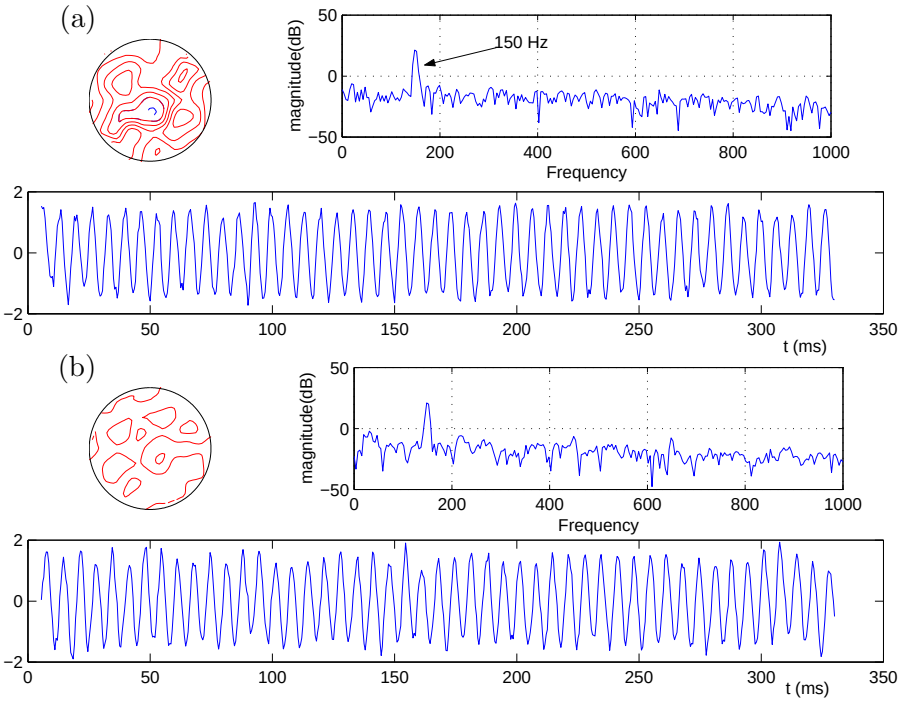

Figure 5.2: Identified power-line signals.

Eliminating these artifacts we find that BSS based techniques yield similar results as the "gold standard" notch filtering approach in a data-driven manner, provided that the sample size was larger than 400 samples (see Fig. 5.3).

The goodness-of-fit of the dipole model was even better when using BSS methods (cf. Fig. 5.4).

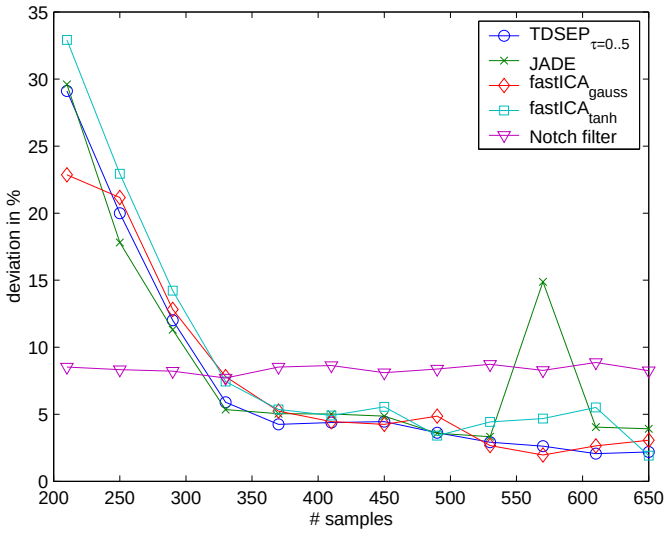

Figure 5.3: Deviation of the normalized field at N20 from the "gold standard" for varying sample size.

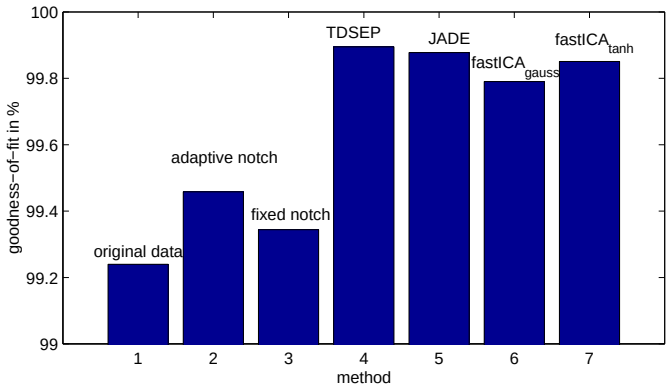

Figure 5.4: Goodness of fit at N20 for various methods.

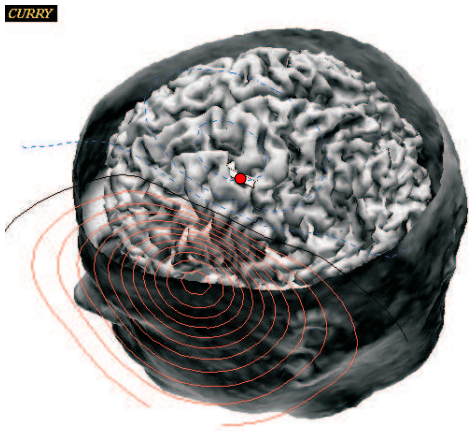

Figure 5.5: Source localization of N20 using the CURRY software with a realistic volume-conductor model from an MRI scan.

5.1.2 DC Magnetometry

In this section we report on the application of the joint diagonalization based BSS methods to data reflecting 'direct current' (DC-) activity in the human brain.

Since previously BSS techniques had been successfully applied to reduce artifacts in multi-channel EEG, MEG and MNG (magnetoneurography) recordings (Makeig et al., 1996; Vigário et al., 1998; Ziehe et al., 2000a) and to analyze evoked responses (Makeig et al., 1997) these methods were also expected to provide interesting decompositions of neuromagnetic data even in the near-DC range.

The identification of near-DC fields with non-invasive neuromagnetic recordings has great relevance for medical applications since slowly varying DC-phenomena have been found e.g. in cerebral anoxia and spreading depression in (invasive) animal studies. Blind source separation techniques have a high potential to become a standard clinical procedure for analysing such data.

Medical Background

Recently, the feasibility of a non-invasive *magnetic* registration of near-DC (below 0.1 Hz) magnetic fields from the human cortex using Superconducting Quantum Interference Devices (SQUIDs) has been shown (Mackert et al., 1999b). Such near-DC phenomena may have importance for metabolic injuries of brain cells in stroke or migraine (Back et al., 1994; Chen et al., 1992; Gardner-Medwin et al., 1991). Being able to perform a DC-coupled brain monitoring is of high medical relevance because many pathophysiological processes have their main energy in the frequency range below 0.1 Hz. Therefore, it is of great importance to further improve the signal extraction from DC-MEG data.

Technical background

The biomagnetic recording technology employed in this application is based on a unique apparatus that mechanically moves the subjects head, respectively, body relative to the sensor array (Wübbeler et al., 1999).

This transposes the near-DC signals of the head/body to the modulation frequency with lower 1/f noise. After signal demodulation this approach has a dynamical sensitivity to detect DC-fields > 30 fT in a 100 sec evaluation period (Wübbeler et al., 1999). This yields a high sensitivity which is both chance and challenge since it will not only enable physicians to detect minute physiological fields (Curio et al., 1993; Mackert et al., 1999a) but also poses problems for data analysis since the magnetic fields of a multitude of different biological processes and noise superimpose the signal of interest. It

is a helpful matter of fact that many of these processes vary in intensity *independently* of each other.

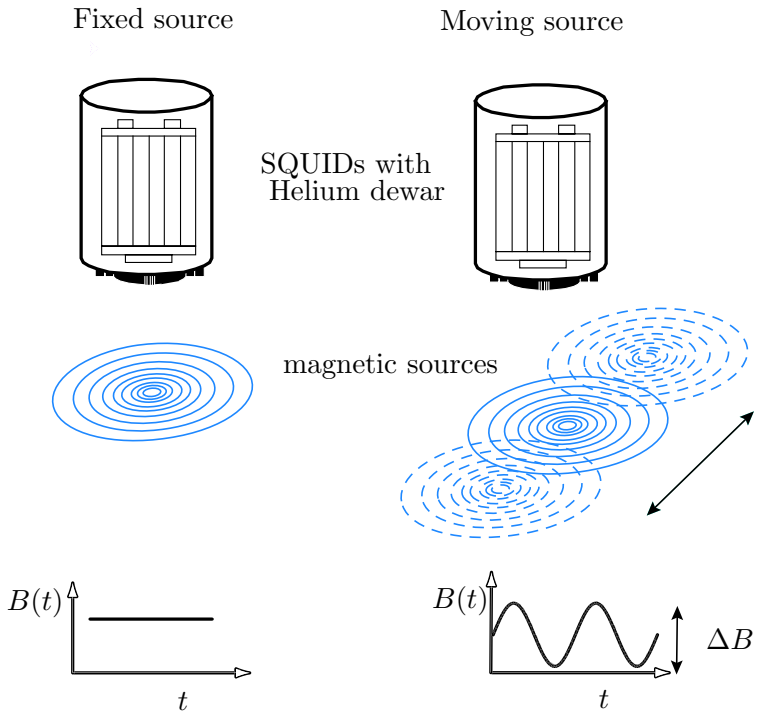

Figure 5.6: Modulation principle for DC measurements. Characteristic magnetometer output for a fixed and a moving source. The sinusodial movement of the source beneath the magnetometer array leads to an oscillating field signal.

Experimental Setup

In Mackert et al. (1999b) a paradigm of prolonged auditory (music) stimulation for DC-MEG was introduced. It consists of presenting 30 seconds of music and 30 seconds of silence to the subjects ear. This experimental setting has the advantages that a physiological DC-source in the brain with an essentially known field pattern[3] can be switch "on" and "off" arbitrarily by external non-invasive stimulation.

Data Acquisition and Validation

The neuromagnetic field data were recorded in a standard magnetically shielded room (AK3b), operated at the "Benjamin Franklin" hospital by the PTB, using 49 low noise first order SQUID gradiometers (70 mm baseline) covering a planar area of 210 mm diameter (Drung, 1995). The sensor

[3]The field patterns were expected to be comparable to patterns of evoked activities of auditory cortices as reported in Pantev et al. (1996).

was centered tangentially approximately over the left auditory cortex. The acoustic stimulation was achieved by presenting alternating periods of music and silence, each of 30 s length, to the subjects right ear during 30 min. of total recording time. The DC magnetic field values were acquired by using a mechanical horizontal modulation of the body position with a frequency of 0.4 Hz and an amplitude of 75 mm. This modulation transposed the DC magnetic field of the subject to the modulation frequency, which is less contaminated by magnetic noise (see also Fig. 5.6). The recorded magnetic field data were processed by digital lock-in techniques in order to extract the modulation induced frequency components (Wübbeler et al., 1998). Then the DC-field of the subject was reconstructed from these frequency components by using a transformation technique based on a virtual magnetic field generator (Mackert et al., 1999b). These reconstructed DC magnetic field values, sampled at the modulation frequency of 0.4 Hz, gave a total number of 720 sample points per channel for the 30 minutes recording time and were used as input for the BSS-algorithms.

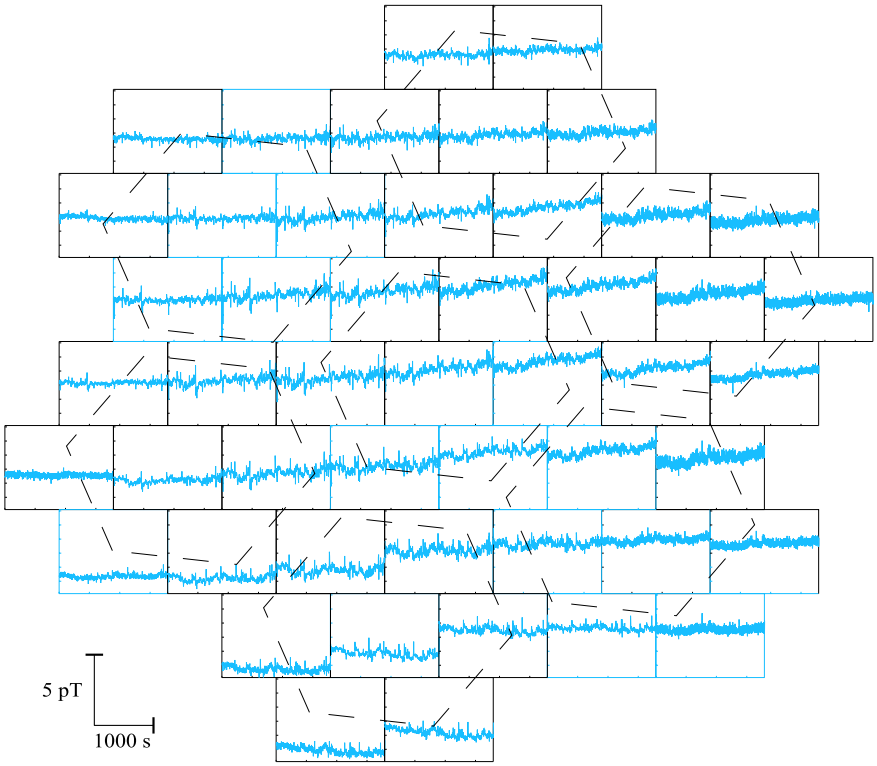

Figure 5.7: Input data (after DC demodulation) arranged according to sensor positions.

Let us examine the time courses of 30 minutes for all 49 channels (cf. Fig. 5.7). At the first glance, the signals have an obvious trend behavior (slow drift) while possible components of interest are covered by other strong signals of

unknown origin, i.e. the response to the stimulus is completely hidden in the data.

The above described experimental paradigm of externally controlled music-related DC-activations of auditory cortices defines a measurement and analysis scenario with almost complete knowledge about both, the spatial pattern and the time course of a cerebral DC-source which on the other hand is fully embedded in the biological and ambient noise background. Hence it may serve as a testbed for a critical comparison of different BSS approaches facing the 'real world' problems of bad signal-to-noise ratio coming along with a limited number of data samples and—on top of that—the presence of outliers.

Matrices to be Diagonalized

First, the originally 49-dimensional sensor data were reduced by PCA and only the 23 most powerful principal components were used.

When applying TDSEP (Ziehe and Müller, 1998) to the preprocessed data, we compute time-lagged correlation matrices of the form

$$\mathbf{C}_\tau(\mathbf{x}) = \langle \mathbf{x}(t)\mathbf{x}^T(t-\tau)\rangle = \begin{bmatrix} \phi_{x_1,x_1}(\tau) & \cdots & \phi_{x_1,x_n}(\tau) \\ \phi_{x_2,x_1}(\tau) & \cdots & \phi_{x_2,x_n}(\tau) \\ \vdots & \ddots & \vdots \\ \phi_{x_n,x_1}(\tau) & \cdots & \phi_{x_n,x_n}(\tau) \end{bmatrix}$$

where $\phi_{x_i,x_j}(\tau) = \langle x_i(t)x_j(t-\tau)\rangle$ denote the respective auto– or cross–correlation functions.

Here 50 time-lagged correlation matrices ($\tau = 1..50$ sample points) were used for approximate joint diagonalization.

Results

The 10 strongest ICA components are shown in Fig. 5.9. Not surprisingly, one component (ICA1) mainly captured the slow drift, that was already visible in the data in Fig. 5.7. While most other components show irregular time courses reflecting the dynamics of undetermined processes it is noteworthy that their field maps feature spatially coherent field patterns which clearly distinguish them from random channel noise patterns.

Remarkably, one component (ICA10) shows a (noisy) rectangular waveform. Its time course and frequency (see Fig. 5.8) clearly resembles the $\frac{1}{30s}$ "on/off" characteristics of the stimulus. The spatial field distribution of ICA10 shows a bipolar pattern, located at the expected position of cortical activity (Mackert et al., 1999b). Both findings give direct evidence that ICA10 represents the response to the acoustical stimulus. Although we do not expect that the cortical response resembles the stimulus completely, computing the correlation coefficient between the stimulus and the ICA time

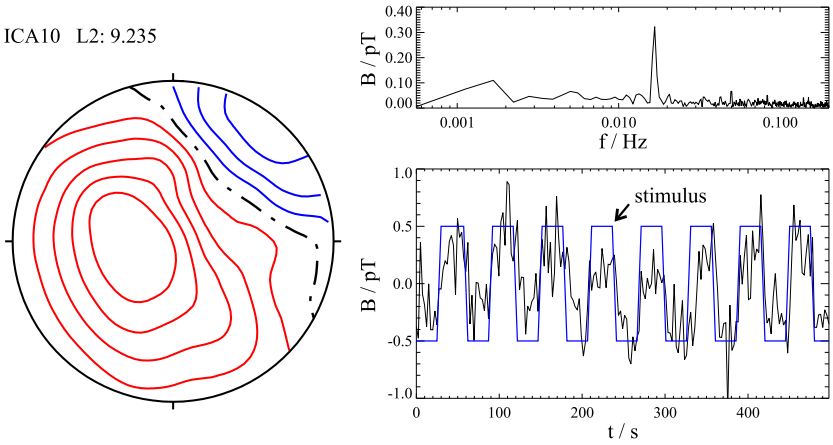

Figure 5.8: Spatial field pattern, frequency content and time course of ICA10.

courses provides a useful measure to evaluate and compare the performance of different separation algorithms. Applying the three algorithms JADE (Cardoso and Souloumiac, 1993), FastICA (Hyvärinen and Oja, 1997) and TDSEP (Ziehe and Müller, 1998), we find that only the temporal decorrelation algorithm TDSEP is able to recover a signal that is highly correlated to the stimulus, while FastICA and JADE fail for this specific task (for correlation coefficients see also Fig. 5.10).

Conclusion

Based on a comparison of several approaches, we found out that methods based on spatio-temporal decorrelation are able to successfully extract a component related to a sustained DC activation in the auditory cortex which was induced by presentation of music. The task is especially challenging due to the limited amount of available data and the occurrence of outliers. Since such a situation is very typical for biomedical measurements, this dataset provides also a useful real-world testbed for evaluating the performance and robustness of different BSS approaches.

It turned out that outliers can strongly decrease the performance of ICA algorithms, in particular methods that use higher-order statistics explicitly (e.g. JADE, FastICA with kurtosis) fail for this dataset, while in contrast spatio-temporal decorrelation methods based on joint-diagonalization of several time-delayed correlation matrices proved to be more robust.

From a general physiological point of view it is interesting to note that when employing these decomposition algorithms it became possible on the single subject level (i.e. without reverting to group statistics) to derive a faithful estimate for the time course of the DC-activation level in a particular area of the brain (i.e. the auditory cortex in the temporal lobe).

Most importantly, this analysis proceeded fully blind to our a priori experimental background knowledge on both the spatial signature of the music-related DC-fields (field map characteristic for auditory cortex activations) and its time course (30 sec. on and 30 sec. off). Both the spatial and the temporal source aspects were adequately captured in one ICA component (ICA10) using TDSEP. It is noteworthy that in contrast to earlier paradigms which identified cortical sources of short-term (2 - 9 sec) "sustained" fields (Pantev et al., 1996) or potentials (Picton et al., 1978) by averaging at least dozens of such repeated activations the present DC-MEG plus ICA approach allows to monitor the time course of cerebral DC-activations without any need for averaging (Fig. 5.8). In principle this is a first step towards "on-line" brain monitoring providing a chance for single trial, resp. single event analysis.

5.2 Summary

In the reported experiments, we have used real-world multi-channel biomagnetic recordings and demonstrated the merits and pitfalls of different BSS approaches. Special emphasis was given to the validation of the ICA/BSS model. The BSS based algorithms were able to isolate artifacts in MEG which can be used to clean the measurements and improve the results of biomagnetic source localization.

What makes blind source separation an appealing method for the analysis of neurobiological data, is that it uses a "weak" model, i.e. statistical independence or temporal decorrelation respectively. Although no strong model (such as the ones based on physiological models) is imposed on the data, the algorithms still extract components, which are neurophysiological plausible.

Recently we have suggested bootstrap based methods for a statistical validation of the reliability of ICA/BSS projections (Meinecke et al., 2001, 2002; Müller et al., 2004) and we expect that this approach will complement the set of tools for data analysis in future clinical applications.

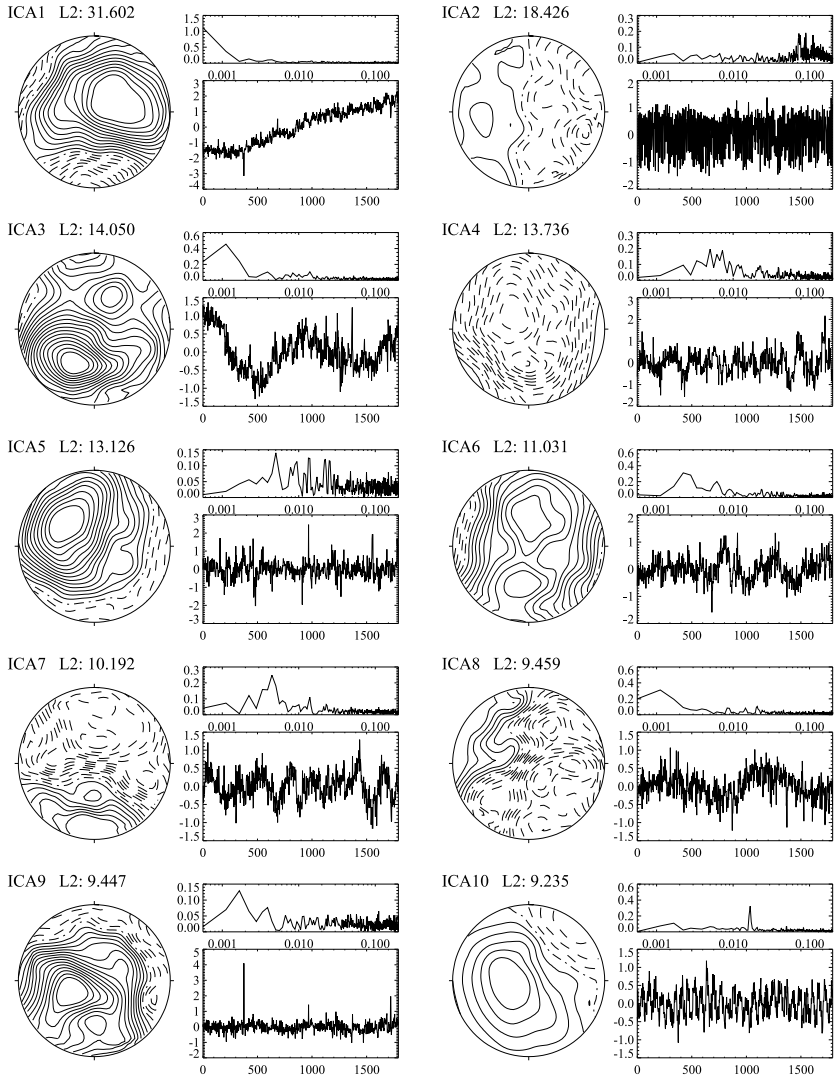

Figure 5.9: Spatial field patterns, waveforms and frequency contents of the first ten components obtained by TDSEP sorted according to the L2-norms. For units and details of ICA10 cf. Fig. 5.8.

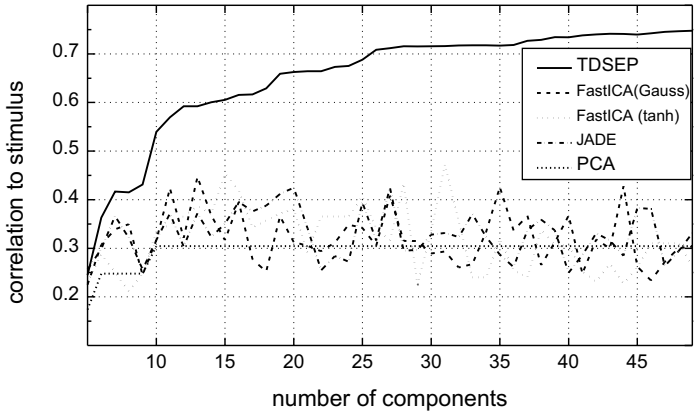

Figure 5.10: A PCA projection to a given number of components is performed prior to ICA in this subspace. We show the correlation coefficient between stimulus and the best matching ICA component vs number of components. The correlation to the best matching PCA component is shown as a baseline.

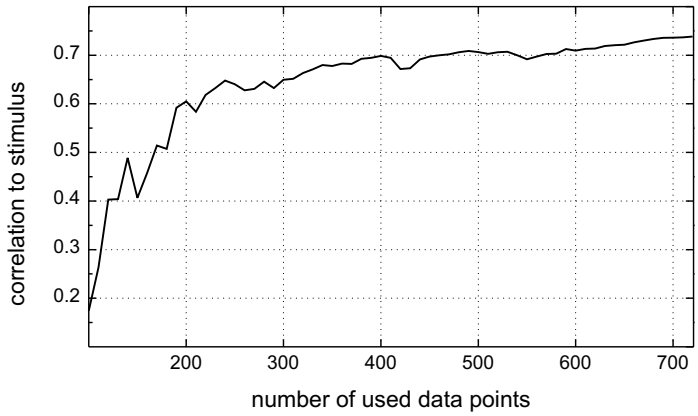

Figure 5.11: Correlation coefficient between stimulus and the best matching ICA component vs number of samples used for TDSEP applied on the full 49-dimensional sensor space.

Chapter 6

Conclusions

Here we summarize our main results and point out some possibilities for future research.

6.1 Summary

In this thesis we addressed the problem of blind source separation (BSS) using approximate joint diagonalization (AJD) of a set of matrices. We have shown that an AJD algorithm provides an efficient estimation procedure for the mixing matrix of the BSS problem. Thus a general AJD method is of great importance. As our main contribution, we proposed two new algorithms to solve the AJD problem numerically. The specific structure of these algorithms implements a numerical optimization procedure on a matrix group by utilizing matrix exponential updates. As in classical optimization methods, the updating can be based on a gradient descent step or on a Newton step. The later is know for a faster convergence rate (quadratic rather than linear), however the computational burden in the classical case is often prohibitively high.

The key features of our new algorithm called FFDiag are the use of a local parameterization of the cost function combined with a further approximation, which results in a block-diagonal Hessian with 2×2 blocks which allows for a closed form inverse and thus yields a highly efficient computation of the Newton update step.

Additional constraints, such as orthogonality, are not required by our algorithm, but if available, such constraints can be naturally incorporated and yield further simplifications of the algorithm.

We have empirically observed that the approximative solution found by the algorithm is of high quality for practical applications, and in particular for BSS problems.

A series of comparisons of the FFDiag algorithm with state-of-the-art diagonalization algorithms was presented under a number of varying conditions. The main conclusions of this comparative evaluation is that our algorithm is competitive with the best algorithms (i.e. extended Jacobi method and Pham's algorithm) in situations where additional constraints either on the class of solutions or the type of input data apply. In large-scale problems of non-orthogonal diagonalization FFDiag exhibits rapid convergence relative to gradient-based methods such as Yeredor's AC-DC algorithm, which is the only AJD method applicable under the same general conditions, i.e. without assuming orthogonality of the diagonalizer or positive-definiteness of the target matrices.

We also noticed that FFDiag may yield suboptimal solutions if the target matrices deviate extremely from a diagonalizable set. In this large-residual case it is recommended to modify the set of target matrices. This is the price that we have to pay to get the low computational complexity and the power to diagonalize matrices of dimensions in the hundreds of rows/columns, without imposing overly restrictive, additional assumptions, in cases where the model holds (small residual case).

Nevertheless, we expect that FFDiag will become a versatile tool for high-dimensional data analysis by taking advantage of the low computational cost to achieve good approximate solutions.

6.2 Future Work

6.2.1 Algorithms

Possible directions for future research are to further develop and tune related optimization algorithms, for example to combine gradient and Newton steps in a Levenberg-Marquardt or conjugate gradient scheme. In addition, it would be worth studying the fundamental differences between the various minimization criteria J_1,J_2 and J_3, seeking for some guidelines for choosing the most appropriate one.

Since the parameter of interest should be constrained to a particular matrix group, e.g. the special orthogonal group SO(N) or the special linear group SL(N), the AJD problem should be treated as a non-linear optimization problem on a manifold with a *group structure*. It turns out that the terminology of differential geometry and matrix algebra provides the right concepts for the development of efficient numerical algorithms that preserve those important features and always stay on the group manifold (Cardoso, 1998a). Thus, studying AJD algorithms as special cases of the isospectral flow methods in Helmke and Moore (1994); Hori (1999); Plumbley (2004) appears to be very promising.

Furthermore, it could be of interest to study the matrix exponential updates in the context of other Lie groups and their corresponding Lie al-

gebras, which opens up new possibilities of research in the intersection of optimization, signal processing and machine learning.

6.2.2 Biomedical Applications

From the biomedical applications point of view it would be an interesting goal for future research to incorporate prior knowledge into BSS models. In some preliminary studies, it has been observed that, as one departs from purely statistically based assumptions, one might get even closer to physiologically meaningful decompositions of electromagnetic brain signals. On the other hand, fitting neural sources in a classical framework, may be hard if some temporal overlap is present in their activations. Hence, a well balanced use of both, the model and appropriate priors, will yield a powerful exploratory decomposition technique that is able to extract meaningful information from high-dimensional biomedical data.

Furthermore it may be possible to apply the approximate joint diagonalization techniques to other problems of neurobiological modeling such as sparse coding in the visual cortex and taking advantage of an underlying group structure.

6.2.3 Other Applications

BSS methods have also been successfully applied to a variety of problems which—beyond biomedical signal processing—include diverse fields as telecommunications, feature extraction for pattern recognition, financial time-series analysis, data mining or image processing (see e.g. Cardoso et al. (1999); Hyvärinen et al. (2001)).

For certain applications, including the famous cocktail-party problem in auditory perception (von der Malsburg and Schneider, 1986), the instantaneous model in equation (2.1) is however too simplistic, since time-delays in the signal propagation are no longer negligible. Extended models to deal with such convolutive mixtures have been considered (e.g. Parra and Spence, 2000; Lee et al., 1998; Murata et al., 2001) and are promising future applications.

Appendix A

Notation

A.1 Abbreviations

AC-DC	Alternating Columns Direct Centers (diagonalization algorithm)
AJD	Approximate Joint Diagonalization
BSS	Blind Source Separation
CHESS	CHaracteristic function Enabled Signal Separation
DOMUNG	Diagonalization Of Matrices Using Natural Gradient
EEG	Electroencephalogram Electroencephalography
EVD	Eigen Value Decomposition
FastICA	Fast Independent Component Analysis
FFDiag	Fast Frobenius Diagonalization
FHG	Fraunhofer Gesellschaft
FIRST	Fraunhofer Institut für Rechnerarchitektur und Softwaretechnik
ICA	Independent Component Analysis
IDA	Intelligent Data Analysis
JADE	Joint Approximate Diagonalization of Eigen-matrices
MEG	Magneto-encephalogram or Magneto-encephalography
MI	Mutual Information
ML	Maximum Likelihood
MNG	Magneto-neurogram or Magneto-neurography
MRI	Magnetic Resonance Imaging
OFI	Optimal Filtering
PCA	Principal Component Analysis
PTB	Physikalisch-Technische Bundesanstalt
pdf	probability density function
SQUID	Superconducting Quantum Interference Device
TDSEP	Temporal Decorrelation SEParation

A.2 Mathematical Notation

Symbols

$\mathbf{A}, \ldots, \mathbf{Z}$	matrices
$\mathbf{a}, \ldots, \mathbf{z}$	vectors
$a, b, \ldots, z \;\; \alpha, \beta, \ldots, \omega$	scalars
$\mathbf{A}^T$	matrix transpose
$\mathbf{A}^{-1}$	matrix inverse
$\mathbf{A}^{-T}$	the inverse of $\mathbf{A}^T$
$\mathbf{A}^{(m)}$	matrix $\mathbf{A}$ in the m-th iteration
A_{ij}	matrix element of $\mathbf{A}$
diag($\mathbf{A}$)	diagonal matrix with same diagonal as $\mathbf{A}$
X	random variable
$\mathbf{I}$	identity matrix
$\mathbf{D}$	diagonal matrix
$\mathbf{P}$	permutation matrix
$\mathbf{A}$	mixing matrix
$\mathbf{V}$	demixing matrix or separating matrix
$\mathbf{X}$	data matrix
$\mathbf{C}$	covariance-like target matrix

Sets and Spaces

$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{R}^{(N\times N)}$	set of real matrices of dimension $N \times N$
$GL(N)$	General linear group
$SL(N)$	Special linear group
$O(N)$	Orthogonal group

Functions

$\mathrm{E}_X\{X\}$	expected value w.r.t X
$\phi_{X,Y}(\tau)$	cross-correlation function
cum$(\cdot, \ldots, \cdot)$	cumulant tensor
MI(X,Y)	mutual information between X and Y
H(X)	Shannon entropy of X
H(X—Y)	conditional entropy of X given Y
$\mathcal{O}(N)$	measure of the complexity of an algorithm

Appendix B

Some basic group theory

B.1 Matrix Lie Groups

The parameter space of the joint diagonalization problem (just as in the related ICA case) is not arbitrary, but possess a very favorable structure, known as a matrix Lie group (Cardoso, 1998a; Hori, 1999; Akuzawa and Murata, 2001; Plumbley, 2004). We have implicitly made use of this fact by introducing the matrix exponential update.

In the following we briefly restate some theoretical concepts taken from Grosche et al. (1995).

Definition (group). *A set $\mathcal{G}$ is a group for the operation $\bullet$ if:*

1. $\forall \mathbf{A}, \mathbf{B} \in \mathcal{G} \Rightarrow \mathbf{A} \bullet \mathbf{B} \in \mathcal{G}$*: $\mathcal{G}$ is closed.*
2. $\forall \mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathcal{G} \Rightarrow (\mathbf{A} \bullet \mathbf{B}) \bullet \mathbf{C} = \mathbf{A} \bullet (\mathbf{B} \bullet \mathbf{C})$*: $\mathcal{G}$ obeys the associative law.*
3. $\exists \mathbf{I} \in \mathcal{G}$ *so that* $\forall \mathbf{A} \in \mathcal{G}\ \mathbf{A} \bullet \mathbf{I} = \mathbf{I} \bullet \mathbf{A} = \mathbf{A}$*: $\mathcal{G}$ has a unit element.*
4. $\forall \mathbf{A} \in \mathcal{G}\ \exists \mathbf{A}^{-1} \in \mathcal{G}$ *so that* $\mathbf{A} \bullet \mathbf{A}^{-1} = \mathbf{A}^{-1} \bullet \mathbf{A} = \mathbf{I}$*: Each element in $\mathcal{G}$ has an inverse.*

The group is called Abelian or commutative if also holds:

5. $\forall \mathbf{A}, \mathbf{B} \in \mathcal{G} \Rightarrow \mathbf{A} \bullet \mathbf{B} = \mathbf{B} \bullet \mathbf{A}$

Definition (Lie group). *A group $\mathcal{G}$ is said to be a Lie group if its multiplication and inversion operation are continuous.*

Definition (Matrix group). *A subset of nonsingular matrices which are closed under matrix multiplication and inversion is called a matrix group.*

It turns out that every matrix group is in fact a Lie group, since the usual matrix multiplication and matrix inversion are smooth maps. The reason that Lie groups are interesting is because this particular entity combines

both algebraic and geometric structures. The most remarkable feature of a Lie group is that the structure is the same in the neighborhood of each of its elements. For this reason Lie group theory can provide powerful tools for designing and analyzing numerical optimization methods in structured parameter spaces. For example, it is possible to confine an iterative numerical algorithm to a certain Lie group by imposing conditions on a related structure, called a Lie algebra.

Definition (Lie algebra). *The set of all tangents at identity of a Lie group $\mathcal{G}$ forms a Lie algebra $\mathfrak{g}$, that is a linear space closed under commutation:*

1. $\mathbf{A}, \mathbf{B} \in \mathfrak{g} \Rightarrow \mathbf{A} + \mathbf{B} \in \mathfrak{g}$;
2. $\mathbf{A} \in \mathfrak{g}, \lambda \in I\!R \Rightarrow \lambda\mathbf{A} \in \mathfrak{g}$;
3. $\mathbf{A}, \mathbf{B} \in \mathfrak{g} \Rightarrow [\mathbf{A}, \mathbf{B}] \stackrel{def}{=} \mathbf{AB} - \mathbf{BA} \in \mathfrak{g}$

The most important relation for Lie groups and Lie algebras involves the matrix exponential function: If $\mathbf{W}$ belongs to a Lie algebra, then $e^{\mathbf{W}}$ is a matrix that belongs to the corresponding Lie group. In other words, there is an unique mapping from the Lie group to the Lie algebra and vice versa. For matrix Lie groups, this mapping is the matrix exponential.

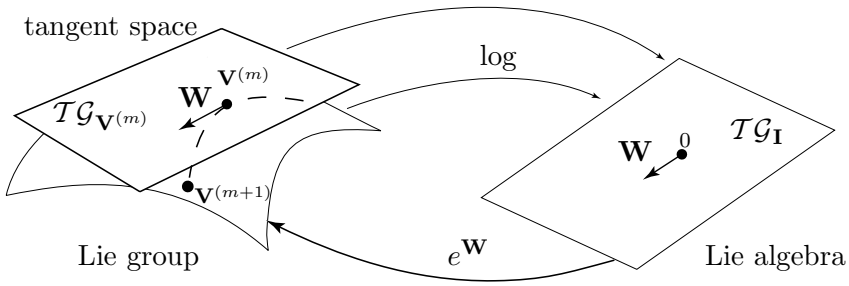

Figure B.1: Concept of a Lie group method: The Lie algebra determines the *local* structure of the Lie group via the exponential map.

Another important concept is the tangent space of a Lie group: it can be shown that at the identity this tangent space has always the structure of a Lie algebra (Helmke and Moore, 1994).

The important feature that we use in our algorithm is the fact that the Lie algebra determines the *local* structure of the Lie group via the exponential map. Thus we obtain a local parameterization of the group in terms of elements of the algebra.

Examples of Lie Groups and Lie Algebras

The classical illustrating example is the unit circle in $\mathbb{C}$. In this case the exponential function maps from the real line (the tangent space) to the circle (see Fig. B.2).

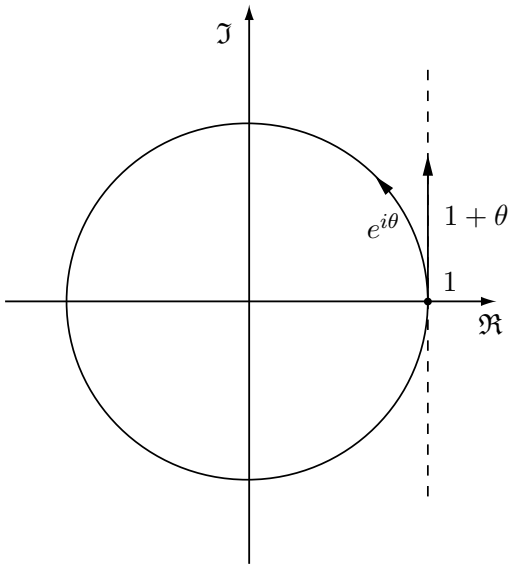

Figure B.2: Mapping from the real line to the circle.

Important examples of matrix Lie groups are:

- **the General Linear Group GL(N)** As indicated by the name, GL(N) is the most general matrix Lie group, in the sense that all other matrix Lie groups are subsets of GL(N). The corresponding Lie algebra $\mathfrak{gl}(N)$ is $I\!R^{N\times N}$.
- **the Special Linear Group SL(N)** is the group of all matrices with determinant one. The dimension is N^2-1.
- **the Orthogonal Group O(N)**) is the group of orthogonal $N \times N$ matrices.
- **the Special Orthogonal Group SO(N)** is the subset of O(N) with determinant one. It has the dimension $N(N-1)/2$.

Bibliography

T. Akuzawa and N. Murata. Multiplicative nonholonomic Newton-like algorithm. *Chaos, Solitons & Fractals*, 12:785f, 2001.

S. I. Amari, T.-P. Chen, and A. Cichocki. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural Computation*, 12:1463–1484, 2000.

S. I. Amari and A. Cichocki. Adaptive blind signal processing – neural network approaches. *Proceedings of the IEEE*, 9:2026–2048, 1998.

S. I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems (NIPS 95)*, volume 8, pages 882–893. The MIT Press, 1996.

S. I. Amari, S. Makino, and K. Matsuoka, editors. *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003. URL `http://ica2003.bsis.brain.riken.go.jp`.

T. Back, K. Kohno, and K.A. Hossmann. Cortical negative DC deflections following middle cerebral artery occlusion and KCl-induced spreading depression: effect on blood flow, tissue oxygenation and electroencephalogram. *J. Cereb. Blood Flow Metab.*, 14(1):12–19, 1994.

A. J. Bell and T. J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on SP*, 45(2):434–44, Feb 1997.

A. Belouchrani and M. Amin. Blind source separation based on time-frequency signal representations. *IEEE Trans. on Signal Processing*, 46 (11):2888–2897, 1998.

A. Bunse-Gerstner, R. Byers, and V. Mehrmann. Numerical methods for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 14(4):927–949, 1993.

J.-F. Cardoso. On the performance of orthogonal source separation algorithms. In *Proc. EUSIPCO*, pages 776–779, 1994.

J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.

J.-F. Cardoso. Learning in manifolds: the case of source separation. In *Proc. SSAP '98*, 1998a.

J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. ICASSP '98. Seattle*, 1998b.

J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, January 1999.

J.-F. Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. ICA 2001, San Diego*, 2001.

J.-F. Cardoso, Ch. Jutten, and Ph. Loubaton, editors. *First International Workshop on Independent Component Analysis and Signal Separation*, Aussois, France, Jan 1999.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings F*, 140(6):362–370, 1993.

J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164, January 1996.

Q. Chen, M. Chopp, H. Chen, and N. Tepley. Magnetoencephalography of focal ischemia in rats. *Stroke*, 23(9):1299–1303, Sep 1992.

S. Choi, A. Cichocki, and A. Belouchrani. Blind separation of second-order nonstationary and temporally colored sources. In *Proc. IEEE Workshop on Statistical Signal Processing (IEEE SSP 2001)*, pages 444–447, Singapore, 2001.

P. Comon. Independent component analysis, a new concept? *Signal Processing, Elsevier*, 36(3):287–314, 1994.

T.M. Cover and J.A. Thomas. *Elements of information theory.* John Wiley & Sons, Inc., New York, 1991.

G. Curio, S.M. Erné, M. Burghoff, K.-D. Wolff, and A. Pilz. Non-invasive neuromagnetic monitoring of nerve and muscle injury currents. *Electroencephalography and clinical Neurophysiology*, 89(3):154–160, 1993.

L. de Lathauwer. *Signal Processing by Multilinear Algebra*. PhD thesis, Faculty of Engineering, K. U. Leuven, Leuven, Belgium, 1997.

D. Drung. The PTB 83-SQUID-system for biomagnetic applications in a clinic. *IEEE Trans. Appl. Supercond.*, 5(2):2112–2117, 1995.

B. Flury and W. Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1): 169–184, January 1986.

K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, 1990. 2nd edition.

A. R. Gardner-Medwin, N. Tepley, G. L. Barkley, J. E. Moran, S. Nagel-Leiby, R. T. Simkins, and K. M. A. Welch. Magnetic fields associated with spreading depression in anaesthetized rabbits. *Brain Res.*, 540(1-2): 153–158, Feb 1991.

G. H. Golub and C. F. van Loan. *Matrix Computation*. The Johns Hopkins University Press, London, 1989.

G. Grosche, V. Ziegler, D. Ziegler, and E. Zeidler, editors. *Bronstein-Semendjajew - Teubner-Taschenbuch der Mathematik - Teil II - Neubearbeitung*, chapter 17. Liegruppen, Liealgebren und Elementarteilchen— Mathematik der Symmetrie, pages 643 – 704. B.G.Teubner Verlagsgesellschaft, Leipzig, 1995.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel feature spaces and nonlinear blind souce separation. In *Advances in Neural Information Processing Systems 14*, pages 761–768, 2001.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, May 2003.

S. Haykin, editor. *Unsupervised Adaptive Filtering, Vol. 1: Blind Source Separation*. Wiley, 2000.

U. Helmke and J.B. Moore. *Optimization and Dynamical Systems*. Springer Verlag, 1994.

G. Hori. Joint diagonalization and matrix differential equations. In *Proc. of NOLTA'99*, pages 675–678. IEICE, 1999.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press., Cambridge, 1985.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, New York, 2001.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

C. G. Jacobi. Über ein leichtes Verfahren, die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *Crelle J. reine angew. Mathematik*, 30:51–94, 1846.

M. Joho and H. Mathis. Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation. In *Proc. of IEEE Sensor Array and Multichannel Signal Processing Workshop SAM*, pages 273–277, 2002.

M. Joho and K. Rahbar. Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation. In *Proc. of IEEE Sensor Array and Multichannel Signal Processing Workshop SAM*, pages 403–407, 2002.

Ch. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

L. V. Kantorovich. On Newton's method. In *Trudy Mat. Inst. Steklov*, volume 28, pages 104–144. Interperiodica publishing, 1949. Translation: Selected Articles in Numerical Analysis by C. D. Benster, 104.1–144.2.

B. Laheld and J.-F. Cardoso. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

T. W. Lee, T. P. Jung, S. Makeig, and T. J. Sejnowski, editors. *Proc. 3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, USA, 2001.

T.W. Lee, A. Ziehe, R. Orglmeister, and T. .J. Sejnowski. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In *Proc. ICASSP98*, volume 2, pages 1249–1252, Seattle, May 1998.

K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, pages 164–168, 1944.

B.-M. Mackert, J. Mackert, G. Wübbeler, F. Armbrust, K.-D. Wolff, M. Burghoff, L. Trahms, and G. Curio. Magnetometry of injury currents from human nerve and muscle specimens using superconducting quantum interferences devices. *Neuroscience Letters*, 262(3):163–166, Mar 1999a.

B.-M. Mackert, G. Wübbeler, M. Burghoff, P. Marx, L. Trahms, and G. Curio. Non-invasive long-term recordings of cortical 'direct current' (DC-) activity in humans using magnetoencephalography. *Neuroscience Letters*, 273(3):159–162, Oct 1999b.

S. Makeig, A.J. Bell, T.-P. Jung, and T.J. Sejnowski. Independent component analysis of electroencephalographic data. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems (NIPS'95)*, volume 8, pages 145–151. The MIT Press, 1996.

S. Makeig, T-P. Jung, D. Ghahremani, A.J. Bell, and T.J. Sejnowski. Blind separation of event-related brain responses into independent components. *Proc. Natl. Acad. Sci. USA*, 94:10979–10984, 1997.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of SIAM*, 11(2):431–441, jun 1963.

K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8:411–419, 1995.

F. C. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. Estimating the reliability of ICA projections. In *Advances in Neural Information Processing Systems 14*, pages 1181–1188, 2001.

F. C. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A Resampling Approach to Estimate the Stability of one- or multidimensional Independent Components. *IEEE Trans. on Biomedical Engineering*, 49(12): 1514–1525, 2002.

F. C. Meinecke, A. Ziehe, J. Kurths, and K.-R. Müller. Measuring phase synchronization of superimposed signals. *Physical Review Letters*, 94(8): 084102, March 2005.

L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23): 3634–3637, 1994.

E. Moreau. A generalization of joint-diagonalization criteria for source separation. *IEEE Trans. on Signal Processing*, 49(3):530–541, March 2001.

K.-R. Müller, N. Murata, A. Ziehe, and S.-I. Amari. *On-line learning in neural networks*, chapter On-line learning in Switching and Drifting environments with application to blind source separation, pages 93–110. Cambridge University Press, 1998.

K.-R. Müller, R. Vigário, F. C. Meinecke, and A. Ziehe. Blind source separation techniques for decomposing event-related brain signals. *International Journal of Bifurcation and Chaos*, 14(2):773–791, 2004.

N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4): 1–24, August 2001.

N. Murata, M. Kawanabe, A. Ziehe, K.-R. Müller, and S. I. Amari. On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks*, 15(4-6):743–760, 2002.

B. Noble and W. Daniel. *Applied matrix algebra*. Prentice Hall, Inc., Englewood Cliffs, NJ, 1977.

G. Nolte, A. Ziehe, and K. R. Müller. Noise robust estimates of correlation dimension and K_2 entropy. *Physical review E*, 64(1):016112, June 2001.

P. Pajunen and J. Karhunen, editors. *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Source Separation*. Helsinki Univ. of Technology, Lab. of Computer and Information Science, Espoo, Finland, June 19-22, 2000. 647 pages.

C. Pantev, C. Eulitz, S. Hampson, B. Ross, L.E., and Roberts. The auditory evoked off response: source and comparison with the on and the sustained responses. *Ear & Hearing*, 17(3):255–265, Jun 1996.

L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, 8(3): 320–327, 2000.

L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller. Unmixing hyperspectral data. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 942–948, 1999.

D.-T. Pham. Joint approximate diagonalization of positive definite matrices. *SIAM J. on Matrix Anal. and Appl.*, 22(4):1136–1152, 2001.

D.-T. Pham. Exploiting source non-stationary and coloration in blind source separation. In *Proceedings of the DSP2002 conference*, pages 151–154, Santorini, Greek, July 2002.

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non-stationary sources. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 187–193, Helsinki, Finland, 2000.

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Sig. Proc.*, 49(9):1837–1848, 2001.

D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7):1712–1725, 1997.

T.W. Picton, D.L. Woods, and G.B. Proulx. Human auditory sustained potentials: part I and II. *Electroencephalography and clinical Neurophysiology*, 45(2):186–210, Aug 1978.

M. D. Plumbley. Lie group methods for optimization with orthogonality constraints. In *Proc. ICA 2004*, volume 3195 of *Lecture Notes in Computer Science*, pages 1245 – 1252, Oct 2004.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipies in C*. Cambridge University Press., Cambridge, 1992.

C. G. Puntonet and A. Prieto, editors. *Independent Component Analysis and Blind Signal Separation: Fifth International Conference (ICA 2004)*, volume 3195 of *Lecture Notes in Computer Science*, Granada, Spain, 2004. Springer Verlag.

L. Tong, V.C. Soon, and Y. Huang. Indeterminacy and identifiability of identification. *IEEE Trans. on Circuits and Systems*, 38(5):499–509, 1991.

M. Uusitalo and R. Ilmoniemi. Signal-space projection method for separating meg or eeg into components. *Med. & Biol. Eng. & Comput.*, 10(35): 135–140, Oct. 1997.

A.-J. van der Veen. Joint diagonalization via subspace fitting techniques. In *Proc. ICASSP*, volume 5, 2001.

R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

R. Vigário, A. Ziehe, K.-R. Müller, J. Särelä, E. Oja, V. Jousmäki, G. Wübbeler, L. Trahms, B.-M. Mackert, and G. Curio. *Advances in Exploratory analysis and data modeling in functional neuroimaging*, chapter Blind decomposition of multimodal and DC evoked responses. MIT Press, Cambridge MA., December 2002. ISBN 0-262-19481-3.

R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE transactions on biomedical engineering*, 47(5):589–593, 2000.

C. von der Malsburg and W. Schneider. A neural cocktail-party processor. *Biological Cybernetics*, 54:29–40, 1986.

H. A. Van der Vorst and G. H. Golub. 150 years old and still alive: Eigenproblems. In *The State of the Art in Numerical Analysis*, volume 63, pages 93–120. Oxford University Press, 1997.

O. Weiss, A. Ziehe, and H. Herzel. Optimizing property codes in protein data reveals structural characteristics. In *Proc. International Conference on Artificial Neural Networks*, pages 245–252. Springer Verlag, 2003.

H.-C. Wu and J.C. Principe. Simultaneous diagonalization in the frequency domain (SDIF) for source separation. In *Proc. First International Conference on Independent Component Analysis and Blind Source Separation ICA 99*, pages 245–250, Aussois, France, January 11–15, 1999.

G. Wübbeler, B.-M. Mackert, F. Armbrust, M. Burghoff, P. Marx, G. Curio, and L. Trahms. Measuring para-DC biomagnetic fields of the head using a horizontal modulated patient cot. *Biomed Tech (Berl)*, Suppl(43):232–233, 1999. in german.

G. Wübbeler, J. Mackert, F. Armbrust, M. Burghoff, B.-M. Mackert, K.-D. Wolff, J. Ramsbacher, G. Curio, and L. Trahms. SQUID measurements of human nerve and muscle near-DC injury-currents using a mechanical modulation of the source position. *Applied Superconductivity*, 6(10-12): 559–565, 1998.

G. Wübbeler, A. Ziehe, B.-M. Mackert, K.-R. Müller, L. Trahms, and G. Curio. Independent component analysis of non-invasively recorded cortical magnetic DC-fields in humans. *IEEE Transactions on Biomedical Engineering*, 47(5):594–599, 2000.

A. Yeredor. Blind source separation via the second characteristic function. *Signal Processing*, 80(5):897–902, 2000.

A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, July 2002.

A. Yeredor, A. Ziehe, and K.-R. Müller. Approximate joint diagonalization using a natural gradient approach. In *Proc. ICA 2004*, Lecture Notes in Computer Science, pages 89–96, 2004.

M. Zibulevsky. Relative Newton method for quasi-ML blind source separation. In *Proc. 4th Intern. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 897–902, Nara, Japan, 2003.

A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller. Blind separation of post-nonlinear mixtures using gaussianizing transformations and temporal decorrelation. In *Proc. ICA 2003*, pages 269–274, Nara, Japan, Apr 2003a.

A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller. Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation. *Journal of Machine Learning Research*, 4:1319–1338, Dec 2003b.

A. Ziehe, P. Laskov, K.-R. Müller, and G. Nolte. A linear least-squares algorithm for joint diagonalization. In *Proc. ICA 2003*, pages 469–474, Nara, Japan, Apr 2003c.

A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, 5: 777–800, 2004.

A. Ziehe and K.-R. Müller. TDSEP–an efficient algorithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skövde, Sweden, 1998.

A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Trans. Biomed. Eng.*, 47(1):75–87, January 2000a.

A. Ziehe, G. Nolte, G. Curio, and K.-R. Müller. OFI: Optimal filtering algorithms for source separation. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 127–132, Helsinki, Finland, 2000b.

A. Ziehe, G. Nolte, T. Sander, K.-R. Müller, and G. Curio. A comparison of ICA-based artifact reduction methods for MEG. In Jukka Nenonen, editor, *Recent Advances in Biomagnetism, Proc. of the 12th International conference on Biomagnetism*, pages 895–898, Espoo,Finland, 2001. Helsinki University of Technology.