

Helsinki University of Technology  
Dissertations in Computer and Information Science  
Espoo 2004

Report D6

## **EXPLORATORY SOURCE SEPARATION IN BIOMEDICAL SYSTEMS**

Jaakko Särelä

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 29th of October, 2004, at 12 o'clock noon.

Helsinki University of Technology  
Department of Computer Science and Engineering  
Laboratory of Computer and Information Science  
P.O.Box 5400  
FIN-02015 HUT  
FINLAND

Distribution:

Helsinki University of Technology

Laboratory of Computer and Information Science

P.O.Box 5400

FIN-02015 HUT

FINLAND

Tel. +358-9-451 3272

Fax +358-9-451 3277

<http://www.cis.hut.fi/>

Available in pdf format at <http://lib.hut.fi/Diss/2004/isbn9512273438/>

© Jaakko Särelä

ISBN 951-22-7342-X (printed version)

ISBN 951-22-7343-8 (electronic version)

ISSN 1459-7020

Otamedia Oy

Espoo 2004

Särelä, J. (2004): **Exploratory source separation in biomedical systems**. Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D6, Espoo, Finland.

**Keywords:** exploratory source separation, independent component analysis, blind source separation, denoising, denoising source separation, biomedical systems, biomedical data, magnetoencephalograms, electroencephalograms.

## Abstract

Contemporary science produces vast amounts of data. The analysis of this data is in a central role for all empirical sciences as well as humanities and arts using quantitative methods. One central role of an information scientist is to provide this research with sophisticated, computationally tractable data analysis tools.

When the information scientist confronts a new target field of research producing data for her to analyse, she has two options: She may make some specific hypotheses, or guesses, on the contents of the data, and test these using statistical analysis. On the other hand, she may use general purpose statistical models to get a better insight into the data before making detailed hypotheses.

Latent variable models present a case of such general models. In particular, such latent variable models are discussed where the measured data is generated by some hidden sources through some mapping. The task of *source separation* is to recover the sources. Additionally, one may be interested in the details of the generation process itself.

We argue that when little is known of the target field, *independent component analysis* (ICA) serves as a valuable tool to solve a problem called *blind source separation* (BSS). BSS means solving a source separation problem with no, or at least very little, prior information. In case more is known of the target field, it is natural to incorporate the knowledge in the separation process. Hence, we also introduce methods for this incorporation. Finally, we suggest a general framework of *denoising source separation* (DSS) that can serve as a basis for algorithms ranging from almost blind approach to highly specialised and problem-tuned source separation algorithms. We show that certain ICA methods can be constructed in the DSS framework. This leads to new, more robust algorithms.

It is natural to use the accumulated knowledge from applying BSS in a target field to devise more detailed source separation algorithms. We call this process *exploratory source separation* (ESS). We show that DSS serves as a practical and flexible framework to perform ESS, too.

Biomedical systems, the nervous system, heart, etc., constitute arguably the most complex systems that human beings have ever studied. Furthermore, the

contemporary physics and technology have made it possible to study these systems while they operate in near-natural conditions. The usage of these sophisticated instruments has resulted in a massive explosion of available data. In this thesis, we apply the developed source separation algorithms in the analysis of the human brain, using mainly magnetoencephalograms (MEG). The methods are directly usable for electroencephalograms (EEG) and with small adjustments for other imaging modalities, such as (functional) magnetic resonance imaging (fMRI), too.

## Preface

This work has been carried out at the Neural Networks Research Centre, hosted by the Laboratory of Computer and Information Science at Helsinki University of Technology. In addition to the funding from HUT, this work has been funded by the Finnish Academy, European Union (IST-1999-14190) and Tekniikan edistämissäätiö.

The laboratory and the Neural Networks Research Centre has been founded by Academician, Professor Emeritus Teuvo Kohonen. He is one of the pioneers in neural networks and machine learning. His work has given me great inspiration. I am also grateful for the outstanding facilities provided for researchers here.

During my stay in this laboratory, I have worked under the supervision of Academy Professor Erkki Oja. As the head of the laboratory and later as the head of the research centre, he has created a truly magnificent atmosphere for conducting top-quality research. Especially I would like to thank him for giving the researchers an opportunity to concentrate on the research and not so much in administrative issues such as funding. This is a rare luxury nowadays.

The research in this thesis has been conducted in very close collaboration with my instructors, Dr. Ricardo Vigário and Dr. Harri Valpola. I have found working with them seamless and effective, but most of all fun. I also want to thank them for sharing non-work related life with me.

The data for the application in biomedical systems has been collected at the Brain Research Unit in the Low Temperature Laboratory in this university. I would like to thank the colleagues there for the rare possibility to participate in designing the experiments. In particular, I would like to thank Academy Professor Riitta Hari, Dr. Matti Hämäläinen and Dr. Veikko Jousmäki.

I have worked with many people in this laboratory during my ten-year stay here. Even more numerous are the people with whom I have shared social life, such as lunch hours, coffee breaks and occasional billiard evenings. As it would be impossible to mention all of their names, I thank them collectively.

This work has been pre-examined by Professor Te-Won Lee and Dr. Ole Jensen. Their comments have lead to many improvements on the quality of this thesis.

I would also like to thank my parents and my brothers and sisters for all the exciting moments we shared solving some puzzles and riddles we used to give to each other. I am sure they have a major role in my decision to enter a research-oriented career.

Finally, I am extremely grateful to my wife, Katja. Without her, this project would never have finished.

Otaniemi, October 2004



Jaakko Särelä

---

# Contents

<b>Mathematical notation</b>	<b>10</b>
<b>Abbreviations</b>	<b>11</b>
<b>1 Introduction</b>	<b>12</b>
1.1 Motivation and overview . . . . .	12
1.2 Publications of the thesis . . . . .	14
<b>2 Mathematical modelling</b>	<b>18</b>
2.1 Data . . . . .	18
2.2 Modelling of the data . . . . .	19
2.3 Probabilistic models . . . . .	21
2.3.1 Events . . . . .	21
2.3.2 Probabilities of continuous variables . . . . .	22
2.3.3 Likelihood of the data . . . . .	26
2.3.4 Bayesian modelling . . . . .	26
2.3.5 Full distribution approaches to Bayesian modelling . . . . .	27
2.4 Gradient based optimisation methods . . . . .	28
2.4.1 Gradient descent and ascent . . . . .	29
2.4.2 Fixed point algorithms . . . . .	30
<b>3 Separation of linearly mixed sources</b>	<b>31</b>
3.1 Principal component analysis . . . . .	33
3.1.1 Power method . . . . .	36
3.1.2 Extracting several components . . . . .	37
3.1.3 Spectral shift . . . . .	37
3.1.4 Nonlinear principal component analysis . . . . .	38
3.2 Independent component analysis . . . . .	39
3.2.1 FastICA: ICA by maximisation of non-Gaussianity . . . . .	40
3.2.2 ICA by maximum likelihood estimation . . . . .	43

---

3.2.3	Some other methods for ICA . . . . .	46
3.2.4	ICA models considering noise explicitly . . . . .	47
3.2.5	Bayesian ICA using ensemble learning . . . . .	48
3.3	Temporal methods . . . . .	50
3.4	Dynamical factor analysis . . . . .	51
3.5	Relaxing the ICA assumptions . . . . .	52
<b>4</b>	<b>Denoising source separation, a new approach</b>	<b>54</b>
4.1	A source separation example using DSS . . . . .	55
4.2	Linear DSS . . . . .	57
4.3	Nonlinear DSS . . . . .	59
4.4	Denoising functions in practice . . . . .	60
4.4.1	Detailed linear denoising functions . . . . .	61
4.4.2	ICA using DSS . . . . .	63
4.4.3	Other denoising functions . . . . .	64
4.5	Speedup in DSS . . . . .	65
4.6	Separation of artificial signals: comparison of DSS algorithms . . . . .	68
4.6.1	Linear denoising . . . . .	68
4.6.2	Nonlinear exploratory denoising . . . . .	69
4.6.3	Separation results . . . . .	69
<b>5</b>	<b>Overfitting</b>	<b>72</b>
5.1	Overfitting in marginal-distribution-based ICA . . . . .	73
5.1.1	Are we saved if $T > M$ ? . . . . .	74
5.1.2	Bumps emerge when low frequencies dominate . . . . .	75
5.1.3	Bumps are the overfitting in magnetoencephalograms . . . . .	76
5.2	Attempts to solve the problems in ICA . . . . .	77
5.2.1	Proper estimate of the ICA model . . . . .	77
5.2.2	Additions to the model . . . . .	78
5.3	Bayesian analysis of the problems of spikes and bumps . . . . .	80
5.3.1	Avoiding spikes . . . . .	80
5.3.2	Reducing the effects of bumps . . . . .	81
5.4	Conclusions on overfitting in marginal-distribution-based ICA . . . . .	82
5.5	Overfitting in DSS . . . . .	83
<b>6</b>	<b>Biomedical systems</b>	<b>84</b>
6.1	Brain imaging techniques . . . . .	84
6.1.1	Early techniques . . . . .	84
6.1.2	Modern techniques . . . . .	85
6.1.3	Basics of magnetoencephalograms . . . . .	86
6.2	Analysis-synthesis cycle . . . . .	88



---

6.3 Analysis-synthesis in extraction of MEG sources . . . . .	89
6.3.1 Adaptive extraction of the component with wandering frequency content . . . . .	91
6.3.2 Adaptive extraction of cardiac subspace in MEG . . . . .	92
<b>7 Conclusions and future trends</b>	<b>95</b>
<b>References</b>	<b>98</b>

## Mathematical notation

lower- or upper-case		scalar, constant or scalar function
bold-face lower-case		column or row vector, vector-valued function
bold-face upper-case		matrix, matrix-valued function
<b>A</b>	$M \times N$	mixing matrix
<b>a</b>	$M \times 1$	a column mixing vector
<b>a<sub>j</sub></b>	$M \times 1$	<i>i</i> th column mixing vector
<i>a<sub>ij</sub></i>	scalar	mixing coefficient of the <i>j</i> th source in <i>i</i> th observation
<b>B</b>	$N \times M$	demixing matrix
<b>D</b>	$T \times T$	linear denoising matrix applied to the source estimate
<b>D*</b>	$T \times T$	linear denoising matrix applied to the whole data
<b>E</b>	$L \times M$	matrix of the eigenvectors
<b>e</b>	$M \times 1$	a column eigenvector
<b>e<sub>l</sub></b>	$M \times 1$	<i>l</i> th column eigenvector
<b>f(.)</b>	$1 \times T$	denoising function
<b>g(.)</b>	scalar	objective function
<i>i, j, l</i>	scalar	general purpose indices, usually <i>i</i> refers to observations, <i>j</i> to sources and <i>l</i> to sphered data
<i>L</i>	scalar	number of retained principal components <b>y<sub>l</sub></b>
<i>M</i>	scalar	number of observations <b>x<sub>i</sub></b>
<i>N</i>	scalar	number of sources <b>s<sub>j</sub></b>
<i>ν</i>	scalar	Gaussian variable
<b>ν<sub>i</sub></b>	$1 \times T$	<i>i</i> th additive noise term
<b>ν</b>	$M \times T$	additive noise matrix
<b>S</b>	$N \times T$	matrix of <i>N</i> sources with <i>T</i> samples
<b>s</b>	$1 \times T$	a row vector consisting of a source
<b>s<sub>j</sub></b>	$1 \times T$	a row vector consisting of the <i>j</i> th source
<i>s<sub>j</sub>(t)</i>	scalar	value of the <i>j</i> th source at (time) index <i>t</i> .
<b>s(t)</b>	$N \times 1$	a column vector containing the values of all of the sources at time instance <i>t</i>
<i>T</i>	scalar	number of samples in sources <b>s<sub>i</sub></b> , and observations <b>x<sub>i</sub></b>
<b>θ</b>	vector	a set of model parameters
<b>V</b>	$L \times M$	sphering matrix
<b>v</b>	$M \times 1$	a column sphering vector
<b>v<sub>l</sub></b>	$M \times 1$	<i>l</i> th column sphering vector
<i>v<sub>li</sub></i>	scalar	sphering coefficient of the <i>i</i> th observation in the <i>l</i> th principal component
<b>W</b>	$N \times L$	demixing (separating) matrix (from the sphered data)
<b>w</b>	$M \times 1$	a column demixing vector

---

$\mathbf{w}_j$	$M \times 1$	$j$ th column demixing vector
$w_{ji}$	scalar	demixing coefficient of the $i$ th observation in the $j$ th source
$\mathbf{X}$	$M \times T$	matrix of $M$ observations with $T$ samples
$\mathbf{x}$	$1 \times T$	a row vector consisting of an observation
$\mathbf{x}_i$	$1 \times T$	a row vector consisting of the $i$ th observation
$x_i(t)$	scalar	value of the $i$ th observation at (time) index $t$ .
$\mathbf{x}(t)$	$M \times 1$	a column vector containing the values of all of the observations at time instance $t$
$\mathbf{Y}$	$L \times T$	matrix of sphered components
$\mathbf{y}$	$1 \times T$	a row vector consisting of a sphered component $\mathbf{y}_1$
$\mathbf{y}_l$	$1 \times T$	a row vector consisting of the $l$ th sphered component
$\mathbf{Z}$	$M \times T$	denoised data

## Abbreviations

BSS	blind source separation
CLT	central limit theorem
CT	computer axial tomogram
DSS	denoising source separation
ECD	equivalent current dipole
EEG	electroencephalogram
ESS	exploratory source separation
FA	factor analysis
fMRI	functional magnetic resonance imaging
ICA	independent component analysis
MAP	maximum a posteriori
MEG	magnetoencephalogram
ML	maximum likelihood
MLP	multi-layer perceptron
MoG	mixture of Gaussians
MRI	magnetic resonance imaging
NPCA	nonlinear principal component analysis
PCA	principal component analysis
pdf	probability density function
PET	positron emission tomogram
SNR	signal-to-noise ratio

## Chapter 1

# Introduction

*Complete knowledge always involves an apparent circle, that each part can be understood only out of the whole to which it belongs, and vice versa.*

–Chladenius (1742)

### 1.1 Motivation and overview

Science mainly advances through continuous alternation between experimentation and suggesting of new explanations for the data that is observed in the experiments. One of the biggest questions of philosophy of science throughout centuries, even millennia, has been how the explanations are arrived at from the observations. This is the realm of modelling.

Classical statistics usually advances from the observations to the explanations by generating detailed models or hypotheses. The validity of these hypotheses is then tested against contradicting null-hypotheses. While this kind of hypothesis testing may be very reliable and useful in certain situations, it does not provide a researcher with much information. Basically only one binary decision can be made: either to accept the hypothesis or to discard it, i.e. to accept the null-hypothesis. Modern statistics offers a researcher with better alternatives. Especially in Bayesian probability theory, uncertainty can be taken into account in a flexible manner, allowing one to estimate more general models from the data. This is very useful for a researcher who wants to get good insight into the data but has no good binary hypotheses a priori. Such research is often called *exploratory data analysis*.

Linear models constitute a special class of general models because of their tractable analytical properties. In this thesis, we discuss the problem of *linear source separation*. In linear source separation, the model consists of two parts: a

set of sources and a linear mapping that links the sources to the observations. In case we want to solve the source separation problem in an exploratory manner, when little is known of the target field, we call the problem *exploratory source separation* (ESS). There is a need to perform the first phase of the ESS process blindly. This means that one wants to fit a general linear model to the data without knowing almost anything of the sources nor of the linear mapping. This process is usually called *blind source separation* (BSS).

We discuss the use of *independent component analysis* (ICA) to solve the BSS problem. We show that it is a reliable and robust way to solve BSS and that it provides the researcher with good insight into the data for subsequent modelling.

Often, the researcher already knows more because of the accumulated research in the target field. She may know some source characteristics, e.g. the nature of their marginal distributions or their time structure, or she may possess some prior knowledge of the linear observation mapping. In this case, the incorporation of this prior knowledge should lead to a more accurate solution of the source separation problem. Furthermore, the use of the prior knowledge often makes it possible to obtain the results faster. We suggest ways to incorporate this prior knowledge in the source separation algorithms. In particular, we suggest a novel framework of denoising source separation (DSS), where this incorporation is simple and practical to achieve and which leads to fast algorithms. In DSS, the source separation algorithms are constructed around denoising methods of the source estimates. Implementation of a denoising for the source estimates is often suggested by the prior knowledge one possesses. We also suggest DSS algorithms for BSS. In fact, we show that certain ICA algorithms can be derived under the DSS framework, leading to improved stability and speedup. This makes DSS a good candidate as a valuable general framework for ESS as well.

Biomedical systems are arguably the most complex systems science has ever studied. For example, the human body consist of several complex subsystems such as the central nervous system (CNS), the heart and the lungs. Several fields of science, such as neuroscience, biology and biochemistry, have concentrated on some particular parts of the biomedical systems. Furthermore, the human behaviour has interested researchers for a long time and there is substantial evidence that it is closely related to the CNS, especially to the brain.

Recent advances in physics and technology have made it possible to study these systems while they operate in near-natural conditions. The usage of these sophisticated instruments has resulted in a massive explosion of available data. In this thesis, we concentrate on developing analysis tools for this data and studying biomedical systems using them. We mainly apply these ESS algorithms in the analysis of the human brain, especially using magnetoencephalograms (MEG). It is straightforward to extend the use of the developed methods for electroencephalograms (EEG) and with small adjustments to data from other measuring

devices, such as (functional) magnetic resonance imaging (fMRI).

The analysis of biomedical systems using source separation algorithms leads to an accumulation of knowledge. This knowledge may be used to generate functional models for the very systems that were studied, leading to a synthesis. This *analysis-synthesis* process plays a central role in most scientific research. We show that this synthesis may as well lead to more accurate subsequent analysis.

The introduction part of this thesis is organised as follows: In the very beginning, there is a list of the most central mathematical notations and some abbreviations. In Ch. 2, we review some needed mathematical concepts and introduce the notation used in this thesis. In Ch. 3, the linear source separation problem is reviewed. We concentrate especially in describing the problem and reviewing previously existing work used to solve it. In Ch. 4, the DSS framework for source separation algorithms is introduced and its connection to some existing algorithms is discussed. In Ch. 5, an overfitting problem arising in practical uses of source separation algorithms is discussed in detail in the case of ICA and somewhat more generally in the case of DSS. Finally, in Ch. 6, we consider the study of biomedical systems and apply the algorithms presented and developed in earlier sections. Illustrations will mainly use magnetoencephalograms (MEG).

The author's contribution in the introduction part is concentrated in the Chs. 4–6 whilst the first chapters mainly consist of a review of existing work. The author's contribution and literature survey are intertwined in the biomedical-systems chapter (6) for belletristic reasons. Following the introduction part, there is a set of published papers. These and the author's contribution in them are introduced next.

## 1.2 Publications of the thesis

The thesis consists of six publications and an introduction part. The introduction aims to give a general description of the problem and the proposed solutions without going into detailed derivations. At the appropriate places, the publications are referred to from the introduction. This does not mean that in order to understand the introduction part, one needs to read the articles. On the contrary, in cases of forward references to the articles, we aim to give pointers to where deeper analysis of the issues can be found.

The following six publications describe the development of methods suitable for exploratory source separation in biomedical systems.

**Publication 1.** 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*, vol. 47, no. 5, pp. 589 – 593, 2000.

**Publication 2.** A. Hyvärinen, J. Särelä, and R. Vigário, “Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size”, *Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation, ICA ’99*, (Aussois, France), pp. 425 – 429, 1999.

**Publication 3.** J. Särelä and R. Vigário, ”Overlearning in marginal distribution-based ICA: analysis and solutions”, *Journal of Machine Learning Research*, vol. 4 (Dec), pp. 1447 – 1469, 2003.

**Publication 4.** J. Särelä, H. Valpola, R. Vigário and E. Oja, ”Dynamical Factor Analysis of Rhythmic Magnetoencephalographic Activity”, *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA ’01* (San Diego, California, USA), pp. 451 – 456, 2001.

**Publication 5.** J. Särelä and H. Valpola, ”Denoising source separation”, *Journal of Machine Learning Research*, accepted with minor revision, revised, 2004.

**Publication 6.** H. Valpola and J. Särelä, ”Accurate, fast and stable denoising source separation algorithms”, *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA ’04*, (Granada, Spain), pp. 65 – 72, 2004.

The content and the contribution of the present author in the above-mentioned papers are as follows:

In **Publication 1**, the applicability of ICA in analysis of MEG is covered thoroughly. The suitability of the assumptions of the ICA model for the physical and functional description of the MEG measurements and corresponding brain activity is extensively discussed. Several already published results are reviewed. For instance, it is demonstrated that ICA is capable of identifying several non-brain-activity related artefacts. Furthermore, ICA is shown to be useful in segmenting event related MEG data to physiologically meaningful components. Finally, ICA is shown to be capable of separating activity from different modalities such as responses to auditory and somatosensory stimulation.

In this paper, the author was highly involved in the experiments and discussed extensively the significance of different results with Dr.Vigário. He also participated in the writing of the manuscript.

**Publication 2** is the first paper ever discussing an overfitting problem in ICA. It is noticed that with insufficient sample sizes all ICA algorithms tend to produce

results where there is a single spike or bump and little activity elsewhere. Some preliminary solutions to this problem are suggested.

The present author noticed the problem together with Dr. Vigário in the analysis of real MEG data. He participated in analysing the problem and finding solutions. He was responsible for carrying out the experiments with artificial data and participated in editing the manuscript.

In **Publication 3**, a comprehensible and thorough analysis of the problem of overfitting (overlearning) suggested in Publication 2 is presented. The impact of different data characteristics such as the length of the data, the number of dimensions and correlations between different samples are analysed both with mathematical formulation and extensive experiments. Several solutions are presented.

The present author was responsible for the mathematical formulation of the problem and the analysis of potential solutions as well as all of the experiments conducted in the paper. He was mainly responsible in writing the manuscript, too. Dr. Vigário participated in the research very actively in all of its stages.

**Publication 4** proposes the use of a Bayesian technique of ensemble learning in the dynamical analysis of rhythmic MEG activity. In addition to the usual linear observation mapping, a nonlinear feedforward network is used to model the dynamics of the underlying sources. The network for the dynamics was not fully connected, but rather connected in smaller blocks. The results of the paper show that the frequency content of the oscillatory activity in the brain has several significant frequencies. This makes the use of linear dynamics insufficient.

The present author suggested the use of the nonlinear state-space model introduced earlier by Dr. Valpola for the analysis of rhythmic MEG. Dr. Valpola had suggested a simplification of the model dynamics. Based on that, the author and Dr. Valpola developed the block-wise dynamics. The author was responsible for carrying out all of the experiments as well as writing the manuscript, Dr. Valpola and the other authors participating in the editing of the paper.

In **Publication 5**, a denoising-source-separation framework is proposed. It is shown that source separation algorithms can be constructed around denoising principles. This framework allows for easy incorporation of prior knowledge to guide the search for the sources, which makes it possible to design source-separation algorithms ranging from highly specialised to almost blind approaches. It is shown that some ICA algorithms can be seen as special cases of this framework. In particular, some extensions to the FastICA algorithm are proposed. The proposed DSS algorithms are extensively applied to both artificial and real MEG data.

This paper shows seamless collaboration between the authors and it is therefore difficult to pinpoint the actual contributions of the present author. While Dr. Valpola initially suggested the framework and is responsible for most of



---

the theoretical contribution, the present author has been actively participating in developing it. Furthermore, the present author is responsible for conducting almost all of the experiments in the paper, though the other author has significantly contributed in them as well. The manuscript has been written in very close collaboration between the authors.

**Publication 6** proposes several practical algorithms derived from the DSS framework. In particular, the extensions to the FastICA algorithm in Publication 5 are developed further. In addition, the accuracy, stability and the speed of convergence of the different algorithms are studied. It is shown that the proposed extensions achieve stability and fast convergence, even faster than with FastICA, simultaneously.

The present author participated in developing the extensions to the particular DSS algorithm, conducted part of the experiments and participated actively in the writing of the manuscript.

## Chapter 2

# Mathematical modelling

*All models are wrong but some are useful.*

–Box (1979)

In this section, we cover some basic mathematical concepts that are necessary for the rest of this thesis. We aim to generality but we introduce some notation for further use, as well.

## 2.1 Data

Mathematical description of data is almost essential for understanding complex phenomena in nature. Consider a set of  $\mathcal{T}$  observations:  $\boldsymbol{\xi} = [\xi_1 \cdots \xi_i \cdots \xi_{\mathcal{T}}]$ . The set of observations  $\boldsymbol{\xi}$  is unordered, i.e. nothing is assumed of the order in which the observations come. Often, however, the measuring process implies some structure for the data. For instance, in many cases, there exist several measurement devices and it is beneficial to structure the data in different sets, one row vector for each measurement device:  $\mathbf{x}_i = [x_i(1) \cdots x_i(t) \cdots x_i(T_i)]$ . It is allowed that different measurement devices have different amounts of data samples  $T_i$ .  $T$  often refers to time, i.e. the measurements are ordered in chronological order, but this is not mandatory, it can refer to other structure such as space or just serve as a general indexing variable. Throughout this thesis, we call this index time for simplicity reasons.

When the samples of the different measurement devices can be grouped, it may be beneficial to have a mathematical notation for those groups too. We denote such groups by a column vector  $\mathbf{x}(t) = [x_1(t) \cdots x_i(t) \cdots x_M(t)]^T$ , where  $M$  different measurement devices are assumed. Note that this is a column vector in contrast to the row vector  $\mathbf{x}_i$  that has been used to denote the observations of one particular measuring device. Furthermore the notation of the vector  $\mathbf{x}(t)$

contains the index  $t$  to indicate that the observations at time  $t$  are collected in the vector. In this thesis, we mainly use the first notation using column vectors. The second notation, with row vectors, is used only when the column-vector notation would lead to cumbersome formulae.

When each measurement device produces the same amount of samples, it is convenient to collect all of the data in a matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_M \end{bmatrix} = [\mathbf{x}(1) \quad \cdots \quad \mathbf{x}(t) \quad \cdots \quad \mathbf{x}(T)], \quad (2.1)$$

where both the column and the row vector notations have been shown, for clarity. The total amount of data elements in the data matrix  $\mathbf{X}$  equals to  $\mathcal{T} = M \times T$ .

## 2.2 Modelling of the data

Mathematical description of the data often requires the use of models. Models serve as simplifications of the underlying, often too complex, phenomenon that generates the data. By use of models, it also becomes possible to answer some specific questions regarding the data. One possibility to answer such specific questions is to generate specific hypotheses. It can then be tested by classical statistics whether these should be accepted or rejected.

If one is interested in more general questions regarding the data origin and its characteristics, one would not gain much by the classical hypothesis testing. To get a comprehensive picture of the data characteristics, one would need an enormous amount of different hypotheses since each test provides one with only little information. In this case, one may want to use more general models that can answer questions such as what kind of a process has generated the data and what kind of features build a good representation of it.

One way to build general descriptions of data is to use *latent variable* models. This is to say that the model includes auxiliary variables that cannot be measured directly, but which affect the data that is measured. A general mathematical description of the latent variable models is given by

$$\mathbf{X} = \mathbf{f}(\boldsymbol{\theta}), \quad (2.2)$$

where  $\boldsymbol{\theta}$  contains the model parameters and  $\mathbf{f}$  is some mapping relating the model parameters to the observations. This model is generative, because  $\mathbf{f}$  explicitly tells how the observations are generated.

We divide the possible generative latent variable models in two classes: 1) linear and 2) nonlinear models. Linear models have been widely studied because of their convenient mathematical properties. For this reason, they are the main focus in this thesis. Additionally, we apply the models in an application, where the linear models are justified from the knowledge of the underlying phenomenon (see Ch. 6 and Publication 6).

In linear models, an observation  $x_i(t)$  at a particular time  $t$  is given as a weighted mixture of, say  $N$  latent, i.e. hidden variables  $s_j(t)$ , plus some additive noise  $\nu_i(t)$ :

$$x_i(t) = \sum_{j=1}^N a_{ij}s_j(t) + \nu_i(t), \quad (2.3)$$

where  $a_{ij}$  is called the mixing coefficient and  $s_j(t)$  can be called source or factor. Note that the source  $s_j(t)$  may have different values depending on time  $t$  but the weighting  $a_{ij}$  is independent of it. Furthermore, the mixing is assumed instantaneous, i.e.  $x_i(t)$  at time  $t$ , only depends on the sources at that same time. For example, convolutive linear models do not fall under this restriction. Throughout this thesis, we only consider this stationary and instantaneous mixing since it is justified in our field of application (see Ch. 6 and Publication 1 for further details).

The noise  $\nu_i(t)$  accounts for all of the data in  $x_i(t)$  that is not fully modelled by the weighted linear sum  $\sum_{j=1}^N a_{ij}s_j(t)$ . We stress that although usually only measurement noise is considered as noise,  $\nu_i(t)$  accounts for all of the other inaccuracies as well. These can be caused by inaccuracies of the generative model, such as the nonlinearity or non-stationarity of the observation mapping. Sometimes the noise  $\nu_i(t)$  is also called sensor noise because it is additive in each sensor  $x_i(t)$  separately.

It is possible to collect the sources and the mixing coefficients in matrices in a similar manner that was done for the data  $\boldsymbol{\xi}$ . Let us define a row vector  $\mathbf{s}_j = [s_j(1) \cdots s_j(t) \cdots s_j(T)]$  that contains all of the values of the  $j$ th source. Then all of the observations in one measurement device, i.e.  $\mathbf{x}_i$  are given by

$$\mathbf{x}_i = \sum_{j=1}^N a_{ij}\mathbf{s}_j + \boldsymbol{\nu}_i. \quad (2.4)$$

Introduction of a column mixing vector  $\mathbf{a}_j = [a_{1j} \cdots a_{ij} \cdots a_{Mj}]^T$  then leads to a more compact form:

$$\mathbf{X} = \sum_{j=1}^M \mathbf{a}_j^T \mathbf{s}_j + \boldsymbol{\nu}. \quad (2.5)$$

Note that the index  $i$  has been dropped from the noise  $\boldsymbol{\nu}$  to indicate that it is a full matrix of size  $N \times T$ .

Furthermore, one can collect all of the sources in one matrix, similar to the data matrix:

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_j \\ \vdots \\ \mathbf{s}_N \end{bmatrix} = [\mathbf{s}(1) \quad \cdots \quad \mathbf{s}(t) \quad \cdots \quad \mathbf{s}(T)]. \quad (2.6)$$

Finally, a mixing matrix  $\mathbf{A} = [\mathbf{a}_1 \mathbf{a}_2 \cdots \mathbf{a}_M]^T$  lets one to describe the whole linear model using one matrix equation:

$$\mathbf{X} = \mathbf{AS} + \nu. \quad (2.7)$$

## 2.3 Probabilistic models

In modelling, the values of the unknown model parameters should be estimated from the data. However, usually the data does not unambiguously define the values of the model parameters. Consider for example a case where the sum of two positive integers below 10 is observed. Let 16 be observed. It is not possible to infer which of the pairs  $\{8,8\}$  or  $\{7,9\}$  has generated the observation. In other words, there is uncertainty of the model parameters. A convenient and mathematically grounded way to account for this uncertainty is to use probabilistic models and Bayesian probability theory (Cox, 1946, Gelman et al., 1995, Jordan, 1999). In this section, we first review some Bayesian concepts. After that, we proceed into possible ways to estimate model parameters under the Bayesian framework.

### 2.3.1 Events

Let  $A$  denote some event. Then the probability that the event  $A$  occurs is denoted by  $P(A)$ . Let  $B$  denote another event. Then the probability that both events occur is  $P(A \wedge B)$ . In the general case, this joint probability can be calculated by

$$P(A \wedge B) = P(A)P(B|A) = P(B)P(A|B), \quad (2.8)$$

where  $P(B|A)$  denotes the *conditional probability* that  $B$  occurs when it is known that  $A$  occurs and vice versa. If the event  $A$  is *independent* of event  $B$ , the joint probability becomes the product of the individual probabilities

$$P(A \wedge B) = P(A)P(B). \quad (2.9)$$

This means that the occurrence of  $A$  gives no knowledge of the probability of occurrence of  $B$ .

In the classical probability theory, it is not possible to speak about the probability of a single event, for example the probability of whether it rains tomorrow. This is because, in the classical theory, probability is defined through a frequency of occurrence in an infinite sample, but tomorrow only occurs once. In Bayesian theory, it is quite natural to speak about probabilities of single events, because the probability actually describes the *degree of belief* in the occurrence of the event. This also makes probability inherently *subjective*.

Often there are several events that are mutually exclusive. One example of such a case is a single coin toss that has the two alternative results: 'heads' or 'tales'. In this case, it is convenient to describe the relative occurrence of the events with a probability distribution. Let  $A$  represent the coin toss. Then the outcome is described by the probability distribution  $P(A = \text{'heads'}) = p$  and  $P(A = \text{'tales'}) = q$ . The sum of the probabilities of all possible mutually exclusive (independent) outcomes has to be one.

### 2.3.2 Probabilities of continuous variables

Real world data does not usually have a distinctive set of possible values but rather an infinite set. In this case, it is not possible to enumerate the probabilities of all possible outcomes. Continuous variables should rather be described by continuous distribution of probabilities. For instance, a Gaussian variable is described by the probability density function (pdf):

$$p_g(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right), \quad (2.10)$$

where the distribution is parameterised by two parameters:  $\mu$  and  $\sigma^2$ , the mean and variance of the variable  $a$ , respectively. From time to time, we use a simplified notation for the Gaussian distribution:  $N(a; \mu, \sigma^2)$ .

Other often used density function are the *Laplacian* and the *uniform* distributions, defined as:

$$p_l(a) = \frac{1}{2\beta} \exp\left(-\frac{|a-\mu|}{\beta}\right), \quad (2.11)$$

$$p_u(a) = \begin{cases} 1/\Delta, & \text{when } \mu - \Delta/2 \leq a \leq \mu + \Delta/2 \\ 0, & \text{otherwise,} \end{cases} \quad (2.12)$$

where  $\mu$  is again the mean of the distributions. Additionally, the Laplacian distribution has a parameter  $\beta$  that defines the variance of the distribution. In the uniform distribution,  $\Delta$  is the interval of possible values for  $a$  around the

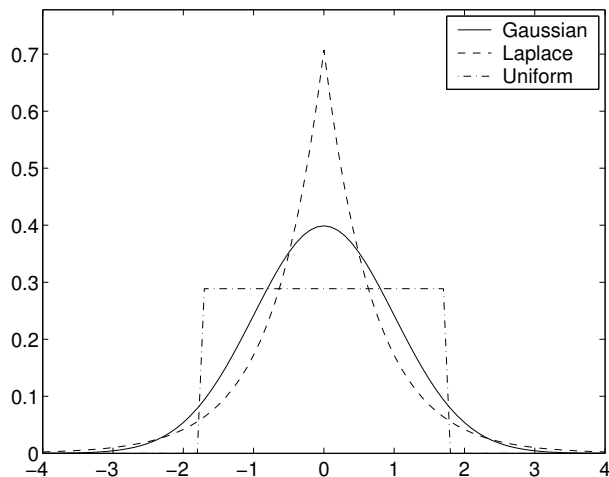


Figure 2.1: *Gaussian, Laplacian and uniform distributions. Their kurtoses are 0, 3 and -2, respectively.*

mean  $\mu$ . The three distributions are plotted in Fig. 2.1 with zero mean ( $\mu = 0$ ) and unit variance ( $\sigma^2 = 1$ ). The parameters  $\beta$  and  $\Delta$  are set to yield unit variance for the Laplacian and uniform distributions respectively.

Notice that the concept of probability of  $a$  is commonly used in two distinctive meanings: First, to mean the value of the density function  $p(a)$  at certain value  $a = a^*$ ; second, to mean the probability that  $a$  lies in certain interval  $a^- \leq a \leq a^+$ . To avoid confusion, when the value of the density function is meant, we use the words *density function* or *probability distribution*. On the other hand, if the latter is meant, we use the words *probability mass* or simply probability. This is because, a particular value of the probability density function does not have any natural interpretation. Actually the probability of any exact particular value of  $a$  occurring is zero. This becomes understandable in the Bayesian framework, because the interpretation of the probability mass is the subjective degree of belief in the variable to have its value on certain interval. Furthermore, it is impossible to observe any exact particular real number because of the uncertainty principle.

In many cases, we describe more than one variable using probability distributions and it is convenient to have notation for the total joint distribution of the variables. In this case we use the vector notation  $p(\mathbf{a}) = p(a_1, \dots, a_i, \dots, a_N)$  to denote the joint distribution. Often, we also need to denote the marginal distribution of all of the variables  $\mathbf{a}$ . This is done by using a subscript to indicate that the distribution is one dimensional:  $p_a(\mathbf{a})$ . When misunderstanding is feared, we

use a subscript for the joint distribution, too:  $p_{\mathbf{a}}(\mathbf{a})$ . In a similar manner, we may use the probability distributions for matrices. The subscript is also sometimes used to denote which distribution function is used. For example, if  $a$  and  $b$  have different distributions, we use  $p_a(\cdot)$  and  $p_b(\cdot)$  to denote their respective distributions. Usually however, we omit the subscript in this case.

We conclude the section on concepts needed for probabilistic modelling by reviewing several methods to describe distributions and dependencies between random variables. In Sec. 2.3.3, we return to the probabilistic modelling of the data.

### Cumulants: simple descriptors for distributions

Often there is a need to describe distributions using some simple scalar functions. One practical way is to use *cumulants* (c.f. Kendall and Stuart, 1958). They are defined as follows:

Consider a random variable  $a$  with  $E\{a\} = 0$ . The characteristic function  $\hat{h}(t)$  of  $a$  is defined as  $\hat{h}(t) = E\{e^{ita}\}$ . If the logarithm of the characteristic function is expanded into Taylor series:

$$\log \hat{h}(t) = \kappa_1(it) + \frac{\kappa_2(it)^2}{2} + \dots + \frac{\kappa_r(it)^r}{r!} + \dots, \quad (2.13)$$

the Taylor coefficients  $\kappa_r$  define the cumulants (of the distribution) of  $a$ . The first two cumulants are the *mean* and the *variance*

$$\kappa_1 = E\{a\} \quad (2.14)$$

$$\kappa_2 = E\{a^2\}. \quad (2.15)$$

They define perfectly a Gaussian distribution. The rest of cumulants are zero for Gaussian distributions, but may be non-zero for non-Gaussian distributions. The third cumulant, defined as

$$\kappa_3 = E\{a^3\}, \quad (2.16)$$

is called *skewness*. The skewness is zero for symmetrical distributions and non-zero for others. The fourth cumulant is perhaps the most interesting cumulant in our framework. It is called *kurtosis* (Kendall and Stuart, 1958, Nikias and Mendel, 1993), and is defined as

$$\text{kurt}(a) = \kappa_4 = E\{a^4\} - 3(E\{a^2\})^2. \quad (2.17)$$

Distributions that have kurtosis greater than zero are called super-Gaussian distributions, Laplacian distribution being one example. Super-Gaussian distributions are more peaked around the mean and have heavier tails than Gaussian



distributions. Negative-kurtosis-valued distributions are called sub-Gaussian distributions and they are flatter than Gaussian distributions. The flattest distribution is the uniform distribution that has  $\text{kurt}(a) = -2$ . Note that the kurtosis is bounded to -2 in sub-Gaussian distributions, but unbounded for super-Gaussian distributions. This makes kurtosis rather tuned to the tails of the distribution.

The higher than fourth order cumulants are not usually named mainly because they become too complex to analyse.

### Negentropy: a general measure of structure of the distribution

Differential entropy  $H$  of a random variable is a measure of disorder and is dependent on the variance of the variable:

$$H(a) = - \int p_a(x) \log p_a(x) dx, \quad (2.18)$$

where the variable  $a$  has a density function  $p(x)$ . For variables of fixed variance, the Gaussian distribution gives the highest entropy and is thus most unstructured distribution. A measure of structure, independent of variance, can be derived from the differential entropy by calculating the difference between the differential entropy of a variable  $a$  and a Gaussian variable with the same variance  $\nu$ :

$$N(a) = H(\nu) - H(a). \quad (2.19)$$

This is called *negentropy* and it is zero for the Gaussian distribution and non-negative for all distributions.

The calculation of the differential entropy, needed for the negentropy, assumes the distribution of the variable to be known. This is seldom the case in practice and the estimation of the distribution is often difficult and computationally demanding. Thus the negentropy is usually approximated with some simple measures easily calculatable from a signal  $\mathbf{s}$ . For example approximations based on the cumulants have been suggested (Kendall and Stuart, 1958) but these often provide a poor approximation for the negentropy (Hyvärinen, 1998b).

### Mutual information

Often two variables and their dependencies are compared. A good measure for the comparison is the *mutual information*. It measures the information that is common between the variables:

$$I(a_1, a_2, \dots, a_m) = \sum_{i=1}^m H(a_i) - H(\mathbf{a}), \quad (2.20)$$

where  $\mathbf{a} = [a_1, a_2, \dots, a_m]^T$  and  $H(\mathbf{a})$  is the total differential entropy of the variables  $a_1, a_2, \dots, a_m$ .

### Kullback-Leibler Divergence

A good measure for the dissimilarity between two distributions is the *Kullback-Leibler divergence* (Luenberger, 1969):

$$D_{\text{KL}}(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \quad (2.21)$$

This measure is always non-negative and it is zero *if and only if* the distributions  $p$  and  $q$  are equal. For this reason, it is sometimes called Kullback-Leibler distance. However, the measure is not symmetric, i.e.  $D_{\text{KL}}(p, q) \neq D_{\text{KL}}(q, p)$  for which reason it does not constitute a metric.

KL divergence can be used to measure the dependence between two variables  $a$  and  $b$  by calculating the dissimilarity  $D_{\text{KL}}(p(a, b), p(a)p(b))$ .

### 2.3.3 Likelihood of the data

When a data matrix  $\mathbf{X}$  has already been observed, it makes little sense to speak about its probability. The probability should be one, because the data has been observed as it is<sup>1</sup>. In that case, the word *likelihood* is used. Usually, the likelihood of the data is given under some model  $\mathcal{H}$  having parameters  $\boldsymbol{\theta}$ , i.e.  $p(\mathbf{X}|\boldsymbol{\theta}, \mathcal{H})$ . This likelihood describes how probable it would have been to observe this particular data  $\mathbf{X}$  if it would have been generated by model  $\mathcal{H}$  with parameter values  $\boldsymbol{\theta}$ .

In *maximum likelihood* (ML) estimation, such a set of model parameters  $\boldsymbol{\theta}$  of the particular model  $\mathcal{H}$  are sought that maximise the likelihood  $p(\mathbf{X}|\boldsymbol{\theta}, \mathcal{H})$  of the data.

### 2.3.4 Bayesian modelling

When a data  $\mathbf{X}$  has been observed, such a model  $\mathcal{H}$  and such a set of parameters  $\boldsymbol{\theta}$  should be selected that fit the data. In ML estimation, the likelihood of the data given the model is maximised. But actually, it would be more natural to maximise the probability of the model given the data, i.e.  $p(\mathcal{H}, \boldsymbol{\theta}|\mathbf{X})$ , not the other way around.

The equality on the right side of the Eq. (2.8) can be used to tell the conditional probability of  $P(B|A)$  using the prior probabilities  $P(A)$  and  $P(B)$  and the reversed conditional probability  $P(A|B)$  using the Bayes' theorem (Cox, 1946):

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}. \quad (2.22)$$

---

<sup>1</sup>To be precise, there is an uncertainty principle that states that it is impossible to observe any real values exactly. This would generate uncertainty in the data, too.

This can be directly used for the estimation of the model parameters. Let us denote by  $p(\mathbf{X}|\mathcal{H})$  some prior probability distribution for the data  $\mathbf{X}$  given the model  $\mathcal{H}$  and by  $p(\boldsymbol{\theta}, \mathcal{H})$  the prior distribution of the model and its parameters. Furthermore,  $p(\mathbf{X}|\boldsymbol{\theta}, \mathcal{H})$  is the likelihood of the data given the model  $\mathcal{H}$  having parameters  $\boldsymbol{\theta}$ . Then the probability of the model  $\mathcal{H}$  having the parameters  $\boldsymbol{\theta}$ , given the data is

$$p(\boldsymbol{\theta}, \mathcal{H}|\mathbf{X}) = \frac{p(\boldsymbol{\theta}, \mathcal{H})p(\mathbf{X}|\boldsymbol{\theta}, \mathcal{H})}{p(\mathbf{X}|\mathcal{H})}. \quad (2.23)$$

This is called the posterior probability distribution of the model  $\mathcal{H}$  having parameters  $\boldsymbol{\theta}$ . The model  $\mathcal{H}$  is often dropped from the notation since it is usually possible to infer it from the set of parameters  $\boldsymbol{\theta}$ .

In *maximum a posteriori* (MAP) estimation, such a set of parameters  $\boldsymbol{\theta}$  is chosen that maximise the density of the posterior probability distribution  $p(\boldsymbol{\theta}|\mathbf{X})$ . However, note that the density itself does not have an interpretation, only probability mass does. Hence, MAP estimation behaves badly when the posterior has narrow but high peaks, having little probability mass. This leads to overfitting of the model and overfitted models are not useful for generalisation, i.e. do not fit well to future data. Overfitting is discussed in more detail in Sec. 5, Publication 2 and Publication 3.

Note that the ML estimation discussed earlier is equivalent to the MAP estimation when the prior probabilities for the model parameters  $p(\boldsymbol{\theta})$  are assumed uniformly distributed. This is because  $p(\mathbf{X})$  is always constant and the Bayes' theorem results in  $p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})$ .

### 2.3.5 Full distribution approaches to Bayesian modelling

In correct Bayesian modelling, one should not only use a single set of parameters  $\boldsymbol{\theta}$ . Instead one should use the total posterior distribution for any inferences. Thus, only one model should not be selected as in MAP but instead all of the models should be taken into account. Moreover, their contribution should be weighted according to their respective densities in the posterior distribution.

Almost in all of the practical applications, the calculation of the full posterior distribution is intractable. Thus some methods are needed to approximate it. Perhaps the most popular approximation methods base inferences on some finite sample of the true posterior distribution. These are usually called Markov-Chain-Monte-Carlo (MCMC) methods. These sampling methods fall out of the scope of this thesis, but we point the reader to a good textbook on them by Gelman et al. (1995). The Bayesian methods discussed here approximate the posterior distribution with some analytically tractable distribution.

### Ensemble learning

Ensemble learning (EL, Wallace, 1990, Hinton and van Camp, 1993, Lappalainen and Miskin, 2000) is a recently developed variational method for fitting a parametric approximation to the exact posterior density function  $p(\boldsymbol{\theta}, \mathcal{H}|\mathbf{X})$ . The true posterior distribution is approximated by a density  $q(\boldsymbol{\theta})$  having a simple form. The misfit of the approximation is measured by the Kullback-Leibler divergence (2.21) between the approximation and the true posterior distribution:

$$D(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}, \mathcal{H}|\mathbf{X})) = E_{q(\boldsymbol{\theta})} \left[ \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{H}|\mathbf{X})} \right]. \quad (2.24)$$

The posterior distribution can be written as  $p(\boldsymbol{\theta}, \mathcal{H}|\mathbf{X}) = p(\boldsymbol{\theta}, \mathcal{H}, \mathbf{X})/p(\mathbf{X}|\mathcal{H})$ . The normalising term  $p(\mathbf{X}|\mathcal{H})$ , called evidence for the model  $\mathcal{H}$ , cannot usually be evaluated because it would involve an intractable integration over all of the model parameters. However, this term is constant, when the parameters are estimated for a fixed model. Hence, in EL, the evidence term is neglected and the actual cost function is

$$C_{KL} = D(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}, \mathcal{H}|\mathbf{X})) - \log p(\mathbf{X}|\mathcal{H}) \geq -\log p(\mathbf{X}|\mathcal{H}). \quad (2.25)$$

The last inequality follows from the fact that Kullback-Leibler divergence is always non-negative.

This cost function is enough to determine the model parameters. However, in addition to parameter estimation, it is important to be able to compare different models, e.g. with different number of sources. Unfortunately, the most natural measure for the comparison is the neglected evidence term. But fortunately, minimisation of the cost function (2.25) for a given model  $\mathcal{H}_i$ , maximises a lower bound of the evidence  $p(\mathbf{X}|\mathcal{H}_i)$  for that model. In particular, the lower bound is given by

$$p(\mathbf{X}|\mathcal{H}_i) \geq e^{-C_{KL}}. \quad (2.26)$$

Hence Eq. (2.25) can also be used for comparing and selecting different models. It is to be kept in mind, though, that because of the simple form of the posterior approximation, ensemble learning favours simpler models.

## 2.4 Gradient based optimisation methods

In the previous section, we reviewed three methods that give grounds on which the parameter estimation could be performed: ML and MAP estimation, and ensemble learning. Each of them define an objective function to be optimised. However, it is usually not possible to calculate the optima analytically, but iterative methods are needed. A good handbook for iterative optimisation methods

has been written by Luenberger (1969). All of the methods reviewed below can be found therein.

Let  $g(\boldsymbol{\theta})$  be the function that should be optimised (minimised or maximised). Generally, in the optimum, it holds that

$$\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta}) = \mathbf{0}, \quad (2.27)$$

where  $\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta})$  is a column vector having  $\partial g(\theta_i)/\partial \theta_i$  as its  $i$ th element and  $\mathbf{0}$  is a column vector of the same size having zeroes as its elements.

In case the parameters  $\boldsymbol{\theta}$  have some constraints, the condition in the optima includes the Lagrange multipliers:

$$\nabla_{\boldsymbol{\theta}}[g(\boldsymbol{\theta}) - \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\theta})] = 0, \quad (2.28)$$

where the column vectors  $\boldsymbol{\lambda}$  and  $\mathbf{h}$  denote the Lagrange multipliers and the corresponding constraints under which the optimisation is performed, respectively.

Note that the gradient of the function  $g$  points in the direction where  $g$  grows maximally, in the *Euclidean coordinate system*. However, Amari (1998) has pointed out that the parameter space often has a Riemannian structure. In the scope of this thesis, it is not possible to discuss this issue any further. We abide for noting that in this case, the so-called *natural gradient* serves as a more suitable choice for the gradient-based algorithms.

In the following, we review three optimisation methods based on the gradient.

### 2.4.1 Gradient descent and ascent

The gradient  $\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta})$  points in the direction where the objective function grows maximally. Hence, an infinitesimal step in the direction of the gradient always increases the objective function. By taking successive step, always in the direction of the gradient, the following update rule is obtained:

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta} + \gamma \nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta}) \quad (2.29)$$

where  $\gamma$  makes the step size infinitesimal. Under certain conditions (Luenberger, 1969), this algorithm is guaranteed to converge to the nearest local maximum of the objective function  $g(\boldsymbol{\theta})$ . Convergence is of course awfully slow, but it can be greatly improved by using considerably bigger step size  $\gamma$ . Then it is usual to call it the *learning rate*. When non-infinitesimal  $\gamma$  is used, it must usually be taken adaptive. Otherwise the convergence of the gradient-ascent algorithm cannot be guaranteed.

In case one needs to minimise  $g(\boldsymbol{\theta})$ , one should always move in the opposite direction of the gradient. This results in a gradient descent algorithm. This may be implemented by changing the addition to subtraction in Eq. (2.29).

### 2.4.2 Fixed point algorithms

In many cases, the estimation of the parameters  $\boldsymbol{\theta}$  can be performed in a subset of the parameters. Let the subset of the parameters under which the optimisation is performed to be denoted by  $\mathbf{w}$ . Furthermore, it is common to restrict the optimisation on a unit sphere, i.e.  $\|\mathbf{w}\| = \mathbf{w}^T \mathbf{w} = 1$ .

Consider the following general iterative algorithm to optimise some objective function  $g(\mathbf{w})$  on the unit sphere:

$$\mathbf{w}^+ = \mathbf{f}(\mathbf{w}) \quad (2.30)$$

$$\mathbf{w}_{\text{new}} = \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}, \quad (2.31)$$

where the explicit normalisation has been added to ensure the satisfaction of unit norm. In other words, the current estimate  $\mathbf{w}$  is updated using function  $\mathbf{f}$  and normalised. This iteration has stable points that satisfy the condition:

$$\mathbf{w}^+ = \mathbf{f}(\mathbf{w}) \propto \mathbf{w}. \quad (2.32)$$

These points are stable because the normalisation step then renders  $\mathbf{w}_{\text{new}} = \mathbf{w}$ . Such points are called the *fixed points* of the algorithm.

Now consider the Lagrange equation (2.28). Addition of  $\mathbf{w}$  on both sides of the equation results in

$$\mathbf{w} = \nabla_{\mathbf{w}}[g(\mathbf{w}) - \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{w})] + \mathbf{w}, \quad (2.33)$$

which can be directly used as the fixed point iteration step (2.30). The fixed points of this algorithm are clearly exactly the points that satisfy Eq. (2.28). But note that there are other algorithms that result in the same fixed points. For instance, the right hand side of Eq. (2.33) can be multiplied by any constant because of the explicit normalisation (2.31). Furthermore, any multiple of  $\mathbf{w}$  can be added without changing the fixed points for similar reasons. These modifications result in the general fixed point algorithms optimising  $g(\boldsymbol{\theta})$  having constraints  $\mathbf{h}(\mathbf{w})$ :

$$\mathbf{w}^+ = \alpha(\mathbf{w}) \nabla_{\mathbf{w}}[g(\mathbf{w}) - \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{w})] + \beta(\mathbf{w}) \mathbf{w}. \quad (2.34)$$

$$\mathbf{w}_{\text{new}} = \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}. \quad (2.35)$$

where  $\alpha(\mathbf{w})$  and  $\beta(\mathbf{w})$  are scalar-valued functions.  $\beta(\mathbf{w})$  can be used for speeding up the convergence of the fixed-point algorithm. Such speedup for the estimation of linear models using particular optimisation algorithms are further discussed in Secs. 3.1.3 and 4.5.

## Chapter 3

# Separation of linearly mixed sources

*Pluralitas non est ponenda sine necessitas.*

–Ockham (14th century)

Let us consider modelling of data  $\mathbf{X}$  using the linear model given in Eq. (2.7):

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \boldsymbol{\nu}. \quad (3.1)$$

To recall the notation, we remind the reader that  $\mathbf{X}$  consists of  $M$  observations  $T$  samples each, resulting in an  $M \times T$ -matrix and  $\mathbf{S}$  of  $N$  sources also  $T$  samples each, similarly. These are related by the matrix  $\mathbf{A}$  which is the  $M \times N$  mixing matrix. Finally, there is some additive noise  $\boldsymbol{\nu}$  that takes care of all of the modelling inaccuracies.

The problem addressed in this thesis is to recover the unknown parts  $\mathbf{A}$  and  $\mathbf{S}$  of Eq. (3.1). This problem is called *linear source separation*. Since we mainly consider linear mixtures in this thesis, we frequently drop the word linear. Recall also that we concentrate on instantaneous and stationary mixing.

Solving of the source separation problem is not possible if there is no information on some of the variables  $\mathbf{A}$  and  $\mathbf{S}$ , in addition to the known data  $\mathbf{X}$ . If the mixing is assumed to be known and the noise to be negligible, the sources can be estimated by finding a matrix  $\mathbf{B}$ , for which  $\mathbf{B}\mathbf{A} = \mathbf{I}$ . Then  $\mathbf{B}\mathbf{X} = \mathbf{B}\mathbf{A}\mathbf{S} = \mathbf{S}$ . If  $\mathbf{A}$  is a square matrix, i.e. there are as many observations  $\mathbf{X}$ , as there are sources  $\mathbf{S}$ , and  $\mathbf{A}$  has full rank, the solution to this is simply  $\mathbf{B} = \mathbf{A}^{-1}$ . The above full-rank assumption is the necessary and sufficient condition for the existence of the inverse matrix  $\mathbf{A}^{-1}$ . If instead, there are more observations than sources, there exist several matrices  $\mathbf{B}$  that satisfy  $\mathbf{B}\mathbf{A} = \mathbf{I}$ . In that case, as long as we are interested in some features of  $\mathbf{S}$  only, disregarding  $\mathbf{A}$ , any such  $\mathbf{B}$  can be

chosen. If the rank of  $\mathbf{A}$  is less than the amount of sources, the problem has no unique solution, if further assumptions are not made. This is the case when there are less observations than sources or when there are some redundancies in the mixing.

If, on the other hand, no non-trivial prior information of the mixing  $\mathbf{A}$  is assumed, the problem of estimating the unknowns,  $\mathbf{A}$  and  $\mathbf{S}$ , is referred to as *blind source separation* (BSS). Notice that something of the sources  $\mathbf{S}$  still has to be "seen", to make the estimation possible. Consider for instance the multiplication of the source estimates with some constant  $\alpha$ . Then, the relation  $\mathbf{X} = \mathbf{A}\mathbf{S}$  can be preserved by multiplying the mixing with  $1/\alpha$ . This presents a simple case of indeterminacy, though usually harmless, that cannot be solved without additional assumptions. In this thesis, we restrict the sources to have zero mean and unit variance, without loss of generality (Hyvärinen et al., 2001b). If the observations  $\mathbf{x}_i$  are not zero mean, the sample means  $\sum_t x_i(t)$  are removed. Thus, for the rest of this thesis, we assume  $\mathbf{X}$  zero mean, as well.

There exists one common formulation for the source separation problem in addition to the formulation (3.1):  $\mathbf{s}$  is interpreted as a vector-valued random variable. Then the model is expressed as  $\mathbf{x} = \mathbf{A}\mathbf{s}$ , where  $\mathbf{s}$  contains the sources (now a column vector of length  $N$ ) and  $\mathbf{x}$  the data (a column vector of length  $M$ ). The source separation model (3.1) used in this thesis is achieved from this random-vector model by collecting the instances of the observation vector  $\mathbf{x}$  in an observation matrix  $\mathbf{X}$  and the instances of the source vector  $\mathbf{s}$  in a source matrix  $\mathbf{S}$ . The choice between the vector- and the matrix-form models is mainly notational. All of the algorithms can be easily developed in both notations. Usually, algorithms using the vector notation can be molded from the algorithms having the matrix form by changing the sums and normalisations with  $T$  into expectations and vice versa. We mainly use the matrix notation to be consistent.

So far, the noise  $\boldsymbol{\nu}$  has been assumed to be non-existent or negligible. Then it is evident that the solution to the source separation problem is found in a form  $\hat{\mathbf{S}} = \mathbf{B}\mathbf{X}$ . In case the noise is not negligible, the demixing matrix  $\mathbf{B}$  can often be identified, but the estimated sources still contain noise (c.f., Hyvärinen et al., 2001b). In particular, the demixing matrix can be identified when the noise covariance  $\boldsymbol{\Sigma}$  has the specific form:

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T\sigma^2. \quad (3.2)$$

In this case, it is possible to use a transformation  $\tilde{\boldsymbol{\nu}} = \mathbf{A}^{-1}\boldsymbol{\nu}$  and the linear model results in  $\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{A}\tilde{\boldsymbol{\nu}} = \mathbf{A}(\mathbf{S} + \tilde{\boldsymbol{\nu}})$ . Or equivalently, the demixing matrix can be solved from  $\mathbf{S} + \tilde{\boldsymbol{\nu}} = \mathbf{B}(\mathbf{X} + \tilde{\boldsymbol{\nu}})$ . Note that in order to solve the sources, additional methods are needed to account for the noise term  $\tilde{\boldsymbol{\nu}}$ . To conclude, even in the noisy case, it is sometimes possible to identify the demixing matrix



and the noisy sources by

$$\hat{\mathbf{S}} = \mathbf{B}\mathbf{X}. \quad (3.3)$$

In the following, we discuss source separation methods that concentrate on estimating the separating matrix  $\mathbf{B}$ . Only in Secs. 3.2.4, 3.2.5 and 3.4, we discuss methods that try to estimate the noise in the sources or where the noise covariance does not follow the model (3.2) and the identification of the separating matrix is not directly possible.

This lays the foundation and the notation for the problem of source separation used in this thesis. In the following sections, we review several ways to restrict the unknown variables to make the estimation of the interesting features of the sources possible.

### 3.1 Principal component analysis

Consider the two two-dimensional data sets shown in Fig. 3.1. Substantial correlations exist between the components  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . This hints that there is a more informative representation for the data via latent variables. A good attempt to recover the possible original sources  $\mathbf{s}_1$  and  $\mathbf{s}_2$  is to find a linear mapping  $\mathbf{B}$  that makes the produced sources uncorrelated. This can be achieved by calculating the eigenvalue decomposition of the covariance matrix  $\mathbf{X}\mathbf{X}^T/T = \mathbf{E}^T\mathbf{\Lambda}\mathbf{E}$ , and making the transformation:

$$\mathbf{Y} = \mathbf{V}\mathbf{X} = \mathbf{\Lambda}^{-1/2}\mathbf{E}\mathbf{X}, \quad (3.4)$$

where  $\mathbf{\Lambda} = \text{diag}(\lambda_1 \lambda_2 \cdots \lambda_L)$  has the  $L$  largest eigenvalues in decreasing order and  $\mathbf{E} = [\mathbf{e}_1 \cdots \mathbf{e}_l \cdots \mathbf{e}_L]^T$  consists of the eigenvectors corresponding to the eigenvalues<sup>1</sup>. The transformation  $\mathbf{E}$  makes the components in  $\mathbf{Y}$  decorrelated to each other and the diagonal matrix  $\mathbf{\Lambda}^{-1/2}$  renders all of the variances to unity. It is common to call the transformed components  $\mathbf{Y}$  whitened or sphered. The latter name is used in this thesis.

The eigenvectors  $\mathbf{E}$  can be calculated for instance by the classical *power method* (Wilkinson, 1965). Other possibilities include neural methods (c.f. Amari, 1977, Oja, 1982, Oja and Karhunen, 1985). Transformation  $\mathbf{E}\mathbf{X}$  without the normalisation  $\mathbf{\Lambda}^{-1/2}$  has many names, the most popular perhaps being *principal component analysis* (PCA). Other names include Hotelling transformation (Hotelling, 1933) or Karhunen-Loève transformation (Loève, 1955).

The principal eigenvalue corresponds to the variance of the data  $\mathbf{X}$  projected to the direction of the principal eigenvector. The nut-shaped curves in Fig. 3.1

---

<sup>1</sup>Note that an unconventional direction has been selected for the eigenvectors  $\mathbf{e}_l$  in the eigenmatrix  $\mathbf{E}$  to produce an  $L \times M$  matrix. This has been done to keep the notation coherent with that of the matrices  $\mathbf{A}$ , and  $\mathbf{B}$ .

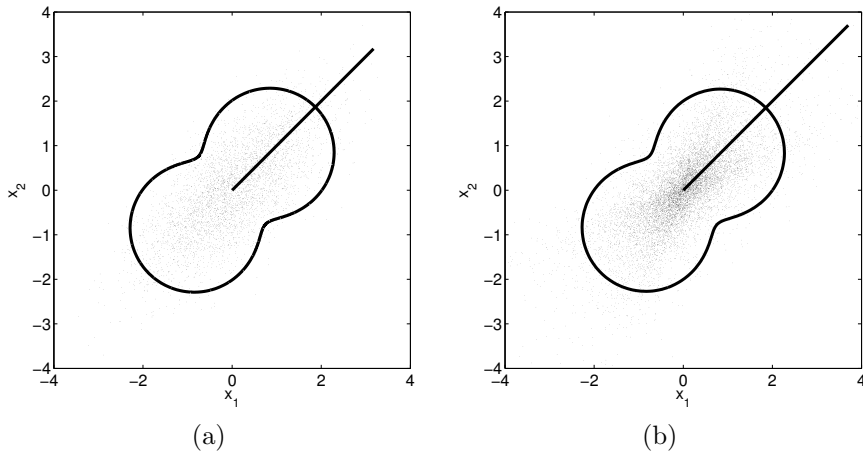


Figure 3.1: a) Some 2-dimensional data b) Other 2-dimensional data. The nut-shaped curves show the standard deviation of the projections of the data in each direction. The straight lines show the direction of the maximal variance.

are in proportion to the standard deviation of the data when projected in that particular direction<sup>2</sup>. The straight lines show the principal eigenvectors. The transformed or sphered data are shown in Fig. 3.2. The marginal distributions of the sphered components  $\mathbf{y}_1$  and  $\mathbf{y}_2$  in Fig. 3.2a are Gaussian and hence the components are independent as well.

In practical use of PCA, it is common to reduce the dimensionality. Then the principal components enable the optimal reconstruction of the original data in mean-squared-error sense.

PCA is not the only way to achieve a transformation matrix  $\mathbf{V}$  that makes the components  $\mathbf{Y}$  mutually uncorrelated, and independent in case of Gaussian sources. Actually, any further orthogonal rotation  $\mathbf{U}\mathbf{Y}$  gives uncorrelated components:  $\mathbf{U}\mathbf{Y}(\mathbf{U}\mathbf{Y})^T/T = \mathbf{U}\mathbf{Y}\mathbf{Y}^T\mathbf{U}^T/T = \mathbf{I}$ . This is illustrated by the fact that in Fig. 3.2 the projection of the data in any direction gives unit variance.

The rotational indeterminacy means that it is not possible to recover the originally mixed Gaussian i.i.d sources in the first data set. Thus, in strict sense, PCA does not solve the source separation problem. It can be argued, though, that from ESS point of view, no further insight to the structures of the data  $\mathbf{X}$  can be achieved. It is usual to use a common name of *factor analysis* (FA,

<sup>2</sup>To be exact, the double of the standard deviation is shown. The scaling is done for illustrative purposes only, and similar scaling is used in all subsequent examples following the same logic.

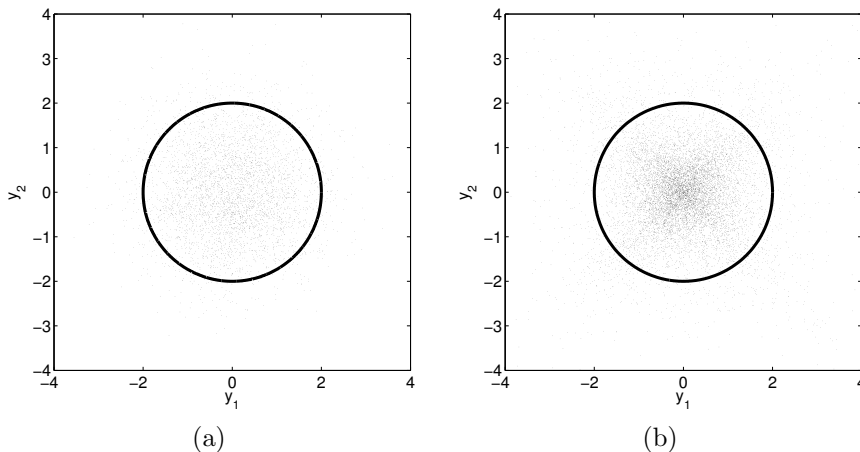


Figure 3.2: *The sphered components of Fig 3.1. The unity of variances is highlighted by the solid circles.*

see Spearman, 1904, Holzinger and Harman, 1951, Cattell, 1952, Horst, 1965, for classical references), for all of the decorrelating projections.

In the case of the other data set shown in Fig. 3.1b, sphering does not remove all of the structure, but the star shape is persistently visible in Fig. 3.2b. This is an indication that though PCA has removed the simple linear correlations, it has not been able to recover the original sources. In contrast to the previous case, the marginal distributions are not Gaussian. Furthermore, some higher-order correlations exist and the sphered components are thus not independent. Hence, PCA has not been able to solve the source separation problem from the ESS point of view either.

Despite the strong rotational indeterminacy described above, PCA and FA methods are frequently applied in biomedical systems (Jokeit and Makeig, 1994, Marder et al., 1997, Levelt, 2001, Tsiptsios et al., 2003, Ito et al., 2003). Even heavier impact they have on psychology and social sciences (Spearman, 1904, Rubin and Thayer, 1982, Gorsuch, 1983, Everitt, 1984, Basilevsky, 1994).

To conclude, FA seems like a useful step towards source separation but it is not able to recover the original sources because of the rotational indeterminacy. In the following, we discuss what additional structure might make the separation possible. We consider sources having non-Gaussian marginal distributions in Sec. 3.2. In Sec. 3.3, we concentrate on sources having some time structure, i.e. individual samples  $s(t_1)$  and  $s(t_2)$  are not independent for at least some pair  $(t_1, t_2)$ .

But first, we review the classical power method approach to calculate the eigenvectors and the eigenvalues of the covariance matrix in Sec. 3.1.1 and some common extensions in subsequent sections. These extensions and the power method will be used to derive fast and robust source separation algorithms in Sec. 4.

### 3.1.1 Power method

The principal eigenvector  $\mathbf{e}_1$  should correspond to the direction of maximum variance of all of the projections  $\mathbf{y} = \mathbf{e}^T \mathbf{X}$ . In other words,

$$g_{\text{lin}}(\mathbf{e}) = \|\mathbf{e}^T \mathbf{X}\|^2 = \mathbf{e}^T \mathbf{X} \mathbf{X}^T \mathbf{e} / T \quad (3.5)$$

should be maximised. Since the eigenvectors have unit length, there is an additional constraint  $h(\mathbf{e}) = \mathbf{e}^T \mathbf{e} - 1 = 0$ . According to the Lagrange equation (2.28), the optima of  $g_{\text{lin}}(\mathbf{e})$ , subject to the constraint  $h(\mathbf{e})$ , satisfy:

$$\nabla_{\mathbf{e}}[g(\mathbf{e}) - \lambda^T h(\mathbf{e})] = 2\mathbf{X} \mathbf{X}^T \mathbf{e} / T - 2\lambda \mathbf{e} = \mathbf{0}. \quad (3.6)$$

Derivation of a fixed point algorithm (see Sec. 2.4.2) from this is simple and results in:

$$\mathbf{e}^+ = \mathbf{X} \mathbf{X}^T \mathbf{e} \quad (3.7)$$

$$\mathbf{e}_{\text{new}} = \frac{\mathbf{e}^+}{\|\mathbf{e}^+\|}, \quad (3.8)$$

where  $\mathbf{e}_{\text{new}}$ , means the new estimate of the principal direction and the normalisation has been added to ensure stable unit length of the eigenvector estimate  $\mathbf{e}$ . The unnormalised principal component is given by  $\mathbf{e}^T \mathbf{X}$ . The algorithm above is the classical *power method* (Wilkinson, 1965). For an intuitive explanation of the power method, consider two consecutive steps of the algorithm:

$$\mathbf{e}^* = \mathbf{X} \mathbf{X}^T \mathbf{e}_{\text{new}} = \mathbf{X} \mathbf{X}^T \frac{\mathbf{X} \mathbf{X}^T \mathbf{e}}{\|\mathbf{X} \mathbf{X}^T \mathbf{e}\|} = \frac{\mathbf{C}^2 \mathbf{e}}{\|\mathbf{X} \mathbf{X}^T \mathbf{e}\|}, \quad (3.9)$$

where  $\mathbf{C} = \mathbf{X} \mathbf{X}^T$  is the unnormalised covariance matrix. Thus in each step,  $\mathbf{C}$  is multiplied by itself. This promotes the principal eigenvector to the lesser eigenvectors.

The power method has some limitations. First, it calculates only the principal direction corresponding to the maximal eigenvalue. Second, its matrix power nature makes the algorithm converge slowly when the principal directions have comparable eigenvalues. There are two common extensions that try to overcome these problems: *deflation* can be used to calculate several principal directions and *spectral shift* can be used to speedup the algorithm. These are discussed next.

### 3.1.2 Extracting several components

In case there is a need to calculate a subset or all of the principal directions, the power method can be applied several times. This is referred to as the *deflation* procedure. However, the subsequent estimates  $\mathbf{e}_i$ ,  $i > j$  need to be restricted so that they do not converge to already estimated principal directions  $\mathbf{e}_j$ ,  $j < i$ . This can be achieved by replacing the normalisation (3.8) with an orthonormalisation procedure:

$$\mathbf{e}_{\text{new}} = \text{orth}(\mathbf{e}^+). \quad (3.10)$$

One possibility for orthonormalisation  $\text{orth}(\mathbf{e}^+)$  was introduced by Hotelling (1933)<sup>3</sup>:

$$\mathbf{e}_{\text{orth}} = \mathbf{e}^+ - \mathbf{E}^T \mathbf{E} \mathbf{e}^+ \quad (3.11)$$

$$\mathbf{e}_{\text{new}} = \frac{\mathbf{e}_{\text{orth}}}{\|\mathbf{e}_{\text{orth}}\|}, \quad (3.12)$$

where  $\mathbf{E}$  contains the already estimated principal directions.

It is also possible to estimate several uncorrelated basis vectors at the same time in a *symmetric* manner. Then the symmetric power method becomes:

$$\mathbf{E}^+ = \mathbf{X} \mathbf{X}^T \mathbf{E} \quad (3.13)$$

$$\mathbf{E}_{\text{new}} = \text{orth}(\mathbf{E}^+). \quad (3.14)$$

The orthonormalisation can be implemented by  $\mathbf{E}_{\text{new}} = (\mathbf{E}^+ \mathbf{E}^{+T})^{-1/2} \mathbf{E}$  (Luenberger, 1969). In this symmetric procedure, it is not guaranteed that the basis vectors correspond to the principal directions but the variance of the covariance matrix may be divided more evenly. However, it does hold that the first  $L$  symmetric basis vectors span the most varying subspace of the whole data  $\mathbf{X}$ . Thus, symmetric power method can be used to perform FA, where the factors come from the most varying subspace.

### 3.1.3 Spectral shift

In the classical power method, the convergence speed depends on the ratio of the largest eigenvalues,  $|\lambda_1/\lambda_2|$ , where  $|\lambda_1| > |\lambda_2|$  (Wilkinson, 1965). If this ratio is close to unity, the matrix multiplication (3.7) does not promote the largest eigenvalue effectively when compared to the second largest eigenvalue.

---

<sup>3</sup>However, Wilkinson (1965) notes that this method may suffer from numerical instability and suggests more stable methods. The method by Hotelling (1933) is used in this thesis for its simplicity, though.

The convergence speed in such cases can be increased by so called spectral shift<sup>4</sup> which modifies the eigenvalues without changing the fixed points. At the fixed point of the classical power method,

$$\lambda \mathbf{e} = \mathbf{X}\mathbf{X}^T \mathbf{e}. \quad (3.15)$$

Then it also holds that  $(\lambda + \lambda_\beta)\mathbf{e} = (\mathbf{X}\mathbf{X}^T + \lambda_\beta)\mathbf{e}$  for any  $\lambda_\beta$ . The additional term simply adds  $\lambda_\beta$  to all eigenvalues. The spectral shift modifies the ratio of two largest eigenvalues to  $|(\lambda_1 + \lambda_\beta)/(\lambda_2 + \lambda_\beta)| > |\lambda_1/\lambda_2|$ , provided that  $\lambda_\beta$  is negative but not much smaller than  $-\lambda_2$ . This can greatly increase the convergence speed of the classical power method.

On the other hand, for very negative  $\lambda_\beta$ , some eigenvalues will become negative. In fact, if  $\lambda_\beta$  is small enough, the absolute value of the originally smallest eigenvalue will exceed that of the originally largest eigenvalue. With this negative spectral shift, the modified power method converges to the minor component (Oja, 1992, Xu et al., 1992).

### 3.1.4 Nonlinear principal component analysis

In the following, we review an algorithm that can realise PCA, but which also has an important application to BSS. Consider a gradient descent algorithm that converges to the vector  $\mathbf{e}$  corresponding to the maximum of an objective function  $g(\cdot)$  (Oja et al., 1991):

$$\mathbf{e}^+ = \mathbf{e} + \gamma(t)\nabla_{\mathbf{e}}g(\mathbf{e}^T \mathbf{x}(t)) \quad (3.16)$$

$$\mathbf{e}_{\text{new}} = \frac{\mathbf{e}^+}{\|\mathbf{e}^+\|}, \quad (3.17)$$

where  $\gamma(t)$  is a learning rate, changing in time to ensure convergence and the gradient is taken with respect to each element of  $\mathbf{e}$  resulting in a column vector.  $\mathbf{x}(t)$  contains the values of the data at a particular time instance  $t$  in a column vector. The gradient is usually called the nonlinearity and denoted by  $\mathbf{f}_{\mathbf{e}}(\cdot) = \nabla_{\mathbf{e}}g(\cdot)$ . It should give negative values for negative arguments and positive values for positive arguments, for stability reasons. The main aim of Oja et al. (1991) is to present a local neuron learning rule to approximate this algorithm. However, the nonlocal algorithm above (3.16) and (3.17) serves our purposes better and we do not present the details of the local algorithm.

When function  $\mathbf{f}_{\mathbf{e}}(\cdot)$  is linear, the iteration (3.16) and (3.17) calculates the principal direction. In other cases, this presents a case of *nonlinear PCA* (NPCA). The reference further argues that a saturating nonlinearity implements a PCA algorithm that is robust against outliers. However, note that the components of

<sup>4</sup>The set of the eigenvalues is often called eigenvalue spectrum.

the NPCA algorithm are not ordered according to the eigenvalues as in ordinary PCA. Rather, they are ordered according to the nonlinear eigenvalues defined as the local maxima of the objective function  $g(\cdot)$ .

## 3.2 Independent component analysis

In the previous section, noncorrelated components were achieved using PCA or FA methods. However, this only solves the source-separation problem for Gaussian sources. In this section, we review a method called independent component analysis (ICA) that can be used to solve the BSS problem in non-Gaussian cases as well. This is done by assuming the sources statistically independent and to have non-Gaussian marginal distributions<sup>5</sup>.

Thus, in this section we assume that the sources are independent, they are linearly mixed with a stationary and instantaneous mixing and there exist at least as many nondegenerate mixtures as there are sources. Furthermore, we assume that at most one of the sources has a Gaussian distribution. We as well assume that there exist an infinite amount of independent samples of the mixtures.

ICA can be used either to solve the BSS problem, or as a feature extraction technique. In BSS, the main focus is the determination of the underlying independent sources. This is the main goal when attempting to identify artefacts and signals of interest in biomedical systems (see Sec. 6.2 for several references). It has also been used to blindly separate audio signals (Torkkola, 1999), or to demix multispectral images (Parra et al., 2000, Funaro et al., 2003).

The other central application for ICA is feature extraction, where it provides a set of bases, which can be used to represent the observed data. So far, some of the main applications of this feature extraction strategy include the study of natural image statistics (Hurri et al., 1996, Bell and Sejnowski, 1997) and the development of computational models for human vision (Hoyer and Hyvärinen, 2000, Hyvärinen et al., 2001a), although it has been as well used for denoising (Hyvärinen, 1999b).

The references given above do not exhaust the applications of ICA, on the contrary. ICA research has found applications in a multitude of fields during its relatively short history. For further applications see the books by Lee (1998), Girolami (2000), Hyvärinen et al. (2001b), Roberts and Everson (2001), Cichocki and Amari (2002), Stone (2004) and the proceedings of the conference-series on ICA and BSS (ICA99, ICA00, ICA01, ICA03, ICA04).

In the previous section, we reviewed an NPCA algorithm, based on an article by Oja et al. (1991). Later on, it was found out by Karhunen and Joutsensalo (1994)

---

<sup>5</sup>To be precise, at most one of the sources can have Gaussian distribution.

and by Oja (1995, 1997) that NPCA actually does more than decorrelates the data  $\mathbf{X}$  in a robust manner. In case the sources  $\mathbf{S}$  have non-Gaussian distribution and the nonlinearity  $\mathbf{f}_e$  is chosen properly, NPCA carries out ICA and thus solves the BSS problem.

In this section, we discuss ICA in more detail and review several other algorithms to implement it. There are several principles that can be used to arrive at an ICA algorithm. For instance, the mutual information (2.20) between the source estimates may be minimised. Or one may minimise the KL divergence (2.21) between the joint probability distribution of the source estimates and the product of the marginal densities. The subsequent reviews of ICA algorithms follow the structure by Hyvärinen et al. (2001b), but other reviews that derive several previously suggested algorithms from a common criterion have also been written (cf. Lee et al., 2000, Cardoso, 2003, Parra and Sajda, 2003).

In Secs. 3.2.2–3.2.3, we review some approaches that assume the noise  $\nu$  negligible. More precisely, they assume that the independent components  $\mathbf{S}$  can be recovered by an demixing or separating matrix  $\mathbf{W}$  from the sphered data  $\mathbf{Y}$  by  $\mathbf{S} = \mathbf{W}\mathbf{Y}$  or by an unsphered separating matrix  $\mathbf{B}$  from the original data  $\mathbf{X}$  by  $\mathbf{S} = \mathbf{B}\mathbf{X}$ . In Sec. 3.2.4, we review some approaches that explicitly take the noise  $\nu$  into account. In Sec. 3.2.5, we review a Bayesian method to implement ICA. The last section 3.5, reviews some existing work to relax the assumptions for ICA.

### 3.2.1 FastICA: ICA by maximisation of non-Gaussianity

Reconsider the second source separation example in Sec. 3.1. The data after PCA and sphering of the variances is shown again in Fig. 3.3a. It was stated already that the sphered components do not have Gaussian marginal distributions. This becomes evident in Fig. 3.3b where the normalised histograms of the sphered components are plotted together with the standardised Gaussian distribution. The distributions of the sphered components are clearly super-Gaussian (recall the definition from the Sec. 2.3.2) because of the higher peaks and heavier tails than the Gaussian distribution. The projection directions where the independent components lie, are shown by the dashed and dot-dashed lines. The distributions of the projections in these directions, depicted in Fig. 3.3c are even peakier around zero and show even greater tails.

This indicates that at least in this case, a way to estimate the independent components would be to maximise the non-Gaussianity of the projections  $\mathbf{w}^T\mathbf{Y}$ . Could this be the case more generally as well? The *central limit theorem* (CLT, c.f. Papoulis, 1991) states that the distribution of the sum of independent variables tends to Gaussian when the number of variables increase. Vice versa, it may be thought that the less Gaussian a projection from the sphered data



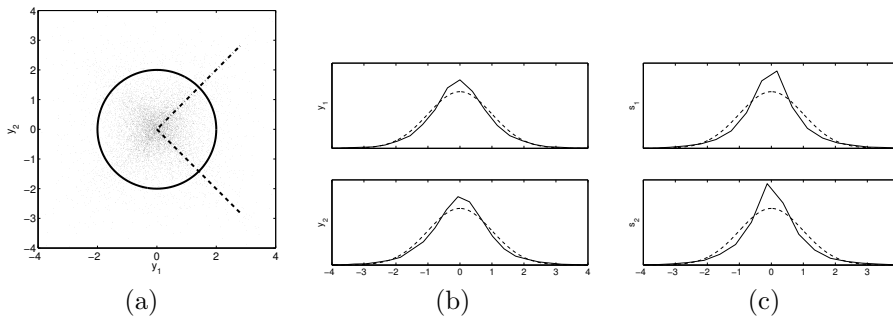


Figure 3.3: *a) Sphered data, reprinted from Fig. 3.2b b). Marginal distributions of the sphered components. The dashed line indicates the Gaussian distribution. c) Marginal distributions of the true sources. Again, the dashed line indicates the Gaussian distribution.*

$\mathbf{Y}$  is, the more independent it is of the other orthogonal projections. Following this inverse intuition, an algorithm for ICA might be designed by maximising a measure of the non-Gaussianity of the source estimates.

Perhaps the simplest method to measure the non-Gaussianity of the source estimates is to use the sample *kurtosis* (see Sec. 2.3.2) defined as:

$$\text{kurt}(\mathbf{s}) = \sum_t s^4(t)/T - 3 \left( \sum_t s^2(t) \right)^2 / T. \quad (3.18)$$

For a signal with zero mean and unit variance, this simplifies to  $\text{kurt}(\mathbf{s}) = \sum_t s^4(t)/T - 3$ . More generally, non-Gaussianity can be measured by *negentropy* (again, see Sec. 2.3.2). This would lead to maximising

$$g(\mathbf{s}) = N(s) = H(\nu) - H(s), \quad (3.19)$$

where  $H(s)$  and  $H(\nu)$  are the differential entropies of the random variable  $s$  and a Gaussian variable  $\nu$  with same mean and variance as  $s$ , respectively. However, this definition does not lend itself for easy implementation, because it needs the estimation of the marginal distribution  $p_s(\mathbf{s})$  from the source estimate  $\mathbf{s}$ . Thus some approximations are usually used.

FastICA (Hyvärinen and Oja, 1997, Hyvärinen, 1999a) is a family of algorithms derived from a general objective function:

$$N_g(\mathbf{s}) = |g(\mathbf{s}) - g(\nu)|^p, \quad (3.20)$$

where  $g$  is said to be any even, non-quadratic, sufficiently smooth scalar function.  $\boldsymbol{\nu}$  is a standardised Gaussian data vector.  $p$  is a positive integer, usually 1 or 2. Compared to Eq. (7) by Hyvärinen (1999a), we have used the data-vector notation and thus dropped the expectations. Additionally, the upper-case letter  $G$  has been replaced by the lower-case letter  $g$ . This contrast function can be seen as an approximation of *negentropy* (Hyvärinen, 1998b). The objective function is usually called *contrast function* since it measures the deviation of the distribution of the source estimate  $\mathbf{s}$  from the Gaussian distribution.

Hyvärinen (1999a) suggests several all-purpose functions  $g$  for the basis of the contrast function  $N_g$  (the derivatives  $\mathbf{f}_i$  of the contrast functions are given for future use):

$$g_1(\mathbf{s}) = \frac{1}{aT} \sum_t \log \cosh(as(t)), \quad \mathbf{f}_1(\mathbf{s}) = \tanh(as) \quad (3.21)$$

$$g_2(\mathbf{s}) = \frac{1}{4T} \sum_t s^4(t), \quad \mathbf{f}_2(\mathbf{s}) = \mathbf{s}^3, \quad (3.22)$$

where  $1 \leq a \leq 2$  and  $\mathbf{s}^3 = [s^3(1) s^3(2) \dots s^3(T)]$ . It is said that  $g_1$  is generally usable while  $g_2$  is justified only for sub-Gaussian distributions when no outliers exist. This is because  $s^4$  grows very rapidly and heavy tails or outliers would then dominate the measure. Application of contrast function  $g_2$  actually maximises the *kurtosis* (3.18). In general, it does not provide a very robust method to estimate the independent components. On the other hand, it has some nice analytical properties for which reason it is introduced here (Hyvärinen, 1999a).

In FastICA, the contrast function (3.20) is maximised using an approximate Newton iteration. The details of the derivation fall outside the scope of this introduction and we only give the algorithm:

$$\mathbf{w}^+ = \mathbf{Y}\mathbf{f}^T(\mathbf{w}\mathbf{Y})/T - f'(\mathbf{w}^T\mathbf{Y})\mathbf{w} \quad (3.23)$$

$$\mathbf{w}_{\text{new}} = \text{orth}(\mathbf{w}^+), \quad (3.24)$$

where  $f'(\mathbf{w}^T\mathbf{Y})$  is the derivative of function  $\mathbf{f}(\cdot)$ . Note also that the usual simple normalisation has been replaced by the orthonormalisation (3.24). This is done to be ready to apply the deflation (Sec. 3.1.2). The first term  $\mathbf{Y}\mathbf{f}^T(\mathbf{w}\mathbf{Y})/T$  determines the stable points of the algorithms while the second term  $f'(\mathbf{w}^T\mathbf{Y})\mathbf{w}$  significantly speeds up convergence. It has been proven that the asymptotic convergence of the FastICA algorithm, i.e. when there is an infinite amount of samples, is at least quadratic, usually cubic (Hyvärinen, 1999a, Oja, 2002) when the ICA model holds. This is much faster than simple gradient-ascent-based optimisation algorithms. Furthermore, Hyvärinen and Oja (1997) prove that with kurtosis as the contrast function, the FastICA converges asymptotically to the independent components.

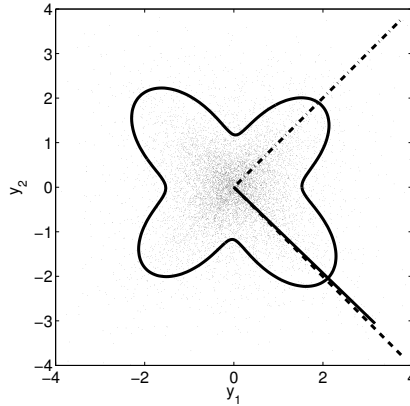


Figure 3.4: *The curve indicates the kurtosis of different projections  $\mathbf{w}^T \mathbf{Y}$ . The solid line corresponds to the direction of maximal kurtosis, whereas the dashed and dot-dashed lines indicate the true original sources.*

Let us now finally solve the second source separation example in Sec. 3.1 using FastICA. In contrast to the variance of the sphered data, the value of the contrast function  $N_{g_2}(\mathbf{w}^T \mathbf{Y})$  (3.22 applied to 3.20 with  $p = 1$ ) is not constant but depends on the direction of the projection  $\mathbf{w}$ . This is illustrated by the curve in Fig. 3.4. The direction giving the maximum value of the contrast function is indicated in a solid line. The dashed and the dot-dashed lines depict the directions where the original sources can be found. It is evident that maximisation of the kurtosis is able to recover the original sources. FastICA converged in the first independent component in five iterations. The second component, achieved using deflation (see Sec. 3.1.2), was fixed by the orthogonalisation procedure and thus converged in one iteration.

### 3.2.2 ICA by maximum likelihood estimation

For any linear model  $\mathbf{X} = \mathbf{A}\mathbf{S}$ , there holds the following relation between the marginal probability distribution  $p_{\mathbf{x}}(\mathbf{X})$  of the data  $\mathbf{X}$ , and  $p_{\mathbf{s}}(\mathbf{S})$  of the latent variables  $\mathbf{S}$ :

$$p_{\mathbf{x}}(\mathbf{X}) = |\det \mathbf{B}| p_{\mathbf{s}}(\mathbf{S}), \quad (3.25)$$

where  $\mathbf{B} = \mathbf{A}^{-1}$ . The joint marginal distributions have been indicated by the bold subscript, in concordance with Sec. 2.3.2. When the latent variables are

independent of each others, this can be written as

$$p_{\mathbf{x}}(\mathbf{X}) = |\det \mathbf{B}| \prod_j p_j(\mathbf{b}_j^T \mathbf{X}), \quad (3.26)$$

where the relation  $\mathbf{S} = \mathbf{B}\mathbf{X}$  has been made explicit. It is usual to assume the samples  $\mathbf{x}(t)$  independent for all  $t = 1, 2, \dots, T$ . It is then easy to write down the likelihood of the measured data  $\mathbf{X} = [\mathbf{x}(1) \cdots \mathbf{x}(t) \cdots \mathbf{x}(T)]$  in function of the demixing matrix  $\mathbf{B}$  :

$$g(\mathbf{B}) = \prod_{t=1}^T \prod_{j=1}^N p_j(\mathbf{b}_j^T \mathbf{x}(t)) |\det \mathbf{B}|, \quad (3.27)$$

where  $\mathbf{x}(t)$  means all of the observed values at time instance  $t$  collected in a column vector. Usually it is more practical to maximise the logarithm of the likelihood  $g(\mathbf{B})$ . The original likelihood is naturally maximised for the same parameter values. Then the products becomes summations and Eq. (3.27) simplifies to

$$\log g(\mathbf{B}) = \sum_{t=1}^T \sum_{j=1}^N \log p_j(\mathbf{b}_j^T \mathbf{x}(t)) + TN \log |\det \mathbf{B}|. \quad (3.28)$$

The maximisation of this likelihood would be easy by gradient methods. However, there is one problem to it. The likelihood of the data  $\mathbf{X}$  actually also depends on the source distributions and not only on  $\mathbf{B}$ . But the source distributions are usually unknown and cumbersome or even intractable to estimate from the data  $\mathbf{X}$ . Some approximations are thus needed.

However, Hyvärinen et al. (Theorem 9.1, 2001b) prove that the approximation for the source distributions need not be very exact. Let  $\tilde{p}_j$  denote the *assumed* distribution of the  $j$ th independent component and

$$q_j(\mathbf{s}_j) = \frac{\partial}{\partial \mathbf{s}_j} \log \tilde{p}_j(\mathbf{s}_j) = \frac{\tilde{p}'_j(\mathbf{s}_j)}{\tilde{p}_j(\mathbf{s}_j)}. \quad (3.29)$$

Then, for the ML estimator to be locally consistent, it is only needed that

$$\mathbb{E}\{s_j q_j(\mathbf{s}_j) - q'_j(\mathbf{s}_j)\} > 0, \quad (3.30)$$

for all  $j$ . This means that as long as the assumed distributions  $\tilde{p}_j$  do not make the expectation (3.30) negative, the ML estimation converges (locally) to the correct maximum.

Some practical choices for the approximations of the source distributions are then needed. Hyvärinen et al. (Theorem 9.1, 2001b) have further suggested and proven that to reasonably approximate the source distributions, only a very

simple one-parameter approximation is needed. Basically they use one approximation for super-Gaussian source distributions, and another one for sub-Gaussian ones. Then a binary parameter is needed to switch between these two approximations. They suggest the following log-distributions as an example:

$$\log \tilde{p}_j^+(\mathbf{s}) = \alpha_1 - 2 \log \cosh s \quad (3.31)$$

$$\log \tilde{p}_j^-(\mathbf{s}) = \alpha_2 - (s^2/2 - \log \cosh s), \quad (3.32)$$

where  $\tilde{p}_j^+(\mathbf{s})$  and  $\tilde{p}_j^-(\mathbf{s})$  are used to denote the super-Gaussian and the sub-Gaussian distributions respectively and  $\alpha_1$  and  $\alpha_2$  are some suitable constants.

After the derivation of these log-distributions, the expectation (3.30) becomes

$$2 \mathbb{E}\{-\tanh(s_j)s_j + (1 - \tanh^2(s_j))\} \text{ for } \tilde{p}_i^+ \quad (3.33)$$

$$\mathbb{E}\{\tanh(s_j)s_j - (1 - \tanh^2(s_j))\} \text{ for } \tilde{p}_j^-. \quad (3.34)$$

What is important in these equations is that they have always opposite signs. Then, it is (nearly) always possible to choose the approximation that gives the consistent ML estimation. Thus these approximations can be used for practically any source distributions. The only limitation is that the above equations should not give zero for the target distribution. In this case, the ML estimation using these approximations is not possible. This is similar to the case of trying to extract zero-kurtosis sources using kurtosis as the objective function.

It is now possible to construct ICA algorithms using the ML method, based on the log-likelihood (3.28) and the approximations of the source distributions (3.31) and (3.32). At each iteration, one can determine which approximation one should use by checking the signs of Eqs. (3.33) and (3.34).

The derivation of FastICA using CLT in an inverse manner in Sec. 3.2.1 seems somewhat heuristic. It turns out that the FastICA algorithm can be derived from the ML principle, making the FastICA algorithm sound. The details are omitted here but can be found in the comprehensive ICA book by Hyvärinen et al. (2001b).

In the following sections, we review two ICA algorithms that can be seen as ML estimations and acknowledge the connection of FastICA to ML estimation.

### Bell-Sejnowski

One popular ICA algorithm is the Bell-Sejnowski algorithm (Bell and Sejnowski, 1995). It was first derived from the *Infomax* principle that maximises the entropy of outputs of a nonlinear network. Cardoso (1998) showed that the algorithm can be derived as a stochastic gradient method for the log-likelihood (3.28). The algorithm arrives at the following update rule:

$$\Delta \mathbf{B} \propto [\mathbf{B}^T]^{-1} + \mathbf{F}(\mathbf{B}\mathbf{X})\mathbf{X}^T, \quad (3.35)$$

where  $\mathbf{F} = [\mathbf{f}_1^T \dots \mathbf{f}_j^T \dots \mathbf{f}_N^T]^T$  contain the nonlinearities derived as approximations to Eq. (3.29) and  $\mathbf{f}_j(\mathbf{s}_j)$  is a row vector containing the nonlinearities applied to each of its elements  $s_j(t)$ . For each source estimate, we may use different nonlinearities  $\mathbf{f}$ , depending on whether the current source-estimate suggests super- or sub-Gaussian distribution. For instance, Eqs. (3.31) and (3.32) may be used for super- and sub-Gaussians respectively, arriving at the following nonlinearities

$$\mathbf{f}^+(\mathbf{s}) = -2 \tanh \mathbf{s} \quad (3.36)$$

$$\mathbf{f}^-(\mathbf{s}) = \tanh \mathbf{s} - \mathbf{s}. \quad (3.37)$$

The Bell-Sejnowski algorithm usually suffers from slow convergence. Furthermore, calculation of one iteration is rather intensive because of the matrix inversion. This can be avoided by presphering the data. Moreover, a natural gradient can be used instead of the stochastic gradient.

### Natural gradient

In Sec. 2.4, we noted that a natural gradient is often more justified choice for a gradient ascent algorithm. In the case of ICA, applying the natural gradient results in (Amari, 1998)

$$\Delta \mathbf{B} = \left( \mathbf{I} + \mathbf{F}(\hat{\mathbf{S}})\hat{\mathbf{S}}^T \right) \mathbf{B}, \quad (3.38)$$

where  $\hat{\mathbf{S}} = \mathbf{B}\mathbf{X}$  are the source estimates and  $\mathbf{F}$  contains the nonlinearities as in the Bell-Sejnowski algorithm (3.35).

### 3.2.3 Some other methods for ICA

Several other approaches to ICA has been proposed, too. In the following, we review two most frequently used ICA algorithms in addition to already mentioned ones.

#### JADE

In Sec. 3.1, it was seen that eigenvalue decomposition of the covariance matrix leads to non-correlated data. This idea of diagonalisation can be generalised in higher-order correlations than  $\mathbf{X}\mathbf{X}^T$  using the so called cumulant *tensors*. As an example, let us consider the joint approximate diagonalisation of eigenmatrices (JADE, Cardoso, 1990). The idea is to compute several cumulant tensors  $\mathbf{F}(\mathbf{M}_i)$ , where  $\mathbf{F}$  represents the cumulant tensor and  $\mathbf{M}_i$  the corresponding eigenmatrices. These tensors are diagonalised jointly as well as possible. A possible objective

function for the goodness of the joint diagonalisation process is

$$g_{\text{JADE}}(\mathbf{B}) = \sum_i \|\text{diag}(\mathbf{B}\mathbf{F}(\mathbf{M}_i)\mathbf{B}^T)\|^2, \quad (3.39)$$

which is practically the sum of all of the diagonal elements in all of the diagonalised cumulant tensors. Maximisation of this cost function minimises the sum of the off-diagonal terms in the nearly diagonal matrices  $\mathbf{B}\mathbf{F}(\mathbf{M}_i)\mathbf{B}^T$ .

### Jutten-Hérault algorithm

If two sources  $s_1$  and  $s_2$  truly are independent, then correlation between any nonlinear functions of them should be zero:

$$E\{f_i(s_1)f_j(s_2)\} = 0. \quad (3.40)$$

The pioneering algorithm in ICA by Jutten and Herault (1991) uses this approach. The update rule for the algorithm is

$$\begin{aligned} \Delta \mathbf{A}_{ij} &\propto \mathbf{f}_1(\mathbf{s}_i)\mathbf{f}_2^T(\mathbf{s}_j), \text{ for } i \neq j, \\ \Delta \mathbf{A}_{ij} &= 0, \text{ for } i = j, \end{aligned} \quad (3.41)$$

where the  $\mathbf{s}_i$  are computed at each iteration according to  $\mathbf{S} = (\mathbf{I} + \mathbf{A})^{-1}\mathbf{X}$ . The nonlinearity  $\mathbf{f}_k(\mathbf{s}_i) = [f_k(s_i(1)) \cdots f_k(s_i(t)) \cdots f_k(s_i(T))]$  is a row vector containing the nonlinearity  $f_k$  applied to each element of  $\mathbf{s}_i$  separately. The Jutten-Hérault algorithm has also been a fruitful basis for other algorithms (c.f. Cichocki et al., 1994, Cichocki and Unbehauen, 1996, Cichocki et al., 1997).

### 3.2.4 ICA models considering noise explicitly

It became evident when the source separation problem was introduced in the beginning of Sec. 3 that the separating matrix  $\mathbf{B}$  can be identified when the noise covariance has the simple form (3.2). In this case, the demixing matrix cannot be used to recover the original sources, but only noisy versions of them:  $\tilde{\mathbf{S}} = \mathbf{S} + \tilde{\mathbf{v}} = \mathbf{B}(\mathbf{X} + \tilde{\mathbf{v}})$ . In this section, we assume that the separating matrix  $\mathbf{B}$  and the noisy sources  $\tilde{\mathbf{S}}$  have been solved, already. Detailed description of these noisy ICA methods is not possible in the scope of this thesis and we confine to review one of them in a very concise manner and give some further references:

- Douglas et al. (1998), a bias-removal technique.
- Hyvärinen (1998a), a shrinkage method, considered in more detail below.
- Books considering several noisy models: Lee (1998), Hyvärinen et al. (2001b), Cichocki and Amari (2002).

- Bayesian techniques: Knuth (1998), Attias (1999), Lappalainen (1999), Miskin and MacKay (2001), Choudrey and Roberts (2001), Højen-Sørensen et al. (2002), Chan et al. (2003). The Bayesian approach is considered in more detail in Sec. 3.2.5.

### Shrinkage removal of the noise

In this section, we review one shrinkage estimation method, based on Hyvärinen (1999b). The main results therein are given in Hyvärinen et al. (pp. 299–302, 2001b). The intuition behind the idea of shrinkage estimation is simple: strongly super-Gaussian sources have relatively few small values (but many zeroes). Thus, when a small value is encountered at time instance  $t$ , it is probable that it is solely noise and  $s(t) = 0$ . The method itself is rather complex. However, noise covariance of the form (3.2) constitutes an interesting and tractable special case. In particular, the noiseless source estimates are given by

$$\mathbf{S} = \mathbf{F}(\tilde{\mathbf{S}}), \quad (3.42)$$

where the elements of the matrix-valued function  $\mathbf{F}$  are obtained by inverting the relation

$$f_{jt}^{-1}(\tilde{s}_j(t)) = \tilde{s}_j(t) + \sigma^2 q'_j(\tilde{s}_j(t)). \quad (3.43)$$

$\sigma^2$  is the noise variance and  $q'_j$  is the derivative of the logarithm of the marginal pdf of the corresponding source  $\mathbf{s}_j$  (3.29). For details, see Hyvärinen et al. (2001b). The inversion may be impossible analytically. However, Hyvärinen et al. (2001b) considers three special cases where it is possible. We review one of them below.

Consider a source  $\mathbf{s}$  having Laplacian marginal pdf:  $p(s) = \exp(-\sqrt{2}|s|)/\sqrt{2}$ . The derivative of the logarithm becomes  $q'(s) = \sqrt{2}\text{sign}(s)$  and  $f$  is given by

$$f(\tilde{s}_j(t)) = \text{sign}(\tilde{s}_j(t)) \max(0, |\tilde{s}_j(t)| - \sqrt{2}\sigma^2). \quad (3.44)$$

This shrinkage function is plotted in solid line in Fig. 3.5 with noise variance  $\sigma^2 = 0.3$ . The dash-dotted line is  $f(\tilde{s}_j(t)) = \tilde{s}_j(t) - \tanh \tilde{s}_j(t)$  which has the same convergence points as FastICA using nonlinearity (3.21), due to the indeterminacy similar to  $\beta(\boldsymbol{\theta})$  in the general fixed point algorithm (2.34). This actually means that the nonlinearity of FastICA can be used for noise removal as well. This kind of approach has been used by Valpola (2004). However, in case of this shrinkage function, it is not easy to derive the corresponding pdf.

### 3.2.5 Bayesian ICA using ensemble learning

Recently, a Bayesian approach called ensemble learning (EL, recall Sec. 2.3.5 for details) has been applied to ICA by several researchers (Attias, 1999, Lap-



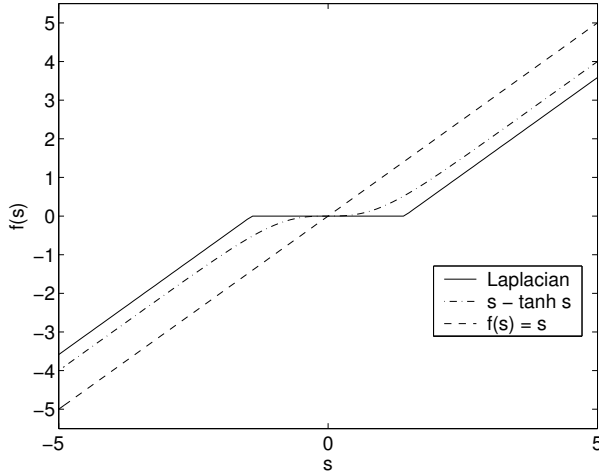


Figure 3.5: *Shrinkage functions.* The solid line corresponds to the Laplacian distribution and the dash-dotted line to  $f(\tilde{s}_j(t)) = \tilde{s}_j(t) - \tanh \tilde{s}_j(t)$ . The dashed line corresponding to  $f(\tilde{s}_j(t)) = \tilde{s}_j(t)$  is given for comparison.

palainen, 1999, Miskin and MacKay, 2001, Choudrey and Roberts, 2001, Højensørensen et al., 2002, Chan et al., 2003). The approach is often called the variational Bayes method, as well. Here, we concentrate on the Bayesian ICA (BICA) introduced by Valpola in Lappalainen (1999)<sup>6</sup>. There the form of the posterior approximation is factorised Gaussian. For this end, the noise  $\nu$  is assumed to have Gaussian distribution, with means  $b_i$  and variances  $e^{2\sigma_i}$ . The source distributions are modelled by mixtures of Gaussians (MoG):

$$p(s_j(t)|\mathbf{c}_j, \boldsymbol{\mu}_j, \boldsymbol{\gamma}_j) = \frac{\sum_i e^{c_{ji}} N(s_j(t); \mu_{ji}, e^{2\gamma_{ji}})}{\sum_j e^{c_{ji}}}. \quad (3.45)$$

The parameters  $\mathbf{c}$  are the logarithms of the mixture coefficients,  $\boldsymbol{\mu}$  the means and  $\boldsymbol{\gamma}$  the logarithms of the standard deviations of the Gaussians.  $N(a; b, c)$  denotes a Gaussian distribution over  $a$  with mean  $b$  and variance  $c$ . The variances of the Gaussians are parameterised by the logarithms of the standard deviations in order to make the assumption of roughly Gaussian posterior pdf valid.

Another possibility for posterior approximation is to use the so called conjugate priors. Their advantage is that they ensure that the posterior distribution has

<sup>6</sup>The present name of the author is Harri Valpola.

the same form as the prior. However, their use is problematic for hierarchical models, especially when there are hierarchies including variance parameters.

Valpola et al. (2001) developed several general purpose building blocks for hierarchical Bayesian modelling using ensemble learning. These blocks make it relatively easy to implement the Bayesian ICA. The building blocks enable also further modelling of the various parts of the ICA model. See Valpola and Karhunen (2002) for a detailed construction of some extensions on the ICA model under the EL framework.

Minimising the cost function used in EL (2.25) automatically estimates the correct number of various parameters in the model such as the number of sources. Often this can be implemented by starting with a big number of sources and then pruning away those that were not used, or vice versa by starting from a small number of sources and creating new sources as needed. Naturally, it is as well possible to go through the estimations using several models with different number of sources.

Valpola and Pajunen (2000) developed a fast version of Bayesian ICA (FBICA), using the same model as in BICA (3.45). It first derives FastICA from a low-noise approximation of the expectation-maximisation algorithm (EM, see Dempster et al., 1977, Bermond and Cardoso, 1999)<sup>7</sup>. Then the low-noise restriction is slackened by introducing the EL framework.

### 3.3 Temporal methods

ICA can solve the source separation problem when the sources are independent and have non-Gaussian distributions. However, reconsider the sphered data shown in Fig. 3.2a. After sphering, there is no more structure visible in the scatter-plot. This means that the marginal distributions of the sphered components are Gaussian and there are no correlations between them. Thus, the source estimates are independent, but not separated. However, if there exist correlations between different time instances in the sources, i.e.  $p(s_i(t_1), s_i(t_2)) \neq p(s_i(t_1))p(s_i(t_2))$ , the correlations can be used to separate the sources. Note that this requires that the auto-correlation structures of the sources are different.

Several algorithms (c.f., Tong et al., 1991, Molgedey and Schuster, 1994, Belouchrani et al., 1997, Ziehe and Müller, 1998), have been suggested that achieve source separation by diagonalising jointly several delayed autocorrelation matrices  $\mathbf{X}^{\tau_1} \mathbf{X}^{\tau_2 T} / T$ , where the delays are defined by  $\tau_1$  and  $\tau_2$  and the elements of  $\mathbf{X}^{\tau}$  are the elements of the data matrix  $\mathbf{X}$  delayed by  $\tau$ , i.e.  $x_i^{\tau}(t) = x_i(t - \tau)$ .

In case the sources have both non-Gaussian marginal distributions and time structure, one benefits from combining ICA and the time-delay approach. Müller

---

<sup>7</sup>For connections of EL to EM, see Neal and Hinton (1999).

et al. (1999) combined JADE with the TDsep algorithm. Hyvärinen (2001) modelled the time structure using an AR-model. The remaining innovation process was made as non-Gaussian as possible. Another approach, based on the non-stationarity of the variances of the sources has been proposed by Pham and Cardoso (2001). Note that the time structure can also be used to solve the noisy ICA problem. This has been done for example by Koivunen et al. (2001).

### 3.4 Dynamical factor analysis

In this section, we review another model using time structure called dynamical factor analysis (DFA) that is applied to MEG signal analysis in Publication 4. The parameter estimation of DFA is based on EL that was discussed in Sec. 2.3.5, see also Sec. 3.2.5.

The model is a very general model of complex dynamical processes. Observations  $\mathbf{x}(t)$  are assumed to be generated by linear mixing  $\mathbf{A}$  from hidden states  $\mathbf{s}(t)$  including Gaussian white additive noise  $\boldsymbol{\nu}_x(t)$ . Additionally each state  $\mathbf{s}(t)$  for all  $t$  is generated from the previous states  $\mathbf{s}(t-1)$  by a nonlinear mapping  $\mathbf{f}$  with a Gaussian innovation process  $\boldsymbol{\nu}_s(t)$ . Mappings  $\mathbf{A}$  and  $\mathbf{f}$  are assumed to be independent of time. This results in a two part model:

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{A}\mathbf{s}(t) + \boldsymbol{\nu}_x(t) \\ \mathbf{s}(t) &= \mathbf{f}(\mathbf{s}(t-1)) + \boldsymbol{\nu}_s(t).\end{aligned}\tag{3.46}$$

The nonlinear mapping  $\mathbf{f}$  is modelled by a two-layer multi-layer-perceptron network (MLP, Haykin, 1999) with sigmoidal tanh's as the hidden layer nonlinearities. The overall model structure is shown in Fig. 3.6.

Factor analysis defines the mapping up to a rotation. This means that the learned states can be mixtures of each others, though they are not correlated. The dynamical mapping defines the rotation, but it is very slow to learn, if the MLP network is fully connected (Valpola and Karhunen, 2002). For this reason the dynamics of the factors is forced to be block-wise (see Fig. 3.6), which simplifies the network and encourages the model to find independent source processes. If the factors are modulated sinusoids as is the case in rhythmical activity, blocks of two factors suffice.

The posterior approximation of DFA that is used in EL and the learning algorithm are described in detail in Publication 4. They are quite similar to those of the nonlinear state-space model (Valpola and Karhunen, 2002).

A Bayesian framework for combining temporal and marginal-distribution information has been presented by Attias and Schreiner (1998). However, as is true with DFA as well, the algorithms are computationally rather intensive. A lighter

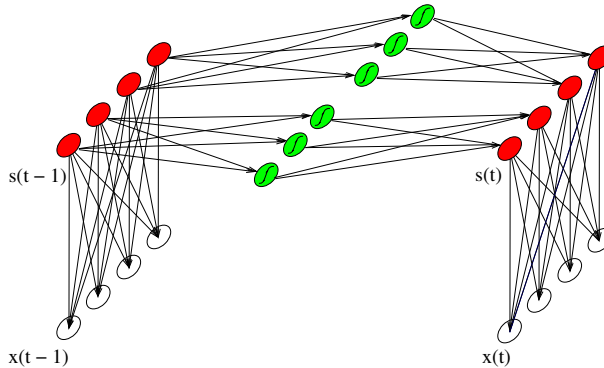


Figure 3.6: *Part of DFA model. Dark units are states, empty units the observations and the ones with a sigmoid inside correspond to the MLP dynamics. Direction of the arrows correspond to the direction of the causality: observations are caused by states and next states are caused by previous states. (from Publication 4)*

Bayesian algorithm for dynamical sources has been suggested by Hansen et al. (2001).

In the next chapter, we show that the procedure that was used to initialise the sources (Sec. 3.2 from the Publication 4) can be extended into full source separation framework. In this framework the usage of dynamics to guide the separation becomes simpler and computationally more efficient than in the algorithms derived from the Bayesian framework.

### 3.5 Relaxing the ICA assumptions

In Sec. 3.2, we assumed that the sources are independent, they are linearly mixed with a stationary and instantaneous mixing and there exist at least as many nondegenerate mixtures as there are sources. Furthermore, we assumed that at most one of the sources has a Gaussian distribution. We as well assumed that there exist an infinite amount of independent samples of the mixtures. In this section, we give the readers some pointers to existing literature that discusses separation of sources under assumptions relaxed from the ones given above.

The assumptions have been relaxed:

- in the sources, more specifically
  - dependent sources have been considered by: Hyvärinen and Hurri (2004), Tanaka and Cichocki (2004).

- Gaussian sources with time structure are discussed in Sec. 3.3.
- in the mixing, more specifically
  - nonlinearity has been considered by: Hyvärinen and Pajunen (1999), Valpola and Karhunen (2002), Almeida (2003),
  - and post-nonlinearity by: Taleb and Jutten (1999), Ziehe et al. (2003).
  - non-stationarity has been considered by: Everson and Roberts (2000).
  - convolution has been considered by: Haykin (2000), Olsson and Hansen (2004), Winter et al. (2004).
  - overcomplete and nonorthogonal representations have been considered by: Amari (1999), Lewicki and Sejnowski (2000), Hyvärinen and Inki (2002), Winter et al. (2004).
  - undercomplete representations have been considered by: Amari (1999), see also Sec. 5 and Publication 3.
- the noise is considered in Sec. 3.2.4

The above lists are not exhaustive. We again refer to the ICA books by Lee (1998), Girolami (2000), Hyvärinen et al. (2001b), Roberts and Everson (2001), Cichocki and Amari (2002), Stone (2004) and the proceedings of the conference-series on ICA and BSS (ICA99, ICA00, ICA01, ICA03, ICA04).

## Chapter 4

# Denoising source separation, a new approach

*He who has ears, let him hear.*

–The Bible: Matthew 13:9

In Secs. 3.2–3.4, we reviewed several solutions to the source separation problem. We noticed that it is possible to solve the source separation problem if the sources do not have Gaussian distributions or they have differing time structures.

In this section, we introduce a novel general framework of denoising source separation (DSS) that gathers the approaches of these two families under the same roof. This framework is more explicitly described in Publication 5. The accuracy, stability and speed of convergence in some of the algorithms developed under this framework are described in more detail in Publication 6.

In Publication 5, the DSS framework is derived from a generative linear model whose parameters are estimated using the EM algorithm (Dempster et al., 1977, Bermond and Cardoso, 1999). Namely, we assume that the data is (pre)sphered and the sources are estimated one-by-one. This results in a fixed point algorithm, similar to FastICA (Hyvärinen, 1999a). The nonlinearity in the algorithm is interpreted as a denoising step.

In this chapter, we take a more tutorial-like approach to DSS. We argue that source separation algorithms can be constructed around denoising principles. We show that some of the source separation algorithms in previous sections can explicitly be seen as carrying out denoising of the source estimates and using simple correlation based learning to estimate the demixing vectors.

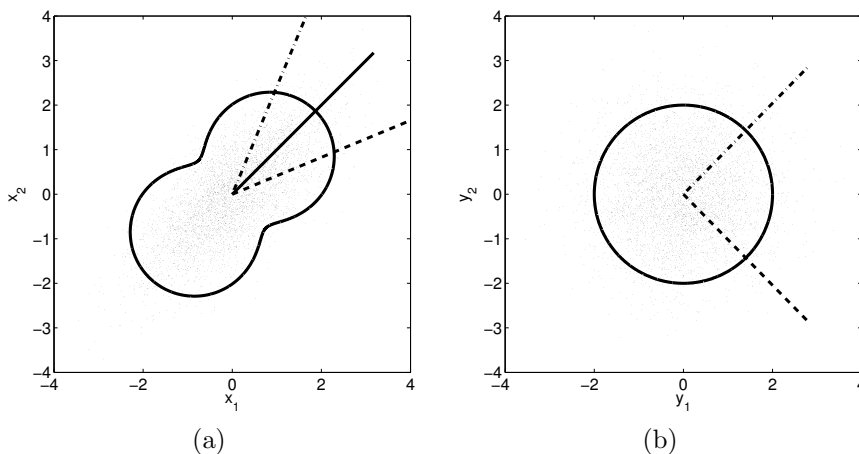


Figure 4.1: *a) The two-dimensional data from Fig. 3.1a. The true mixing coefficients are shown in dashed and dot-dashed lines. b) The normalised principal components of the data. The unity of variance is highlighted by the solid circle. Again, the true directions of the sources are shown in dashed and dot-dashed lines.*

## 4.1 A source separation example using DSS

Let us reconsider the first source separation example in this thesis, in Sec. 3.1. The data is redrawn in Fig. 4.1a and the sphered data in Fig. 4.1b. The projections that would yield the original sources are shown in dashed and dash-dotted lines, respectively. All orthogonal projections from the sphered data yield components which are decorrelated to each others and which have unit variance. Thus FA cannot be used to solve the source separation problem. Neither does the principal direction yield any of the sources.

However, the scatter plot loses any temporal structure the signals might have. A good representation for stationary temporal structure is given by amplitude spectrum computed using the discrete Fourier transform or the discrete cosine transformation (DCT). Consider then the amplitude spectra of the sphered components, shown in Fig. 4.2a. The spectra are dominated by low frequencies, giving rise to a hypothesis of existing slowly varying sources. Thus low-pass filtering, e.g. by a filter whose amplitude response is shown in bottom of Fig. 4.2a, should make the sources clearer, i.e. denoise the source estimates. The denoised data using the particular low-pass filter is shown in Fig. 4.2b.

In general, low-pass filtering decreases the energy of the signals and the re-

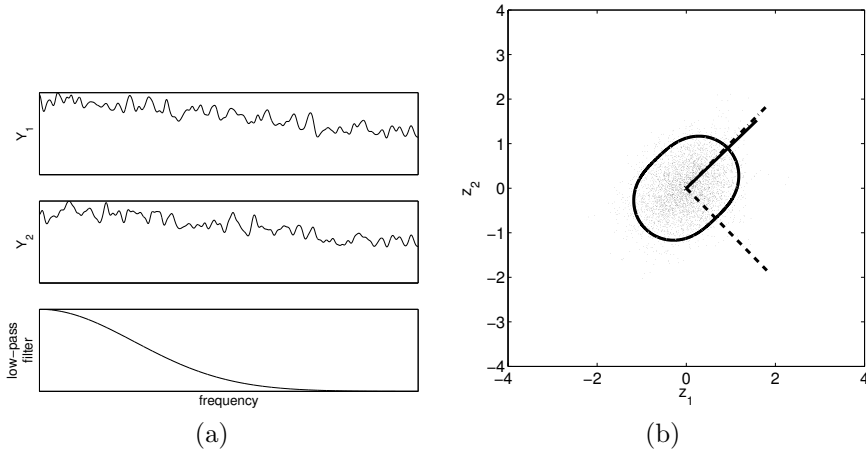


Figure 4.2: a) The amplitude spectra of the sphered data and the amplitude response of a low-pass filter. b) The sphered data after denoising using the low-pass filter, together with the standard deviations in each directions, direction of maximal variance and the true directions, similarly to previous figures.

maining energy depends on their frequency content. This means that projecting the data on a vector with unit length no longer yields a signal with unit variance. This is illustrated by the fact that the data cloud in Fig. 4.2b has shrunk. But more importantly, not all projections result in the same variance, as illustrated by the ellipsoid corresponding to proportion of the standard deviation. After low-pass filtering, it is therefore possible to identify the signals having higher than average proportion of low frequencies by PCA. This is manifested in the fact that the principal direction is aligned with the first source mixing vector and the second mixing vector is perpendicular to that. The projections to the principal and the minor components extract the original sources. The amplitude spectra of the source estimates are shown in Fig. 4.3a. From those, we may conclude that the first source had significant low frequencies not being i.i.d. samples though the marginal distribution is Gaussian. The second source seems to be Gaussian i.i.d. with no time structure.

Usually sphering is used in source separation algorithms because it provides a quick way to restrict the search space of separating vectors on a unit sphere. We stress that in case of DSS, the sphering has much greater role. Only the presphering makes it possible to identify the original sources by PCA on the denoised data. As a case of contrary, consider the denoising through low-pass filtering applied straight to the original data  $\mathbf{X}$ , illustrated in Fig. 4.3b. Note



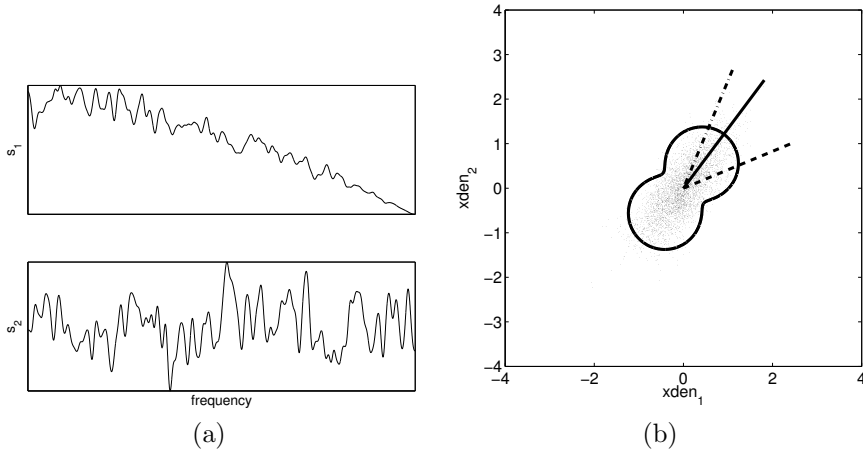


Figure 4.3: a) The amplitude spectra of the estimated sources. b) The original data  $\mathbf{X}$  after denoising using the low-pass filter.

that the maximal-variance direction does not align well with either of the projection directions of the sources, though it is closer to the original source having considerable low frequencies.

## 4.2 Linear DSS

The above source separation had three distinct phases: sphering, denoising and PCA. In the following, we formulate these in a unified framework. We note that instead of applying the denoising to the sphered data, it can be applied to the current source estimate. In Sec. 4.3, we then consider replacing the linear denoising with a nonlinear one.

Linear denoising can be mathematically expressed as matrix multiplication. Then, for denoising matrix  $\mathbf{D}^*$ , the denoised data  $\mathbf{Z}$  becomes:

$$\mathbf{Z} = \mathbf{Y}\mathbf{D}^*. \quad (4.1)$$

Note that  $\mathbf{D}^*$  operates on each signal  $\mathbf{y}_i$  separately, i.e. denoising is defined for one-dimensional signals. Furthermore, the denoising is performed over time, i.e.  $\mathbf{D}^*$  is  $T \times T$ -matrix. We have used the asterisk here to be congruent with Publication 5.

Though other methods exist, we propose to compute the first principal com-

ponent of the denoised data  $\mathbf{Z}$  by the classical *power method* (see Sec. 3.1.1):

$$\mathbf{w}^+ = \mathbf{Z}\mathbf{Z}^T \mathbf{w} \quad (4.2)$$

$$\mathbf{w}_{\text{new}} = \text{orth}(\mathbf{w}^+) \quad (4.3)$$

where  $\mathbf{w}_{\text{new}}$ , a column vector, means the new estimate of the principal direction and we have used the orthonormalisation in case we estimate several components in a deflationary or symmetric manner (see Sec. 3.1.2). The power method applied to the denoised data maximises the objective function:

$$g_{\text{lin}}(\mathbf{w}) = \mathbf{w}^T \mathbf{Z}\mathbf{Z}^T \mathbf{w}, \quad (4.4)$$

subject to the constraint  $\mathbf{w}^T \mathbf{w} = 1$  (again, refer to Sec. 3.1.1 for more details).

Let us now substitute the denoising (4.1) into the power method (4.2):

$$\mathbf{w}^+ = \mathbf{Z}\mathbf{Z}^T \mathbf{w} = \mathbf{Y}\mathbf{D}^* \mathbf{D}^{*T} \mathbf{Y}^T \mathbf{w}. \quad (4.5)$$

Further, let us denote

$$\mathbf{D} = \mathbf{D}^* \mathbf{D}^{*T}. \quad (4.6)$$

Then the classical power method applied to filtered data can be reformulated as follows:

$$\mathbf{s} = \mathbf{w}^T \mathbf{Y} \quad (4.7)$$

$$\mathbf{s}^+ = \mathbf{s}\mathbf{D} \quad (4.8)$$

$$\mathbf{w}^+ = \mathbf{Y}\mathbf{s}^{+T} \quad (4.9)$$

$$\mathbf{w}_{\text{new}} = \text{orth}(\mathbf{w}^+), \quad (4.10)$$

where  $\mathbf{s}$  is used to denote the current source estimate corresponding to the eigenvector estimate  $\mathbf{w}$ . Mathematically, algorithm (4.7)–(4.10) is equivalent to the classical power method applied to the filtered data (4.1)–(4.3). In the classical version, the denoising  $\mathbf{D}^*$  was applied to the whole data, but in Eq. (4.8) the denoising  $\mathbf{D}$  is applied to the current source estimate  $\mathbf{s}$ , instead. Equations (4.9) and (4.10) compute the weight vector which yields a new source which is closest to the denoised  $\mathbf{s}^+$  in the least-mean-squares (LMS) sense. We call Eqs. (4.7)–(4.10) the *linear DSS algorithm*. The corresponding objective function, starting from Eq. (4.4), can be written as

$$g_{\text{lin}}(\mathbf{s}) = \mathbf{w}^T \mathbf{Z}\mathbf{Z}^T \mathbf{w} = \mathbf{s}\mathbf{D}\mathbf{s}^T. \quad (4.11)$$

The deflation procedure (see Sec. 3.1.2) can be used to extract several components. For that reason, an orthonormalisation (4.10) has been used instead of simple normalisation.

### 4.3 Nonlinear DSS

In general, denoising is not restricted to linear operations. Median filtering is a clear example of nonlinear denoising which cannot be implemented as mere matrix multiplication. Another example of nonlinear denoising is encountered when the denoising is tuned adaptively to improving estimates of the source characteristics as the iteration progresses. A good review on nonlinear filtering is given by Kuosmanen and Astola (1997).

One common way to develop nonlinear algorithms<sup>1</sup>, such as ICA, from linear algorithms, such as PCA, is to replace the quadratic criterion (3.5) by a criterion which contains other than second-order moments. However, we argue that it is often easier and more practical to simply replace Eq. (4.8) by a nonlinear denoising step:

$$\mathbf{s}^+ = \mathbf{f}(\mathbf{s}). \quad (4.12)$$

The function  $\mathbf{f}(\mathbf{s})$  denotes the result of denoising, i.e. both  $\mathbf{s}$  and  $\mathbf{f}(\mathbf{s})$  are row vectors of the length  $T$ . In the linear case  $\mathbf{f}(\mathbf{s}) = \mathbf{s}\mathbf{D}$ , but in general, almost any type of denoising procedure can be applied. When more than one sources are estimated, it may be desirable to use the information in all of the sources  $\mathbf{S}$  for denoising any particular source  $\mathbf{s}_i$ . This leads to the following denoising function:  $\mathbf{s}_i^+ = \mathbf{f}_i(\mathbf{S})$ .

The objective function of the above nonlinear DSS algorithm is extensively discussed in Publication 5. Here it suffices to say that the exact derivation of the objective function is not usually necessary, since the algorithm is constructed around the denoising principle, not around optimisation of an objective function. In cases where the objective function would be needed and would be difficult to compute, we have suggested an approximation:

$$\hat{g}(\mathbf{s}) = \mathbf{s}\mathbf{f}^T(\mathbf{s}). \quad (4.13)$$

It is exact in the case of linear denoising and in some special nonlinear cases too, such as the cumulants.

Denoising is useful as such and therefore there is a wide literature of sophisticated denoising methods to choose from (c.f. Anderson and Moore, 1979). Moreover, one usually has some knowledge about the signals of interest and thus possesses the information needed for denoising. In fact, quite often the signals extracted by BSS techniques would be post-processed to reduce noise in any case (c.f. Vigneron et al., 2003, and Sec. 3.2.4). In the DSS framework, the available denoising methods can be directly applied to source separation, producing better results than purely blind techniques do. There are also very general noise

---

<sup>1</sup>By nonlinear, we refer to the nonlinearity imposed on the source estimate, not to the nonlinearity of the observation mapping. This is on a par with the use of the words in NPCA.

reduction techniques such as wavelet denoising (Donoho et al., 1995, Vetterli and Kovacevic, 1995)<sup>2</sup> or median filtering (Kuusmanen and Astola, 1997) which can be applied in BSS. The DSS framework thus suggests new algorithms ranging from BSS to highly detailed methods in specialised applications.

The deflational method (see Sec. 3.1.2) is readily available for the linear DSS algorithms. It turns out that the deflation is directly applicable to nonlinear DSS as well. The orthogonal constraints  $h(\mathbf{w}) = \mathbf{w}^T \mathbf{w}_i = 0$  do not affect the denoising procedure (see Sec. 2.4). Hence, the consecutive runs of the algorithm optimise the same  $g(\mathbf{w})$  as the first run but under the constraint of orthogonality to the previously extracted components. If more than one sources are estimated simultaneously, symmetric orthogonalisation methods can be used.

## 4.4 Denoising functions in practice

DSS is a framework for designing source separation algorithms. The idea is that the algorithms differ mainly in the denoising function  $\mathbf{f}(\mathbf{s})$  while the other parts of the algorithm remain mostly the same. In this section, we discuss both simple but powerful linear and sophisticated nonlinear denoising functions. The goal is to inspire others to try out their own denoising methods. The range of applicability of the examples spans from cases where the knowledge about the signals is relatively specific to almost blind source separation where very little is assumed about the signal characteristics.

Before, we note that it is usually not crucial for the denoising to be very exact. Otherwise DSS would not be very useful because one would only get what is asked from the algorithm in terms of the denoising function. Fortunately, this is not the case: assuming that the signals are recoverable by linear projections from the observations, it is enough for the denoising function  $\mathbf{f}(\mathbf{s})$  to remove more noise than signal (c.f. Hyvärinen et al., 2001b, Theorems 8.1 and 9.1). This is because the reestimation steps (4.9) and (4.10) constrain the source  $\mathbf{s}$  to the subspace spanned by the data. Even if the denoising discards parts of the signal, reestimation steps restore them.

In practice, the observations contain noise which does not fully disappear by any linear projection. Then the quality of the separated signals depends on the accuracy of the denoising. If there is no detailed knowledge about characteristics of the signals to start with, it is useful to bootstrap the denoising functions. This can be achieved by starting with relatively general signal characteristics and then tuning the denoising functions based on analyses of the structure in the noisy signals extracted in the first phase. In fact, some of the nonlinear DSS algorithms can be regarded as linear DSS algorithms where a linear denoising

---

<sup>2</sup>See Hesse and James, 2004, for a very similar approach to DSS, using wavelet denoising.

function is adapted to the sources.

#### 4.4.1 Detailed linear denoising functions

In this section, we consider several detailed, simple but powerful, linear denoising schemes. We introduce the denoisings using the denoising matrix  $\mathbf{D}$  when feasible. We consider effective implementation of the denoisings as well.

##### On/off-denoising

Consider designed experiments, e.g. in the field of biomedical systems. It is usual to control them by having periods of activity and non-activity. In such experiments, a denoising can be simply implemented by

$$\mathbf{D} = \text{diag}(\mathbf{d}), \quad (4.14)$$

where the diagonal matrix  $\mathbf{D}$  refers to the linear denoising in Eq. (4.8) and  $\mathbf{d} = [d_1 \cdots d_t \cdots d_T]$  determine the active periods:

$$d_t = \begin{cases} 1, & \text{for the active parts} \\ 0, & \text{for the inactive parts} \end{cases} \quad (4.15)$$

This amounts to multiplying the source estimate  $\mathbf{s}$  by a binary mask<sup>3</sup>, where ones represent the active parts and zeroes the non-active parts. Notice that this masking procedure actually satisfies  $\mathbf{D} = \mathbf{D}\mathbf{D}^T$ . This means that DSS is equivalent to the power method applied to the filtered data even with exactly the same filtering. In practice, this DSS algorithm could be implemented by PCA applied to the active parts of the data, while the sphering stage would still involve the whole data.

##### Denoising based on the frequency content

If, on the other hand, signals are characterised by having certain *frequency* components, one can transform the source estimate by DCT, mask the spectrum, e.g. with a binary mask, and inverse transform to obtain the denoised signal:

$$\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad (4.16)$$

where  $\mathbf{U}$  is the transform,  $\mathbf{\Lambda}$  is the matrix with the mask on its diagonal, and  $\mathbf{U}^T$  is the inverse transform. Again, a computational implementation of the algorithm needs not resort to matrix multiplications and it is possible to implement DSS by applying PCA on selected parts of the transformed data.

---

<sup>3</sup>By masking we refer to point-wise multiplication of a signal or of a transformation of the signal.

### Spectrogram denoising

Often a signal is well characterised by what frequencies occur at what times. This is evident, e.g. in burst-like oscillatory activity in the brain. An example of source separation in such data is studied in Sec. 6.3. The time-frequency behaviour can be described by calculating DCT in short windows in time. This results in a combined time and frequency representation, i.e. a spectrogram, where the masking can be applied.

There is a known dilemma in the calculation of the spectrogram: detailed description of the frequency content does not allow detailed information of the activity in time and vice versa. In other words, large amount of different frequency bins  $T_f$  will result in small amount of time locations  $T_t$ . Wavelet transforms (Donoho et al., 1995, Vetterli and Kovacevic, 1995) have been suggested to overcome this problem. There, an adaptive or predefined basis, different from the pure sinusoids used in Fourier transform or DCT, is used to divide the resources of time and frequency behaviour optimally in some sense. Another possibility is to use so called multitaper technique (Percival and Walden, 1993, Ch. 7). A simpler approach than these two is to use fixed length windows and to bootstrap their sizes when source characteristics emerge.

### Denoising of quasiperiodic signals

As a final example of denoising based on detailed source characteristics, consider Fig. 4.4a. There a source estimate  $\mathbf{s}$  has been reached. The apparent quasiperiodic structure of the signal can be used to perform DSS to get a better estimate. The denoising proceeds as follows:

1. Estimate the locations of the peaks of the current source estimate  $\mathbf{s}$  (Fig. 4.4b).
2. Chop each period from peak to peak.
3. Dilate each period to a fixed length (linearly or nonlinearly).
4. Average the dilated periods (Fig. 4.4c).
5. Let the denoised source estimate  $\mathbf{s}^+$  be a signal where each period has been replaced by the averaged period dilated back into the original length (Fig. 4.4d).

The denoised signal  $\mathbf{s}^+$  in Fig 4.4d show significantly better signal-to-noise ratio (SNR) compared to the original source estimate  $\mathbf{s}$ , in Fig. 4.4a.

This averaging is a form of linear denoising since it can be implemented as matrix multiplication. Furthermore, it presents another case in addition to the

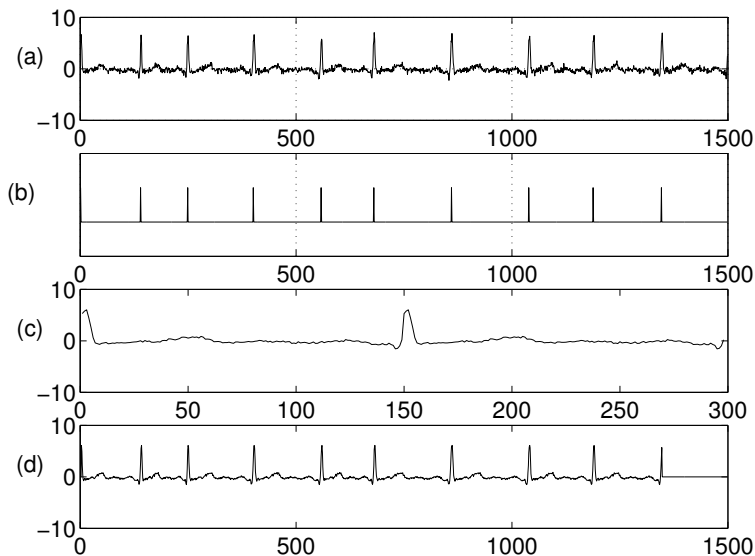


Figure 4.4: *a) Current source estimate  $\mathbf{s}$  of a quasiperiodic signal b) Peak estimates c) Average signal  $s_{\text{ave}}$ , for clarity two averages are shown concatenated. d) Denoised source estimate  $\mathbf{s}^+$ . (from Publication 5)*

binary masking, where DSS is equivalent to the power method for the denoised data even with exactly the same filtering. It would not be easy to see from the denoising matrix  $\mathbf{D}$  itself that  $\mathbf{D} = \mathbf{D}\mathbf{D}^T$ . However, this becomes evident should one consider the averaging of the source estimate  $\mathbf{s}^+$  (Fig. 4.4d) that is already averaged.

Note that there are cases where chopping from peak to peak does not guarantee the best result. This is especially true when the activity does not span the whole section from peak to peak, but there are parts where the response is silent. Then there is need to estimate the lengths of the periods separately.

#### 4.4.2 ICA using DSS

In the previous section, several linear denoising schemes were introduced. In all of them, the details of the denoising were assumed to be known. It is as well possible to estimate the denoising specifications from the data. This makes the denoising nonlinear or adaptive. In this section, we consider a particular ICA algorithm in the DSS framework, suggesting modifications which improve separation results

and robustness.

Consider one of the best known BSS approaches, FastICA optimising the sample kurtosis of the sources:

$$\mathbf{s} = \mathbf{w}^T \mathbf{Y} \quad (4.17)$$

$$\mathbf{s}^+ = \mathbf{s}^3 \quad (4.18)$$

$$\mathbf{w}^+ = \mathbf{Y} \mathbf{s}^{+T} / T - 3\mathbf{w}, \quad (4.19)$$

where  $\mathbf{s}^3 = [s^3(1) \cdots s^3(t) \cdots s^3(T)]$ . The orthogonalisation step has been omitted, because it does not affect the present analysis. This iterative algorithm is equivalent to Eq. (8.20) in Hyvärinen et al. (2001b). The first two steps (4.17) and (4.18) can be considered as intermediate results for the actual update of the projection vector (4.19). The expectation is replaced by matrix multiplication in Eq. (4.19) and thus the normalisation by  $T$  is needed. The term  $-3\mathbf{w}$  does not change the fixed points of the algorithm. The role of it is further discussed in Sec. 4.5.

The above FastICA algorithm can be considered as a case of DSS where the denoising step is  $\mathbf{f}(\mathbf{s}) = \mathbf{s}^3$ . The denoising interpretation of that step arises from the observation that one can interpret  $\mathbf{s}^3$  as being  $\mathbf{s}$  masked by  $\mathbf{s}^2$ , the latter being a somewhat naïve estimate of signal variance and thus relating to SNR.

Kurtosis as an objective function is notorious for being prone to overfitting and producing spiky source estimates (see Sec. 5, Publication 2 and Publication 3 for a more detailed description of the problem). The denoising interpretation gives rise to many improvements to the kurtosis-based ICA, improving its stability and robustness. For instance, it is not necessary to base the instantaneous variance estimate on only one sample, but on several instead. The improvements are discussed in more detail in Publication 5 and Publication 6.

### 4.4.3 Other denoising functions

There are cases where the system specification itself suggests some denoising schemes. One such case is encountered in CDMA signal separation (see Publication 5 for details). Another example is source separation with a microphone array combined with speech recognition. Many speech recognition systems rely on generative models which can be readily used to denoise the speech signals.

Sometimes the sources can be grouped to form interesting subspaces. This could happen, e.g. when all of the sources are not independent of each others, but there exists anyway subspaces that are mutually independent. Some form of subspace rules can be used to guide the extraction of interesting subspaces in DSS. The independence criterion can be further relaxed at the borders of the subspaces. This can be achieved by incorporating a neighbourhood denoising rule



in DSS, resulting in a topographic ordering of the sources. One such topographic rule was used in topographic ICA (Hyvärinen et al., 2001a).

It is possible to combine various denoising functions when the sources are characterised by more than one type of structure. Note that the combination order may be crucial for the outcome. This is simply because, in general,  $\mathbf{f}_i(\mathbf{f}_j(\mathbf{s})) \neq \mathbf{f}_j(\mathbf{f}_i(\mathbf{s}))$  where  $\mathbf{f}_i$  and  $\mathbf{f}_j$  represent two different linear or nonlinear denoisings (compare for instance Rivet et al., 2003 and Rivet et al., 2004).

Finally, a source may be almost completely known. Then it is possible to apply a detailed matched filter to estimate the mixing coefficients or the noise level. Detailed matched filters have been used in Sec. 4.6 to get an upper limit of the SNRs of the source estimates.

## 4.5 Speedup in DSS

As DSS is closely related to the power method, it suffers from slow convergence similar to the case of the power method with comparable principal eigenvalues. In Sec. 3.1.3, we reviewed the method of spectral shift to increase the speed of convergence of the power method in this case. This approach can also be used in DSS. In fact, denoisings of the form:

$$\mathbf{s}^+ = \alpha(\mathbf{s})[\mathbf{f}(\mathbf{s}) + \beta(\mathbf{s})\mathbf{s}] \quad (4.20)$$

result in the same fixed points as  $\mathbf{s}^+ = \mathbf{f}(\mathbf{s})$ .  $\alpha(\mathbf{s})$  and  $\beta(\mathbf{s})$  are some scalar functions.

However, in nonlinear DSS, the denoising is dependent on the current source estimate and this may make finding a suitable spectral shift difficult: DSS either converges slowly with too modest a spectral shift or ends up oscillating between two weight estimates with too enthusiastic a spectral shift. In this section, we first discuss suitable spectral shifts for DSS. Then we introduce a learning-rate term in DSS to ensure fast *and* stable convergence.

In the classical power method, the spectral shift usually applies to the eigenvector estimates of the covariance matrix  $\mathbf{X}\mathbf{X}^T/T$ . However, in DSS the spectral shift can be embedded in the denoising of the source using  $\beta(\mathbf{s})$  according to Eq. (4.20). In particular, the shift of eigenvalues by  $\lambda_\beta$  is implemented by  $\beta(\mathbf{s}) = \lambda_\beta/T$ . We use  $\beta(\mathbf{s})$  to implement the spectral shift in DSS.

A reasonable spectral shift is to move the eigenvalue  $\lambda$  corresponding to a Gaussian signal to zero. This, for instance, most effectively cancels Gaussian noise. This spectral shift can be approximatively implemented by

$$\beta = -\hat{g}(\boldsymbol{\nu})/T = \boldsymbol{\nu}\mathbf{f}^T(\boldsymbol{\nu})/T, \quad (4.21)$$

where  $\boldsymbol{\nu}$  stands for standardised Gaussian variable. The last part of the equation exploits the approximation of the objective function (4.13). One way to improve

the efficiency of this approach is to try to scale the denoising such that a Gaussian noise signal always has a similar contribution in the denoised signal. For example, if the denoising is implemented by masking the source signal, the contribution of a fixed amount of Gaussian noise to the denoised source signal can be equalised by normalising the sum of the masking components.

It is not necessary to base the spectral shift on a global approximation of  $g(\boldsymbol{\nu})$ . An alternative is to linearise  $\mathbf{f}(\mathbf{s})$  around the current source estimate  $\mathbf{s}$  and use this to compute  $\beta(\mathbf{s})$  as follows:

$$\hat{\mathbf{f}}(\boldsymbol{\nu}) = \mathbf{f}(\mathbf{s}) + (\boldsymbol{\nu} - \mathbf{s})\mathbf{J}(\mathbf{s}) \quad (4.22)$$

$$\begin{aligned} \beta(\mathbf{s}) &= -\hat{g}_s(\boldsymbol{\nu})/T = -\boldsymbol{\nu}[\mathbf{f}(\mathbf{s}) + (\boldsymbol{\nu} - \mathbf{s})\mathbf{J}(\mathbf{s})]^T/T \\ &= -\text{tr}\mathbf{J}(\mathbf{s})/T, \end{aligned} \quad (4.23)$$

where  $\mathbf{J}(\mathbf{s})$  is the Jacobian of the source estimate (Luenberger, 1969). The last step follows from the fact that the elements of  $\boldsymbol{\nu}$  are mutually uncorrelated and have zero mean and unit variance. If the denoising is instantaneous, i.e.  $\mathbf{f}(\mathbf{s}) = [f(s(1)) \cdots f(s(t)) \cdots \cdots f(s(T))]$ , the shift can be written as  $\beta = -\sum_t f'(s(t))/T$ . This is the spectral shift used in FastICA (Hyvärinen, 1999a). There, it has been justified as an approximation to Newton's method and DSS thus provides a novel interpretation.

In general, iterations converge faster with the FastICA-type spectral shift (4.23) than with the fixed shift (4.21) but the fixed shift has the benefit that no gradients need to be computed. This is important when the denoising is defined by a complex nonlinear procedure, such as median filtering.

Another well known example where the spectral shift by the eigenvalue of a Gaussian signal is useful is the mixture of both super- and sub-Gaussian distributions. A DSS algorithm designed for super-Gaussian distributions would lead to  $\lambda > \lambda_G$  for super-Gaussian and  $\lambda < \lambda_G$  for sub-Gaussian distributions,  $\lambda_G$  being the eigenvalue of the Gaussian signal. By shifting the eigenvalue spectrum by  $-\lambda_G$ , the most non-Gaussian distributions will result in the largest absolute eigenvalues regardless of whether the distribution is super- or sub-Gaussian. By using the spectral shift it is therefore possible to extract both super- and sub-Gaussian distributions with a denoising scheme which is designed for one type of distributions only.

Consider for instance  $\mathbf{f}(\mathbf{s}) = \tanh \mathbf{s}$  (notice the connection to Eq. (3.21)) which can be used as denoising for sub-Gaussian signals, while  $\mathbf{s} - \tanh \mathbf{s} = -(\tanh \mathbf{s} - \mathbf{s})$  (a shrinkage function, see also Fig. 3.5) is a suitable denoising for super-Gaussian signals. This shows that depending on the choice of  $\beta$ , DSS can find either sub-Gaussian ( $\beta = 0$ ) or super-Gaussian ( $\beta = -1$ ) sources. With the FastICA spectral shift (4.23),  $\beta$  will always lie in the range  $-1 < \beta \leq \tanh^2 1 - 1 \approx -0.42$ . In general,  $\beta$  will be closer to  $-1$  for super-Gaussian sources which shows that FastICA is able to adapt its spectral shift to the source distribution.

None of the above methods always work for nonlinear DSS. Sometimes the spectral shift turns out to be either too modest or strong, leading to slow convergence or lack of it, respectively. For this reason, we suggest a simple stabilisation rule: instead of updating  $\mathbf{w}$  into  $\mathbf{w}_{\text{new}}$  defined by (4.10), it is updated into

$$\mathbf{w}_{\text{adapted}} = \text{orth}(\mathbf{w} + \gamma \Delta \mathbf{w}) \quad (4.24)$$

$$\Delta \mathbf{w} = \mathbf{w}_{\text{new}} - \mathbf{w}, \quad (4.25)$$

where  $\gamma$  is the step size. We propose to start the iteration with  $\gamma = 1$ , but if the consecutive steps are taken in nearly opposite directions, i.e. the angle between  $\Delta \mathbf{w}$  and  $\Delta \mathbf{w}_{\text{old}}$  is greater than  $179^\circ$ , then  $\gamma = 0.5$  for the rest of the iterations. We call this the *179-rule* for adapting the learning rate. A stabilised FastICA has been proposed by Hyvärinen (1999a) as well and a procedure similar to the one above has been used.

The above modification is able to stabilise convergence in case of oscillations but sometimes the spectral shift is too small. Then an increase in step size would be appropriate, i.e.  $\gamma > 1$ . We propose a simple rule for adapting  $\gamma$  which is inspired by predictive controllers used in robotics: a simple, but slow and possibly unstable, reactive controller is used to teach a new, predictive controller. Usually stable and rapid convergence are difficult to achieve simultaneously, but in this setup the new controller can be both faster and more stable.

Translated to our problem, the old slow and unstable controller is the weight modification rule which proposes a modification of weight according to (4.25). The new controller is implemented by (4.24), i.e. it modifies the step size. The new controller tries to do immediately what the old controller would do in the future. The step at the previous time instant was apparently optimal if the step proposed at this time instant is orthogonal with it. If not,  $\gamma$  should have been different and, assuming that the optimal  $\gamma$  is constant, the gamma used at this time step should be

$$\gamma_{\text{new}} = \gamma_{\text{old}} + \Delta \mathbf{w}_{\text{old}}^T \Delta \mathbf{w} / \|\Delta \mathbf{w}_{\text{old}}\|^2. \quad (4.26)$$

As it does not seem productive to take steps in the direction opposite from what is suggested by  $\Delta \mathbf{w}$  or to take extremely short steps, we require that  $\gamma \geq 0.5$ .

The above adaptation of  $\gamma$  has turned out to be very useful and it can both stabilise and accelerate convergence. According to (4.26),  $\gamma$  keeps increasing as long as the steps are taken to the same direction and decreases if they are taken backwards.

## 4.6 Separation of artificial signals: comparison of DSS algorithms

In this section, we demonstrate the separation capabilities of the DSS algorithms presented earlier. Artificial signals were mixed to compare different DSS schemes and JADE (Cardoso, 1999, Sec. 3.2.3). Ten mixtures of five artificially generated sources were produced and independent white noise was added with different SNRs ranging from nearly noiseless mixtures of 50dB to -10dB, a very noisy case. The original sources and the mixtures are shown in Figs. 4.5a and 4.5b respectively. The mixtures shown have SNR of 50 dB.

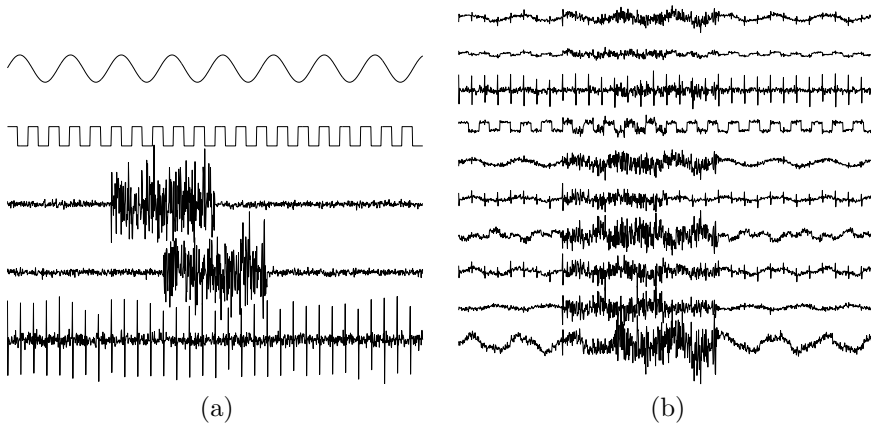


Figure 4.5: (a) Five artificial signals with simple frequency content (signals #1 and #2), simple on/off non-stationarity in time domain (signals #3 and #4) and quasi-periodicity (signal #5). (b) Ten mixtures of the signals in (a). (from Publication 5)

### 4.6.1 Linear denoising

In this section, we show how the simple linear denoising schemes described in Sec. 4.4.1 can be used to separate the artificial sources. These schemes require prior knowledge about the source characteristics.

The base frequencies of the first two signals were assumed to be known. Thus two band-pass filtering masks were constructed around these base frequencies. The third and fourth source estimates were known to have periods of activity and non-activity. Third was known to be active in the second quadrant and the fourth a definite period in the latter half. They were denoised using binary masks

in time domain. Finally, the fifth source had a known quasi-periodic repetition rate and was denoised using the averaging procedure described in Sec. 4.4.1 and Fig. 4.4. Since all of the five denoisings are linear, five separate filtered data sets were produced and PCA was used to recover the principal components. The separation results are described in Fig. 4.6 together with the results of other DSS schemes and JADE.

### 4.6.2 Nonlinear exploratory denoising

In this section, we describe an exploratory source separation of the artificial signals. The present author gave the mixtures to the other author of Publication 5 whose task was to separate the original signals. The author did not receive any additional information, so he was forced to apply a blind approach. He chose to use the masking procedure based on the improvements on the kurtosis-based ICA (see Sec. 4.4.2; for details of the improvements, consult Publication 5). To enable the separation of both sub- and super-Gaussian sources in the ICA-based denoising, he used the spectral shift (4.21). To ensure convergence, he used the 179-rule to control the step size  $\gamma$  (4.24).

Based on the separation results of the ICA-based DSS, he further devised specific masks for each of the sources. He chose to denoise the first source in frequency domain with a strict band-pass filter around the main frequency. The author decided to denoise the second source by a simple denoising function  $\mathbf{f}(\mathbf{s}) = \text{sign}(\mathbf{s})$ . This makes quite an accurate signal model though it neglects the behaviour of the source in time. The third and fourth signal seemed to have periods of activity and non-activity. He found an estimate for the active periods by inspecting the instantaneous variance estimates  $\mathbf{s}^2$ , and devised simple binary masks. The last signal seemed to consist of alternating positive and negative peaks with fixed inter-peak-interval as well as some additive Gaussian noise. The signal model was tuned to model the peaks only.

### 4.6.3 Separation results

In this section, we compare the separation results of the linear denoising (Sec. 4.6.1), improved ICA-based denoising and adapted denoising (Sec 4.6.2) to other DSS algorithms. In particular, we compare to the popular denoising schemes  $\mathbf{f}(\mathbf{s}) = \mathbf{s}^3 - 3\mathbf{s}$  and  $\mathbf{f}(\mathbf{s}) = \tanh(\mathbf{s})$ , suggested for use with FastICA (1998). We compare to JADE (Cardoso, 1999) as well. During sphering in JADE, the number of dimensions was either reduced ( $n = 5$ ) or all of the ten dimensions were kept ( $n = 10$ ).

We restrained from using deflation in all of the different DSS schemes to avoid suffering from cumulative errors in separation of the first sources. Instead one source was extracted with each of the masks several times using different initial

vector  $\mathbf{w}$  until five sufficiently different source estimates were reached (see Himberg and Hyvärinen, 2003, Meinecke et al., 2002, for further possibilities along these lines). Deflation was only used if no estimate could be found for all of the five sources. This was often the case for poor SNR under 0dB.

To get some idea of statistical significance of the results, each algorithm was used to separate the sources ten times with the same mixtures, but with different measurement noises. The average SNRs of the sources are depicted in Fig. 4.6. The straight line above all of the DSS schemes represents the optimal separation. It is achieved by calculating the demixing matrix explicitly using the true sources.

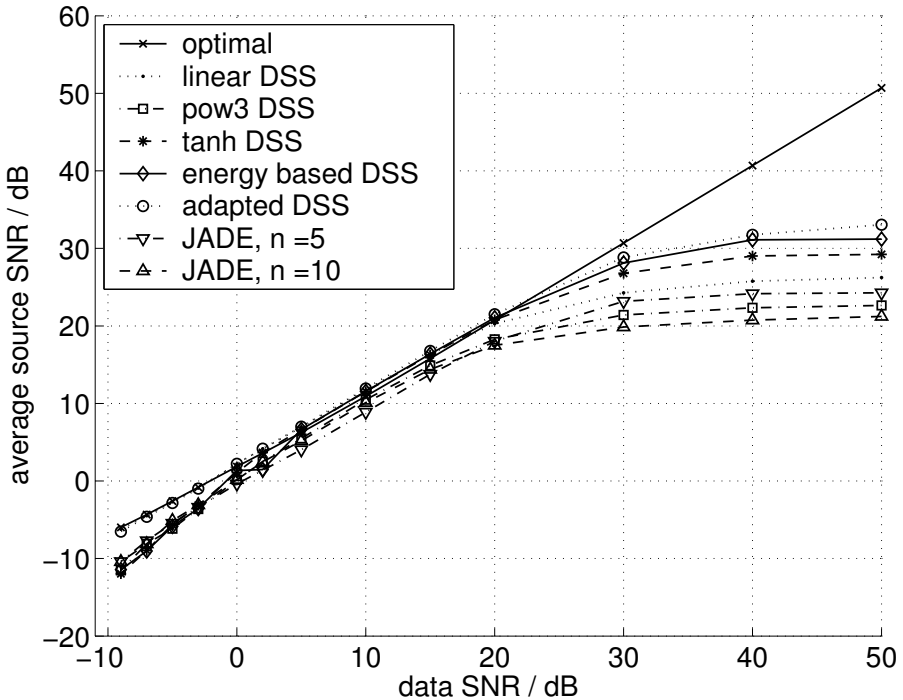


Figure 4.6: Average SNRs for the estimated sources averaged over 10 runs. (from Publication 5)

With outstanding SNR ( $> 20$  dB), linear DSS together with JADE and kurtosis-based DSS seem to perform worst, while the other, nonlinear DSS approaches: tanh-based, improved ICA-based and the adapted one seem to perform better. The gap between these groups is more than two standard deviations of the 10

runs, making the difference statistically significant. In practice, the difference in performance probably does not matter.

With moderate SNRs (between 0 and 20 dB), all algorithms perform quite alike and quite optimally, too. With poor SNR ( $< 0$  dB), the upper group consist of the linear and adapted DSS as well as the optimal one and the lower group consists of the blind approaches. This seems reasonable, since it makes sense to rely more on prior knowledge when the data is very noisy.

## Chapter 5

# Overfitting

*The philosopher, Samuel Alexander, was hard of hearing in his old age, and used an ear trumpet. One day a colleague came up to him in the common room at Manchester University, and attempted to introduce a visiting American philosopher to him. "THIS IS PROFESSOR JONES, FROM AMERICA!" he bellowed to the ear trumpet. "Yes, Yes, Jones, from America," echoed Alexander, smiling. "HE'S PROFESSOR OF BUSINESS ETHICS!" continued the colleague. "What?" "PROFESSOR OF BUSINESS ETHICS!" Alexander shook his head and gave up: "Sorry. I can't get it. Sounds like 'business ethics'!"*

–Dennett (1994)

In the previous chapters, we discussed several source separation algorithms. Often, the convergence analysis of these algorithms is carried out under idealised assumptions. For example, it is usually assumed that there exist an infinite set of samples. Naturally, this is never the case in practice. In fact, we often face a source separation problem with rather limited amount of data. How do these algorithms perform under this more realistic setting?

It is classically known that the linear regression problem (Luenberger, 1969) with equal amount of samples and model parameters results in a serious failure. The fitted curve travels exactly through the data points, achieving a zero mean-squared error. But it may be totally useless for interpolation between the data points and for extrapolation outside the data range. This problem is called the problem of *overfitting* (or overlearning). In a more general optimisation problem, overfitting means that the optimisation result depends less and less on the data and is almost totally determined by the optimisation criterion.

As an example, consider the noiseless linear source separation model with equal amount of samples and mixtures that is  $T = M$ . Then, all of the matrices in the



equation  $\mathbf{X} = \mathbf{A}\mathbf{S}$  are square. Now, by changing the values of  $\mathbf{A}$ , we can give any values for  $\mathbf{S}$ . In this case, the source separation criterion totally determines the resulting estimate of the sources. In other words, whatever the data  $\mathbf{X}$  and the original sources might be, we attain the same source estimates  $\mathbf{S}$ .

In general, there are two reasons for a model to fail in modelling data: 1) the model is not estimated properly, even if it is a probable model for the data; 2) the model is not adequate for the phenomenon causing the data. Both cases have been called overfitting. However, the measures to be taken to solve the overfitting problem differ drastically in the two cases: in the case of the first type of overfitting, more proper ways to estimate the model should be searched for, whilst in the second case the model itself should be affected.

In Sec. 5.1, we first discuss the problem of overfitting in the case of ICA algorithms based on the marginal-distribution information of the sources. Some solutions in the scope of noiseless ICA algorithms are presented in Sec. 5.2. In Sec. 5.3, we broaden the scope by considering the overfitting in a Bayesian marginal-distribution-based ICA. In this section the main emphasis is to find out whether the overfitting problems at hand are of the first type or of the second type. In other words, is it enough to consider a better founded estimate for the model or is it necessary to consider some refinements in the model structure itself. Finally, in Sec. 5.5, the overfitting in DSS is considered.

The overfitting problem in ICA was first reported in Publication 2. A thorough coverage of the problem appeared in Publication 3 and was extended in a Bayesian framework by Särelä and Vigário (2003).

## 5.1 Overfitting in marginal-distribution-based ICA

Consider the FastICA algorithm, using the absolute value of kurtosis as its contrast function (Hyvärinen, 1999a). It has been proven by Hyvärinen that the maximum absolute kurtosis is achieved by signals that have a single spike and are almost zero everywhere else (Publication 3, Appendix A). Thus, in the extreme case where there are only as many samples  $T$  as dimensions in the data  $M$ , the estimate for the source matrix  $\mathbf{S}$  is roughly a permutation matrix, multiplied by  $\sqrt{T}$ .

As an illustration, consider the sources shown in Fig. 5.1a, having 500 samples each. Let 500 mixtures be generated of the sources resulting in a data matrix of size  $500 \times 500$ . Very little noise is also added, to make the covariance matrix of the data to have full rank. From now on, this data set is called the *artificial data set*. Three representative mixtures are shown in Fig. 5.1b. An application of FastICA using the absolute value of kurtosis gives source estimates (Fig. 5.1c) that are utterly different from the original ones, but contain only a single spike, as expected. Similar results have been reported for Bell-Sejnowski algorithm

in Publication 2 and are expected for other ones too.

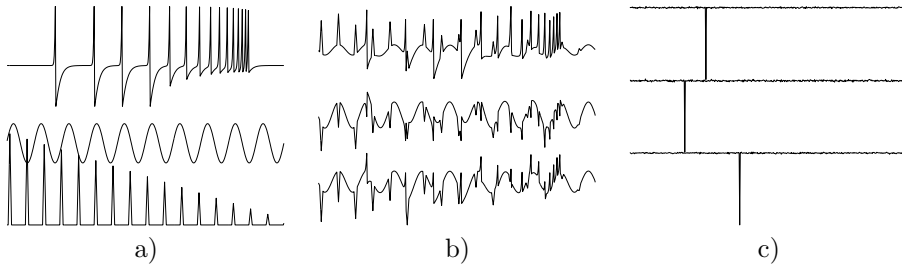


Figure 5.1: a) *Three original sources.* b) *Three representative mixtures of the sources.* c) *Three separated components using FastICA.* (from Publication 3)

### 5.1.1 Are we saved if $T > M$ ?

It does not sound very dangerous to suffer from overfitting when there are only as many samples as there are dimensions. Does the problem vanish, if there exist more? In parameter estimation (e.g. linear regression), a rule of thumb states that one needs, at least, a number of samples equal to ten times the number of free parameters. In the case of ICA, there is a need to estimate the demixing matrix  $\mathbf{W}$ . Assuming presphered data, the number of free parameters is roughly  $N^2/2$ , where  $N$  is the number of sources<sup>1</sup>. Thus, there should exist  $T > 5 \times N^2$  samples.

CLT guarantees that a sum of  $N$  independent variables is more Gaussian than the most non-Gaussian original variable (see e.g. Papoulis, 1991). Thus, it is guaranteed that e.g. the kurtosis (or the absolute value of kurtosis, if you wish) of any linear projection  $\mathbf{s}_i = \mathbf{w}_i^T \mathbf{Y}$  is at most equal to the kurtosis of the most kurtotic independent source<sup>2</sup>. In some sense, this result precludes the emergence of the type of overfitting mentioned earlier. Yet, because infinite realisations of the measurements do not exist, it is conceivable to attain higher values of kurtosis than those of the most kurtotic source, as will be shown next.

Consider a 50 dimensional Gaussian i.i.d data with 12500 samples. Since every channel is Gaussian and independent, every projection, according to the CLT, should be Gaussian. However, it is easy to produce a spike by enhancing one time instance and dampening all others: e.g. by using one time instance as the demixing vector, i.e.  $\mathbf{s}_i = \mathbf{y}^T(t_0)\mathbf{Y}$ , where  $\mathbf{y}(t_0) = [y_1(t_0) \cdots y_i(t_0) \cdots y_M(t_0)]^T$

<sup>1</sup>If  $M > N$ , the spare dimensions can be removed during sphering.

<sup>2</sup> $\mathbf{Y}$  is used for denoting the data to emphasise the fact that the data is uncorrelated already.

corresponds to the data at  $t_0$ . Some spikes generated in this way are shown in Fig 5.2a. Their positive sample kurtoses, ranging between 0.53 and 0.70, make them appear significantly non-Gaussian. FastICA produces similar spikes with comparable kurtoses, which can be seen from Fig. 5.2b.

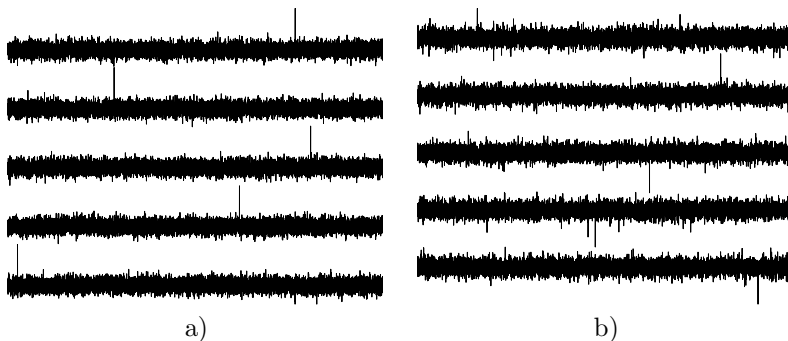


Figure 5.2: *a) Artificially generated spikes from the Gaussian i.i.d. data. b) FastICA estimates for the Gaussian i.i.d. data. (from Publication 3)*

Hence, in the case of finite data sets, the maximisation of measures such as the absolute value of the kurtosis, may result in the generation of spiky components. This is the case especially when the spiky components have greater kurtoses than the actual independent components.

### 5.1.2 Bumps emerge when low frequencies dominate

Consider another data set, having independent channels, but this time with significant dependencies between the samples, i.e.  $p(y(t_1), y(t_2)) \neq p(y(t_1))p(y(t_2))$  for some pairs  $t_1$  and  $t_2$ . In particular, let the power spectra of the components follow a  $1/f$ -curve. This kind of behaviour is faced in many natural data as argued by Bak et al. (1988). Henceforth, this data set is called  $1/f$  data.

Again, the marginal distributions of all channels are Gaussian and thus every projection should, in theory, be Gaussian. Nonetheless, because we are still in presence of finite data samples, we can try to generate spikes using the same strategy as before. Because nearby samples in time are strongly correlated, the result is not any more a single spike, but rather a bump. Figure 5.3a shows five such bumps. Spikes are still visible, but now they emerge in the middle of a small bump. The sample kurtoses of those signals are clearly non-zero (0.74, 0.96, 0.84, 0.65 and 0.92).

A stronger “non-Gaussian” effect can be produced by forcing the demixing

vector to be a weighted average around some time point. Then, the bump is

$$\begin{aligned} \mathbf{s} &= \mathbf{w}^T \mathbf{Y}, & \text{where} \\ \mathbf{w} &= \sum_{t=0}^L d(t) \mathbf{y}(t_0 - L/2 + t). \end{aligned} \quad (5.1)$$

$d(t)$  is the windowing function, normalised to  $\sum d(t) = 1$ .  $L + 1$  is the width of the window, which is, at best, also the width of the bump. Some bumps, using a triangular window of length  $L = 1001$ , are presented in Fig. 5.3b. Here the spike is completely absorbed by the bump and the kurtoses are even greater (1.15, 0.98, 1.48, 0.69 and 1.51). Again, FastICA extracts similar components as seen in Fig. 5.3c.

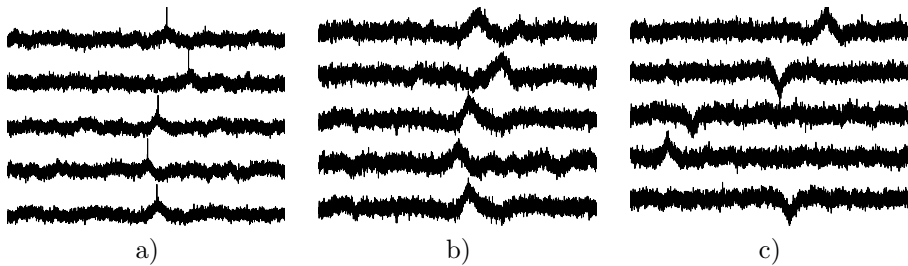


Figure 5.3: *a) Artificially generated bumps from the 1/f-data. b) Artificially generated smooth bumps. c) Components estimated by FastICA. (from Publication 3)*

### 5.1.3 Bumps are the overfitting in magnetoencephalograms

In this section, we show that the bump-type overfitting-problem arises in practice with real data. In particular, magnetoencephalograms (MEG) are used. The data set (Vigário et al., 1997b) has been initially analysed by Vigário et al. (1997a). Three found sources were selected as targets to see the effect of different solution proposals in Secs. 5.2 and 5.3. Magnification of them are shown in Fig. 5.4a. They are called *focus signals* for the rest of this chapter.

Due to its  $1/f$  nature, MEG data tends to show a bump-like overfit, rather than the spike type, observed for the Gaussian i.i.d. data. This fact can be observed in the last five independent component estimates shown in Fig. 5.4b.

Because of the considerable amount of data (over 12000 samples, for a total of 127 channel measurement), as well as the high values of kurtoses of several

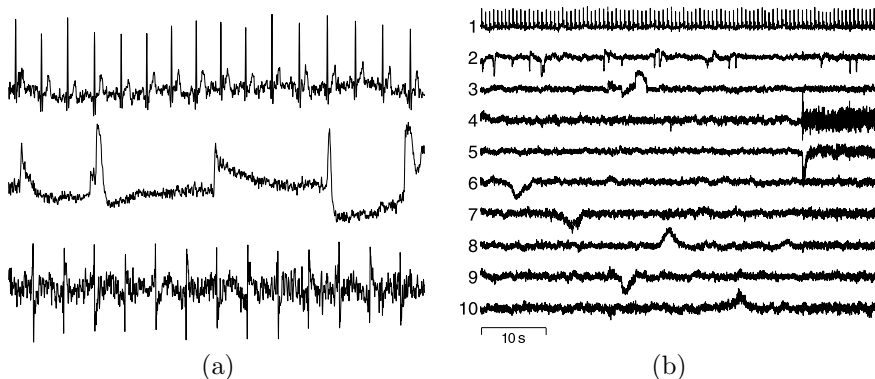


Figure 5.4: *a) Magnifications of the components IC4, IC5 and IC6 from Vigário et al. (1997a). b) Some ICA components, including both true underlying sources and bump-like overfits. (adapted from Publication 3)*

meaningful components, these seem to be allowed to coexist with the overfits (see the first five components in Fig. 5.4b). The kurtoses of the first five components range from 10.30 to 44.46, whereas the last five range from 9.37 to 14.10.

## 5.2 Attempts to solve the problems in ICA

In this section, we present a summary of the solutions proposed for ICA in Publication 3.

### 5.2.1 Proper estimate of the ICA model

If the ICA model is assumed to be a probable one, and the overfitting observed to stem from a modelling failure of the first type, a solution may be attempted by increasing the number of samples per free parameter to be estimated with our ICA algorithm. This can be achieved either by increasing the number of samples, or by reducing the dimensionality of the data set.

We show in Publication 3 that the kurtosis of a spike is inversely proportional to the sample size  $T$ , and directly proportional to the square of the dimensions  $M$ :  $\text{kurt}(\mathbf{s}) \propto M^2/T$ . Hence, decreasing the number of dimensions should be a more efficient way to reduce the overfitting effects than increasing the number of samples. This is illustrated in Fig. 5.5, where results of separation of the components using the *artificial data set* are shown. In particular, the correlations

to the original sources are shown as a) a function of the sample size and b) the compressed dimensionality.

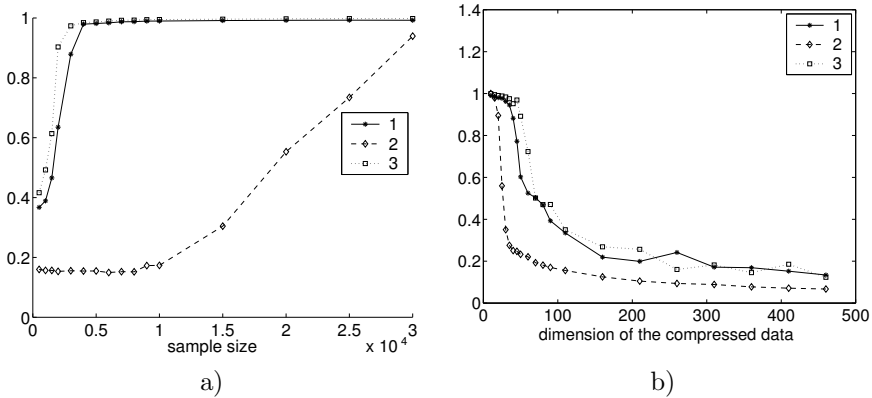


Figure 5.5: a) Correlations between the original signals and the ICA correspond-ing estimates, as a function of the sample size. b) same, as a function of the compressed dimensions. The order of the components is in concordance with Fig. 5.1a. (adapted from Publication 3)

Nevertheless, those considerations assume that there is a strong redundancy between channels, i.e. there exist significantly more sensors than sources, which is not always verified. If this is not the case, dimension reduction cannot be used in principle. In practice, it may be used and hoped that the components found, sums of the underlying independent sources, are still somehow meaningful.

We observed as well that an increase of the sample size or a reduction of the dimension does not yield as clear results in avoiding bumps as when avoiding spikes. This suggests that the problem of bumps may arise also from the inadequacy of the model and not only from the fact the model is not estimated properly. More robust contrast functions suggested by Hyvärinen (1998b) were also tested but they didn't seem to improve the results considerably.

### 5.2.2 Additions to the model

In a model failure of the second type, a solution may lie in additional modelling of the data. It may be possible to divide, linearly, the observations  $\mathbf{X}$  into two terms,  $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ , in such a way that only  $\mathbf{X}_1$  is prone to a bump type of overfitting. Then it should be possible to estimate the original mixing matrix  $\mathbf{A}$  from the relation  $\mathbf{X}_2 = \mathbf{A}\mathbf{S}_2$ . Similarly,  $\mathbf{S}_2$  corresponds now to the portion of

the underlying sources, independent and non-Gaussian, which are associated with  $\mathbf{X}_2$ . This should solve the problem of bumps. This strategy resembles the sensor-noise ICA model considered in Sec. 3 when the noise has the covariance (3.2).

Because bumps are mainly dominated by low frequencies, one such division may consist simply of a high-pass filtering of the data, prior to the application of the ICA algorithm. In experiments, it was observed that the width of the bumps increases with the number of samples. Thus the determination of a fixed cut-off frequency for all the data is not suitable.

A more elegant approach can be derived using an auto-regressive (AR) model to account for the low frequencies in the data:

$$\mathbf{x}(t) = \sum_{\tau=1}^T \mathbf{c}_\tau^T \mathbf{x}(t - \tau) + \mathbf{x}_2(t).$$

The sum corresponds to the AR-process and  $\mathbf{x}_2(t)$  is usually referred to as the *innovation process*. After removing the AR-process from the data, ICA is applied to the residual innovation process  $\mathbf{X}_2$ .

To test this technique with the MEG artefacts data, we selected to model the low frequencies with a one-tap AR process:  $\mathbf{x}(t) = c_1 \mathbf{x}(t - 1) + \mathbf{x}_2(t)$ . The AR-coefficient was estimated from an MEG data using ML resulting in  $c_1 = 0.9$ . This comes close to basic random walking or Brownian movement. The components that best estimate the focus signals are shown in Fig. 5.6. All artefacts are perfectly recovered, better so than with simple high-pass filtering.

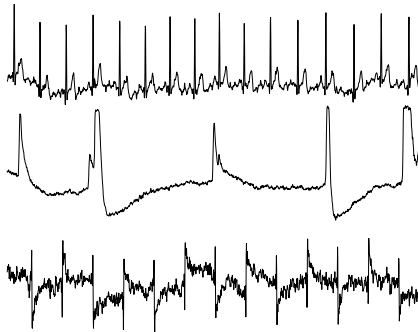


Figure 5.6: *Results with AR-process filtering. (from Publication 3)*

## 5.3 Bayesian analysis of the problems of spikes and bumps

In this section, we discuss the origin of the two overfitting problems encountered in ICA: the spikes and the bumps. The analysis is carried out in Bayesian framework, in particular using analytic EL-approach where a simple factorial approximation is fitted in the true posterior distribution. This approach, together with some ICA implementations were discussed in Secs. 2.3.5 and 3.2.5.

Using the so called bits-back argument, Hinton and van Camp (1993) linked the lower bound of the evidence given by EL (2.26) to the principle of minimum description length (Rissanen, 1978)<sup>3</sup>. Thus, intuitively, EL avoids overfitting due to the fact that overfits, in general, do not provide compact codes for the data. Instead, a big amount of components is needed to fully explain the data. The Bayesian approach has also other assets in avoiding overfitting. For instance, instead of considering one single model, the results are averaged over several models. Therefore it is not sensitive to narrow though high peaks in the posterior. Moreover, EL favours simple models and has an explicit modelling of the noise.

In the experiments below, we use BICA and FBICA (see Sec 3.2.5).

### 5.3.1 Avoiding spikes

The capability of BICA to avoid spikes was tested using the artificial data set. Small amount of additive noise with variance  $\sigma^2 = 0.01$  was added to the mixtures. 10 components were learned from this data. The means of four learned components are shown in Fig. 5.7a. BICA was able to separate the three original independent components. BICA also estimated correctly the number of independent components, by pruning away the remaining 7 components (one of them is depicted in the figure), despite the fact that the three components do not explain all of the data. The unexplained part is Gaussian i.i.d and is thus considered as noise.

The noise estimation capability was tested using the Gaussian i.i.d data set. The smallest value of the cost function  $C_{KL}$ , i.e. the most probable model was quite correctly achieved, when there were no sources but all of the data was considered to be noise. In some cases, EL got stuck in local minima and all of the sources were not killed. Even in those cases the sources were Gaussian and did not contain any spikes.

---

<sup>3</sup>A closely related subject is minimum message length (Wallace and Freeman, 1987).



### 5.3.2 Reducing the effects of bumps

Whether the problem of bumps is of the same origin as the spikes was tested using the  $1/f$  data set, with  $M = 30$  and  $T = 5 \times M^2 = 4500$ . BICA model was compared to modelling the data as Gaussian i.i.d with no sources (null-model). The most probable ICA model contained 5 sources and was approximately  $10^{1500}$  times more probable than the null-model. The means of the sources are pictured in Fig. 5.7b. The results do not have a single bump, but rather several. When the number of sources approaches the number of mixtures, more and more clear bumps emerge. Even if the ICA model with five sources was superior to the 0-model, it does not make it a good model. A model where each channel was modelled using a random walk was more than  $10^{2688}$  times more probable.

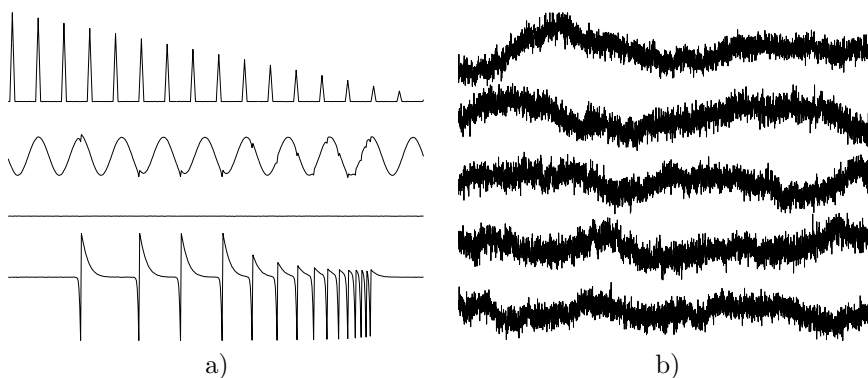


Figure 5.7: a) Four of the 10 estimated independent components by Bayesian ICA in the artificial data. b) ICA estimates from the  $1/f$ -data. (adapted from Särelä and Vigário, 2003)

The performance of FBICA was tested using the MEG data set. The best correspondences to the focus signals are depicted in Fig. 5.8a. FBICA was successful in extracting the first two focus signals. Unfortunately, the third one still contains a bump.

As a final experiment, FBICA was combined with the noise cancelling procedure of ML-based AR filtering described in Sec. 5.2.2. The results are gathered in Fig. 5.8b and show a clear extraction of the cardiac and watch artefacts. Because of the rather slow behaviour of the blink artefact, most of it is erased by the AR-process. That is the reason why the blink artefact is not that well extracted.

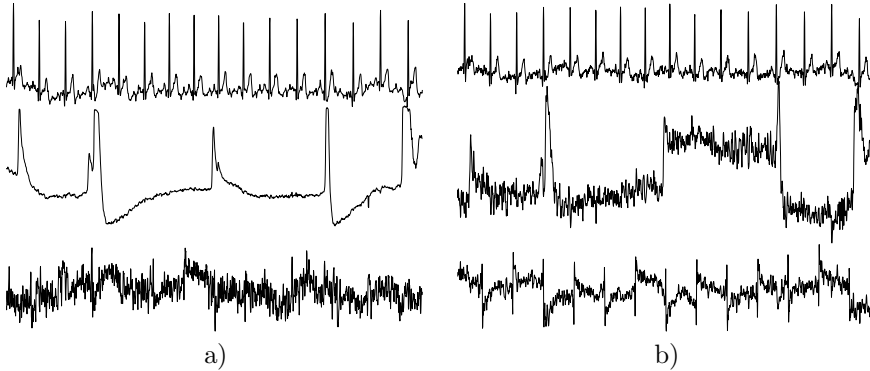


Figure 5.8: a) *Estimated independent components by fast Bayesian ICA in the MEG artefacts data. The first 2000 points of the best correspondences to the focus signals are shown.* b) *AR + FBICA.* (adapted from Särelä and Vigário, 2003)

## 5.4 Conclusions on overfitting in marginal-distribution-based ICA

Bayesian ICA was able to solve the problem of spikes. This is a clear indication that, in the ICA algorithms relying on higher-order statistics, the problem of spikes is originated from the crude estimation of the ICA model, probably due to the coarse approximation of the negentropy. However, in the cases where there existed strong dependencies between the data samples, BICA still produced bumps or bump-like structures. This hints that the ICA model *as such* may not be sufficiently adequate for MEG data nor for any other data with  $1/f$  frequency characteristic. Moreover, the overfitting seems to be of the second type. This suggests that in case of data having strong dependencies between samples, one would benefit from the combination of marginal-distribution-based ICA and time-information-based BSS methods. Note that we were able to recover all of the focus signals without the usage of the Bayesian methods, as well. Furthermore, it is our belief that with suitable preprocessing, the ML- and MAP-based methods can also cope with overfitting more generally. The main aim of using the Bayesian methods here, was to be able to analyse the origin of the two overfitting problems.

## 5.5 Overfitting in DSS

We have shown in Sec. 4 that the marginal-distribution information is effectively combined with the time information in a source separation algorithm under the DSS framework. Thus, we expect DSS to be less prone to overfitting when used properly. However, it is still possible that DSS extracts structures that are not actually present in the data but are generated by the denoising function, i.e. the results may be due to overfitting.

In DSS, the outlook of the overfitted results naturally depends on the denoising criterion and may not resemble spikes or bumps as is typical in ICA. To detect an overfitted result, one should know how it looks like. As a first approximation, DSS can be performed with same amount of i.i.d Gaussian data. Then all of the results present overfits, typical to the denoising function used. Even better characterisation of the overfitting results can be obtained by mimicking the actual data as well as possible. In that case it is important to make sure that the structure assumed by the signal model has been broken. In MEG data, one possibility to realise this is to use  $1/f$ -data with no other structure.

Note that in addition to visual test, the methods described above provide us with a quantitative measure as well. Using the objective function or its approximation (4.13), we can set a threshold under which the sources are very likely overfits and do not carry much real structure. In the case of linear DSS, the value of the objective function is given simply by the corresponding eigenvalue.

We have reported the use of the above-mentioned overfitting tests in Publication 5 where they have been shown to be valuable. Furthermore, we suggest that all of one's source-separation results should be subjected to overfitting tests to investigate their reliability. Additional information on the stability of ICA results can be provided by bootstrapping methods (see Himberg and Hyvärinen, 2003, Meinecke et al., 2002).

## Chapter 6

# Biomedical systems

In this part of the thesis, we apply the exploratory-source-separation methods, developed and reviewed in the first part, to the study of biomedical systems. Biomedical systems arguably form the most complex systems the human kind has ever explored. Living organisms, such as humans, consist of several biomedical subsystems. In humans, there exist the central nervous system (CNS), the heart, and the lungs, just to name a few. Scientific community has devoted serious efforts to the understanding of these systems for centuries, even for millennia. In this thesis, we concentrate on studying CNS, though the methods described in the first part have been successfully applied to other subsystems as well.

This chapter is organised as follows: First, several brain imaging techniques are reviewed in Sec. 6.1. MEG is reviewed in more detail since it is the central imaging technique discussed in this thesis. Then, in Sec. 6.2, we show how the ESS framework, especially the adaptation of the source separation algorithms, can be used in extraction of sources in MEG.

## 6.1 Brain imaging techniques

In this section, we briefly review some brain imaging techniques and discuss MEG in more detail. For more information, see Bankman (2000), Gazzaniga (2000), Robb (2000).

### 6.1.1 Early techniques

Already during the eighteenth century, Luigi Galvani became aware that *electrical nervous stimulation* in a frog caused movement of limbs. During the nineteenth century, many similar studies were done on animal subjects. The most important

study was perhaps the study by Fritsch and Hitzig in the year 1870: by stimulating electrically some brain regions of a dog, they were able to produce muscle activity on the contralateral side to the stimulation.

*Lesion studies* and *post-mortem examination* have brought an enormous amount of information on the functioning of the human brain. Autopsy-studies of people with brain-functional disorders have given us knowledge about brain parts essential to specific functions. Already as early as 1861, Broca was able to show that a lesion located in an area in the left hemisphere, later called the Broca's area, leads to speaking disorder called motoric aphasia.

### 6.1.2 Modern techniques

The twentieth century brought many new techniques for investigating the anatomy and functioning of the human and animal brains. Since ancient times, physicians have used stethoscopes to listen to the breathing and the heart pumping. This has been generalised to phonocardiogram where an array of microphones are used to listen to the heart sounds.

*Electroencephalogram* (EEG) is nowadays perhaps the most used technique in clinical diagnostics. The earliest EEG studies are already from the end of the nineteenth century. In EEG, electric potentials are measured by electrodes on the scalp. Since activation potential of a single neuron is extremely weak, only simultaneous activations of many neurons can be measured. Furthermore, the neuronal activity has to be synchronous, otherwise differently phased activities cancel each other.

*Magnetoencephalogram* (MEG) is a newer technique, in many ways similar to EEG. It measures the magnetic fields on the scalp caused by the electric synchronous activity in the cortex. The next section covers MEG in more detail.

Another measuring technique is *one-cell recording*, where electric potential of a single cell is measured by an electrode on the head of a needle. Since insertion of the needle always kills some neurons, even thousands, this technique has been used much in animal studies, but not so much in humans.

Mostly caused by the enormous progress of physics in last century, many sophisticated measuring techniques have been invented during the latter half of the last century. One of these techniques is the *computer (axial) tomogram* (CT) which combines multiple X-ray shots to scan the human brain.

In another modern technique, *positron emission tomogram* (PET), some radioactive marker, usually oxygen  $^{15}\text{O}$  is injected into the subject's veins. Positrons, which are emitted when  $^{15}\text{O}$  changes to the stable  $^{16}\text{O}$ , annihilate with electrons producing two gamma rays. The gamma rays are measured, giving information about the oxygen consumption in the brain, hence the functional assessment of brain regions.

*Magnetic resonance imaging* (MRI) is a brain imaging technique that takes advantage of nuclear magnetic resonance: Each nucleus has a spin or a momentum. Normally the alignment of the spins in a substance is random. But if the substance is placed in an external magnetic field, the directions of the spins of the nuclei align to it. If the external magnetic field is cancelled, the spins start to return to their normal positions. This relaxation induces gamma radiation, which can be measured, producing anatomical knowledge of different tissues in the brain. The same technology can be used for functional studies resulting in functional MRI (fMRI). Then the blood flow in the brain can be monitored, giving information on the activity distribution in the brain.

One of the newest brain imaging techniques is the *transcranial magnetic stimulation*. There, the cortex is stimulated by strong magnetic field pulses, which induce a post synaptic flow of current leading to the excitation of neurons.

### 6.1.3 Basics of magnetoencephalograms

The following introduction to MEG is heavily based on Hämäläinen et al. (1993).

It is believed that the neural information processing is mainly based on the neurons sending electrical neuroimpulses to other neurons. The neuroimpulses are called action potentials (AP) and they cause post-synaptic currents (PSC) when they arrive at their target neurons. AP cause changing electric potentials, which are measured in EEG. Moving charges cause also magnetic fields. In MEG, the flux due to these fields is measured on the scalp level. However, the magnetic field of an AP of a single neuron is very weak and lasts only about 1 ms. On the other hand, PSC, though weaker in amplitude, lasts for several tens of milliseconds. Hence, if neuron populations fire synchronously, the net PSC caused by them is far greater than the corresponding net AP. This synchronous PSC is measured in MEG.

However, even the magnetic fields caused by net PSCs is several order of magnitudes weaker than the earth's magnetic field. For this reason, MEG devices are placed in magnetically shielded rooms. Furthermore, delicate instruments called superconducting-quantum-interference devices (SQUIDs, Zimmerman et al., 1970) are used to measure the flux. Nowadays it is possible to combine even hundreds of SQUIDs in a same measuring device (c.f. Vectorview<sub>TM</sub>, a 306-channel device). Figure 6.1 presents a simplified picture of one such whole-head MEG measuring device. The device is seen on the left. On the upper-right, one of the sensors is depicted.

MEG is a completely non-invasive brain imaging technique. Traditional definition of non-invasiveness is that the measuring process does not require any surgery. MEG is even safer: it does not need, either, any marking substances (radioactive or other as PET does), nor exposes the studied subject to any radia-

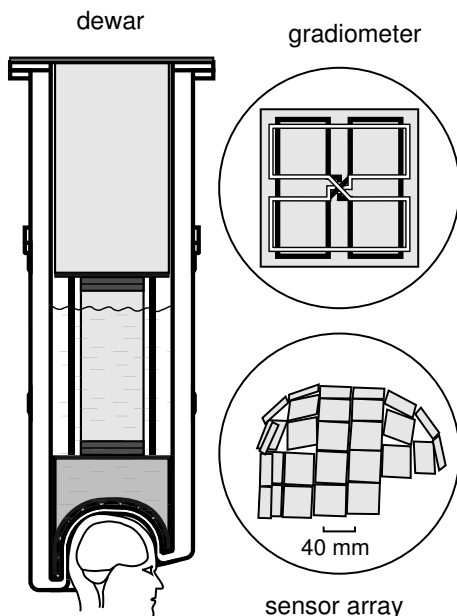


Figure 6.1: A simplified picture of an MEG system using gradiometers. (adapted from Hämäläinen et al., 1993)

tion (like X-rays as CT does) or to strong magnetic fields (as in MRI and fMRI). It only *measures* magnetic flux passing through the scalp. That makes it very safe and useful in clinical medicine as well as in brain imaging research.

MEG data can be acquired arguably fast enough to capture all functionally important neuronal activity in the brain. The spatial precision is quite good too, less than  $1\text{cm}^3$  in favourable circumstances. Spatial resolution, i.e. how close parallel sources are distinguishable, is close to 3cm.

There are two common tasks in electromagnetic field analysis. One is to calculate what is the field contribution on the scalp of electromagnetic sources inside the head. This is called the *forward problem* and can be solved rather precisely, when the electromagnetic field inside the head is known. The other task is the reverse task, i.e. to calculate the electromagnetic field distribution inside the head based on the flow through the scalp. This is called the *inverse problem* or localisation of the sources. Unfortunately, there are always an infinite amount of different field distributions that result in the same flow through the scalp. For this reason, additional assumptions have to be made in order to solve the inverse

problem. Several such assumptions have been proposed, such as the equivalent-current-multipole technique (ECM, Katila, 1983, ECD is used as an acronym for equivalent-current dipole when only one dipole is used), and minimum-norm-estimate technique (Hämäläinen and Ilmoniemi, 1994).

It should be noted that in case of applying ICA to MEG (or EEG) data, both the mixing matrix and the sources give important knowledge of the underlying phenomena. A source naturally describes the temporal activity pattern of the corresponding neuronal population. On the other hand, the corresponding mixing vector contains spatial distribution of the magnetic field on the scalp, enabling the localisation of the active brain region(s), i.e. allows one to solve the inverse problem, together with the additional assumptions.

## 6.2 Analysis-synthesis cycle

Exploratory-source-separation approach to the study of biomedical systems can be described by the diagram shown in Fig. 6.2: The underlying biophysical objects are measured using some devices and the data is *analysed* using source separation methods. Furthermore, the extracted knowledge is used to build a *synthesis*, i.e. to understand the underlying system and to build functional models of it. Finally, these models can be used to generate new hypotheses and to utilise even better source separation methods in order to verify these hypotheses.

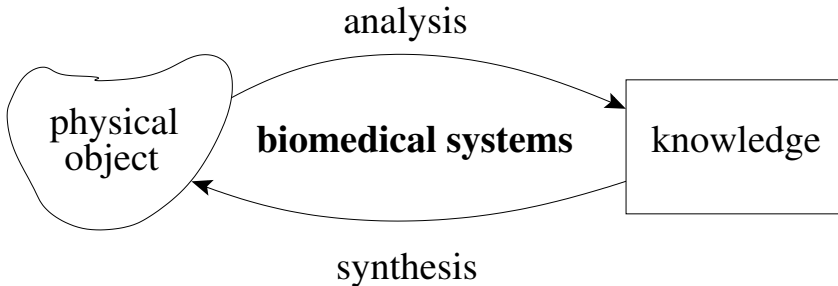


Figure 6.2: *Schematic description of the research in biomedical systems.*

As an example of the analysis-synthesis cycle, consider the following: ICA has been applied to MEG data several times in a blind manner (c.f. Vigário et al., 1998, 1999). This naturally corresponds mainly to the analysis part. These studies suggest that an ECD model (Katila, 1983) explains well the typical magnetic field patterns corresponding to the mixing vectors (synthesis). Vigário (2000) has used this to restrict the mixing mappings to produce dipolar-like field pat-



terns (subsequent analysis). This lead to sparser components, suggesting greater physiological plausibility (subsequent synthesis).

The studies in this thesis concentrate on analysing the data produced by brain imaging methods. However, methods baring resemblance to source separation-methods have been applied to other levels of brain research, too. For instance, ICA has received a growing interest in the computational-neuroscience community. In particular, it has been shown that ICA extracts features from natural images that resemble the receptive fields of the simple cells on the primary visual cortex (V1, c.f. Olshausen and Field, 1996b). Furthermore, such ICA-related methods as slow feature analysis (SFA, Wiskott and Sejnowski, 2002), bubble-ICA (Hyvärinen et al., 2003) and hierarchical DSS (Valpola, 2004) have been shown to extract features similar to complex cells in V1.

ICA has been extensively used in the analysis of biomedical data. For studies on MEG, see for instance Vigário et al. (1997a), Jahn and Cichocki (1998), Tang et al. (2002) and for EEG, Makeig et al. (1996), Vigário (1997), Jung et al. (2000). Both structural and functional MRI have received an increasing interest too (McKeown et al., 1998, Calhoun et al., 2003, McKeown et al., 2003, Karp et al., 2004, Ylipaavalniemi and Vigário, 2004).

## 6.3 Analysis-synthesis in extraction of MEG sources

In this section, we describe experiments where prior knowledge and update of the denoising procedure lead to improved estimates of the underlying sources. We concentrate on spontaneous activity in MEG data.

Since the early EEG and MEG recordings, cortical electromagnetic rhythms have played an important role in clinical research, e.g. in detection of various brain disorders, and in studies of development and aging. It is believed that the spontaneous rhythms, in different parts of the brain, form a kind of resting state that allows for quicker responses to stimuli by those specific areas. For example deprivation of visual stimuli by closing one's eyes induces so called  $\alpha$ -rhythm on the visual cortex, characterised by a strong 8–13 Hz frequency component (c.f. Basar and Schurmann, 1997, Klimesch, 1997, Niedermeyer and Lopes da Silva, 1993, Hämäläinen et al., 1993). For extraction of multiple oscillatory sources in MEG, see Jensen and Vanni (2002).

We examine an MEG experiment where the subject is asked to relax by closing her eyes (producing  $\alpha$ -rhythm). There is also a control state where the subject's eyes are open. The data has been sampled with  $f_s = 200$  Hz, and there are  $T = 65536$  time samples giving a total of more than 300 seconds of measurement. The magnetic fields are measured using a 122-channel MEG device. The data is available in Vigário et al. (1997c). The first source separation results of this data have been reported in Publication 4. Prior to any analysis, the data is high-pass

filtered with cut-off frequency of 1 Hz, to get rid of the dominating very low frequencies.

### Denoising in rhythmic MEG

Examination of the average spectrogram in Fig. 6.3a reveals clear structures indicating the existence of several, presumably distinct, phenomena. The burst-like activity around 10 Hz and the steady activity at 50 Hz dominate the data, but there seem to be some weaker phenomena as well, e.g. on higher frequencies than 50 Hz. To amplify these, we not only sphere the data spatially but temporally as well. This temporal decorrelation actually makes the separation harder, but enables the finding of the weaker phenomena. The normalised and filtered power spectrogram is shown in Fig. 6.3b.

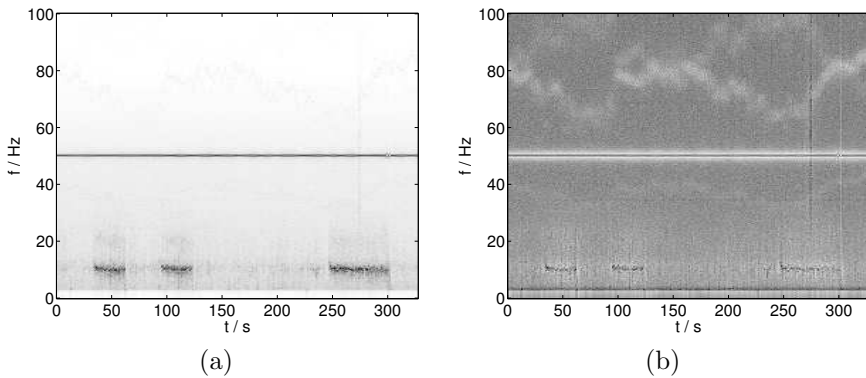


Figure 6.3: (a) Averaged spectrogram of all of the 122 MEG channels. (b) Frequency normalised spectrogram. Time is on the horizontal and frequency on the vertical axis. (from Publication 5)

The spectrogram data seems well suited for demonstrating the exploratory-source-separation use of DSS. We used the time-frequency analysis with lengths of the spectra  $T_f = 256$  (see Sec. 4.4.1). We apply several noise reduction principles based on the estimated variance of the signal and noise discussed in more detail in Publication 6. Specifically, the power spectrogram of the source estimate is smoothed over time and frequency using a 2-D convolution with Gaussian windows. The standard deviations of the Gaussian windows were  $\sigma_t = 8/\pi$  and  $\sigma_f = 8/\pi$ . After this, the instantaneous estimates of the source variances are

decorrelated to get rid of the leakage from other signals. This gives the estimates of the source variances, which are then used to mask the current source estimates to come up with denoised source estimates. Finally, the projection vectors  $\mathbf{w}$  are updated according to the stabilised update-rule (4.24) using the 179-rule (see Sec. 4.5 around Eq. 4.24, for details).

### Separation results

Some extracted sources are shown in Fig. 6.4. Though quite a clear separation of the sources was achieved, some cross-talk between the signals remains. We now turn to more specific masks by taking advantage of the structures uncovered by variance-based masking. In Sec. 6.3.1, the anomalous signal on the upper-right corner of the Fig. 6.4 is inspected further. Furthermore, in Sec. 6.3.2 we show that with specific knowledge it is possible to find even very weak phenomena in MEG data using DSS.

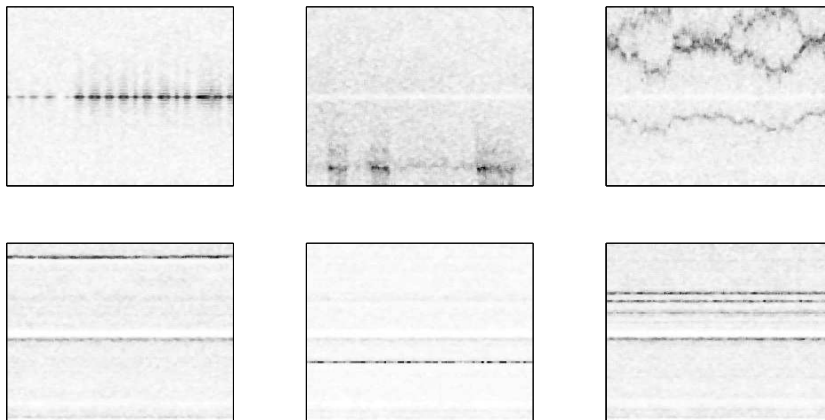


Figure 6.4: *Spectrograms of some of the sources separated using variance sphering. Time is on the horizontal and frequency on the vertical axis. Units as in Fig. 6.3. (adapted from Publication 6)*

#### 6.3.1 Adaptive extraction of the component with wandering frequency content

Exploratory DSS separated some presumable artefacts. In Fig. 6.4, the 3th component on the top row has a curious wandering frequency around 30–40 Hz and

some higher harmonics. In this section, we adaptively maximise SNR of the wandering component. We as well check whether some weaker, but related signals come forward when the masks are adapted.

### **Denoising of the components with wandering frequency contents**

The denoising of the wandering signal is based on the masking of the time-frequency spectrogram. Thus DSS comes close to the approach by Mitra and Pesaran (1999), where a multitaper technique is used together with singular value decomposition. In this example, we confine ourselves to simpler short time DCTs and PCA as suggested for DSS in Secs. 4.2 and 4.4.1.

We adaptively tune a mask in the time-frequency-space so that it takes into account the slow drifting of the base frequency. The very clear 2nd and 3rd harmonics are used to aid in the estimation of the base frequency. The final mask is shown in Fig. 6.5a. Note that the third harmonic surpasses the Nyquist frequency of  $f_s/2 = 100$  Hz at certain locations and causes an aliasing effect.

### **Separation results**

Using the DSS procedure described earlier, we extracted several signals having wandering frequency around 30–40 Hz and higher harmonics. Two of these are shown in Figs. 6.5b and c. As the tuned mask (Fig. 6.5a) is a very narrow one, it can see similar structure in pure Gaussian data already. Comparison of the corresponding eigenvalues revealed that all of the other extracted wandering components, except the two shown, are probably caused by overfitting. The base frequency of the second source is not clearly visible but this appears to be caused by greater noise variance on its frequency range when compared to the higher frequencies where the harmonics are.

## **6.3.2 Adaptive extraction of cardiac subspace in MEG**

Cardiac activity causes magnetic fields as well. Sometimes these are strongly reflected in MEG and can pose a serious problem for the signal analysis of the neural phenomena of interest. In this data, however, the cardiac signals are not visible to the naked eye. Thus, we want to demonstrate the capability of DSS to extract some very weak cardiac signals, using detailed prior information in an adaptive manner.

### **Denoising of the cardiac subspace**

A clear QRS complex can be extracted from the MEG data using standard BSS methods, such as kurtosis- or tanh-based denoising. Due to its sparse nature,

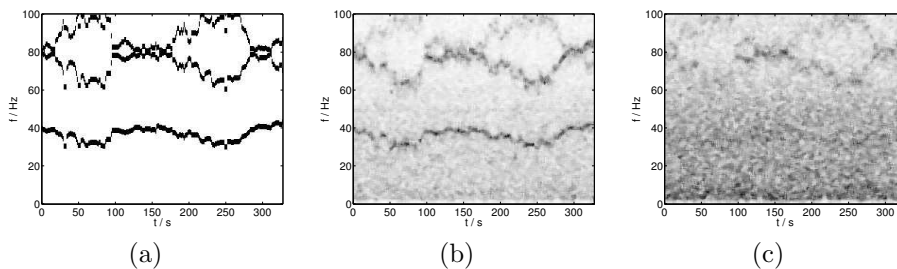


Figure 6.5: (a) The final adapted time-frequency mask for the wandering component. (b) Enhanced spectrogram of one artefact having wandering frequency around 30–40 Hz and harmonics. (c) Enhanced spectrogram of another similar component. Time is on the horizontal and frequency on the vertical axis.

this QRS signal can be used to estimate the onsets of the heart beats. With they are known, we can guide further search using the averaging DSS, as described in Sec. 4.4.1. Every now and then, we reestimate the QRS onsets needed for the averaging DSS.

When the estimation of the QRS locations has been stabilised, a subspace composed of signals having activity phase locked to the QRS complexes is extracted.

### Separation results

Figure 6.6 depicts five signals averaged around the QRS complexes, found using the procedure above<sup>1</sup>. The first signal presents a very clear QRS complex, whereas the second one contains the small but wider P and the T waves. An interesting phenomenon is found in the third signal: there is a clear peak at the QRS onset, which is followed by a slow attenuation phase. We presume that it originates from some kind of relaxation state. Further research may provide one with additional, possibly intriguing, information about the function of the heart.

Two other heart related signals were also extracted. They both show a clear deflection during the QRS complex, but have as well significant activity elsewhere. These two signals might present a case of overfitting, contemplated in Sec. 5. To test this hypothesis, we performed DSS using the same procedure, but for reversed data, i.e.  $t = T, \dots, 2, 1$ . This should break the underlying repetitive structure. Resulting signals should then be pure overfits. They are shown in Fig. 6.6b. The eigenvalues corresponding to the QRS-complex and to the second

<sup>1</sup>For clarity, two identical cycles of averaged heart beats are always shown.

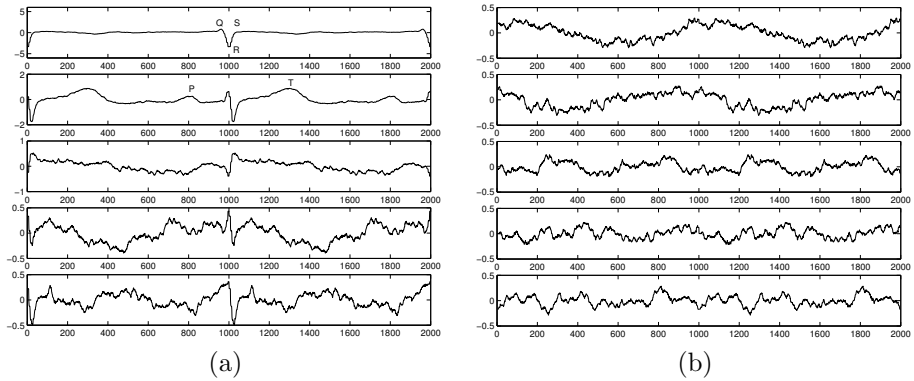


Figure 6.6: *a) Averages of three heart related signals and presumably two overfitting results. b) Averages of five signals from the cardiac control experiment, showing clear overfits. Time on the horizontal axis. (adapted from Publication 5)*

signal having the P and T waves are approximately 10 times higher than the principal eigenvalue of the reversed data. Thus they clearly exhibit some real structure in the data, as already expected. The eigenvalues corresponding to the last three signals are comparable to the principal eigenvalue of the reversed data. Therefore, it is probable that at least part of the structure is due to overfitting. Especially, the slow wavy structure seems to be similar in the control experiment.

It is worth noticing that even the strongest component of the cardiac subspace is rather weakly present in the original data. The other components of the subspace are hardly detectable without advanced methods beyond blind source separation. This clearly demonstrates the power that DSS can provide for an exploring researcher.

## Chapter 7

# Conclusions and future trends

*Space, the final frontier. These are the voyages of the Starship Enterprise, her five-year mission to explore strange new worlds, to seek out new life and new civilization, to boldly go where no [hu]man has gone before.*

–Mission of the Star Trek Enterprise NCC 1701-A, Gene Roddenberry

We considered exploratory source separation in biomedical systems. We introduced denoising source separation as a general framework for both blind algorithms and algorithms where detailed prior knowledge can be utilised. It was suggested that the analysis-synthesis nature of scientific research naturally suggest refinements to the principles source separation is based on. We have shown that DSS is a good candidate for such a task.

We discussed several denoising methods in detail. The methods were divided in two classes: 1) linear and 2) nonlinear denoisings. The linear DSS algorithms were seen to be equivalent to the classical power method applied to the data where comparable linear denoising has been applied to. This leads to very fast, but still powerful, source separation algorithms. Furthermore, a particular nonlinear DSS algorithm was seen to coincide with FastICA, perhaps the fastest existing general ICA algorithm, using kurtosis as the objective function. The denoising interpretation of that particular algorithm has already lead to several improvements, as well.

The performance of different DSS algorithms was examined in simulated data. The best results under poor SNR were achieved using detailed prior knowledge, embedded in linear DSS algorithms. On the other hand, blind approaches implemented using nonlinear DSS algorithms outperformed linear ones with good SNR.

The utility of the DSS algorithms was demonstrated in real magnetoencephalograms. We showed that using DSS, extraction of phenomena having poor SNR becomes possible. In particular, we extracted some weaker parts of the cardiac cycle using linear denoising. The knowledge required for the denoising consisted of the onsets of the cardiac cycles and was acquired in a blind manner. Furthermore, we showed some other cases where adaptation of the denoising procedures leads to improved performance.

Any optimisation method may suffer from overfitting when limited amount of data is available. The issue of overfitting was extensively discussed in case of ICA algorithms. We divided it in two cases: 1) Overfitting is caused by improper or inadequate estimation of the model. 2) The model is not suitable for the data as such but should be revised. In particular, it was noted that the ICA model *as such* is not ideal for MEG data because of the considerable correlations between samples at different time instances. Several solutions were proposed for both cases of overfitting. We discussed also the overfitting in DSS algorithms. There, the type of overfitting encountered depends on the details of the denoising procedure. We proposed quantitative and qualitative measures for the detection of these overfits.

The ESS research described in this thesis, suggests several promising extensions. In the following we mention some of them. While the extensions are worth exploring in a wider context as well, we give suggestions on how the DSS framework could be used therein.

For now, ICA has been applied to several other imaging modalities, including MRI, fMRI, EEG and ECG. Actually, it may be argued that the biomedical signals constitute the most important field of application for ICA. In DSS, the possibility to incorporate prior knowledge may lead to more accurate results and even new findings.

In this thesis, we mainly considered the application of the DSS framework in the study of biomedical systems. We have already extended its use to the separation of CDMA signals. There exist a multitude of other possible applications too. In many fields, denoising tools for the signals of interest already exist. The application of DSS therein should be fairly straightforward.

Source separation is not the only application of ICA-like algorithms. Another important field of application is feature extraction. ICA has been used, for example, for the extraction of features from natural images, similar to those found in the primary visual cortex (Olshausen and Field, 1996a).

Until now, we have only considered extraction of multiple components by forcing the projections to be orthogonal. However, nonorthogonal projections resulting from overcomplete representations provide some clear advantages, especially in sparse codes (Földiák, 1990). Hence, this extension may be found useful if DSS is used for feature extraction.



Linear features can only build a rather limited representation of data. Hence, nonlinear features are often considered. One simple way to generate a nonlinear model is to feed the final source estimates through the same nonlinearity that was used for denoising. This has been used by Valpola (2004) where the nonlinearity was a shrinkage function. Such shrinkage functions can only produce mildly nonlinear mappings. However, building a hierarchy consisting of mildly nonlinear DSS modules can result in very complex mappings. This has been suggested in Publication 5 and discussed further by Valpola (2004).

Similar hierarchy of modules has been suggested for slow feature analysis (SFA, Wiskott and Sejnowski, 2002) which aims at extracting slowly varying complex features from the data. It also has a direct connection to DSS, made explicit in Publication 5.

In addition to temporal slowness in SFA, spatial similarities have been considered, e.g. by Valpola (2004). He uses spatial information in a DSS framework for extraction of complex-cell-like features from a natural scene.

The common feature in SFA and the spatial-DSS approach by Valpola (2004) is that the feature extraction is guided by activations of other parts of the hierarchical network, at different times or at different locations, respectively. This suggests that internal guidance of the feature extraction may be beneficial more generally.

Many researchers have suggested that feedback (top-down, lateral or recurrent) connections from other parts of the brain, often called the context of the input, have an important role in learning (cf., Marr, 1982, Becker and Hinton, 1992, Parga and Rolls, 1998). Recurrent feedback would correspond to temporal and lateral to spatial information. The top-down connections are said to realise attentive mechanisms (cf., Deco and Schürmann, 2000), but they may serve as a feature extraction mechanism as well (Publication 5, Valpola, 2004).

The extensions discussed above are expected to result in new powerful source separation and feature extraction algorithms. They may also suggest new models for computational neuroscience, thus resulting in valuable functional synthesis of neural information processing principles.

## Bibliography

- L. B. Almeida. MISEP – linear and nonlinear ICA based on mutual information. *Journal of Machine Learning Research*, 4 (Dec):1297 – 1318, 2003.
- S.-I. Amari. Neural theory of association and concept formation. *Biological Cybernetics*, 26:175 – 185, 1977.
- S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251 – 276, 1998.
- S.-I. Amari. Natural gradient learning for over- and under-complete bases in ICA. *Neural Computation*, 12:1875 – 1883, 1999.
- B. D. Anderson and J. B. Moore. *Optimal filtering*. Prentice-Hall, 1979.
- H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803 – 851, 1999.
- H. Attias and C. E. Schreiner. Dynamic component analysis. *Neural Computation*, 10:1373 – 1424, 1998.
- P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Physical review A*, 38(1):364 – 374, 1988.
- I. Bankman. *Handbook of Medical Imaging: Processing and analysis*. Academic Press; Elsevier Science & Technology Books, 2000.
- E. Basar and M. Schurmann. Alpha oscillations in brain functioning: an integrative theory. *Int. J. of Psychophysiology*, 26:5 – 29, 1997.
- A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. John Wiley & Sons, New York, 1994.
- S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161 – 163, 1992.

- A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129 – 1159, 1995.
- A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327 – 3338, 1997.
- A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on S.P.*, 45(2):434 – 44, 1997.
- O. Bermond and J.-F. Cardoso. Approximate likelihood for noisy mixtures. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation, ICA'99*, pages 325 – 330, Aussois, France, Jan. 11-15, 1999.
- V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. J. Pekar. ICA of functional MRI data: an overview. In *Fourth Int. Conf. on Independent Component Analysis and Blind Signal Separation*, pages 281 – 288, Nara, Japan, 2003.
- J.-F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *Proc. ICASSP'90*, pages 2655 – 2658, Albuquerque, NM, USA, 1990.
- J.-F. Cardoso. Blind signal separation: Statistical principles. *Proc. of the IEEE*, 86(10):2009 – 2025, 1998.
- J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157 – 192, 1999.
- J.-F. Cardoso. Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4 (Dec):1177 – 1203, 2003.
- R. Cattell. *Factor analysis*. Harper & Brothers, New York, 1952.
- K.-L. Chan, T.-W. Lee, and T. J. Sejnowski. Variational Bayesian learning of ICA with missing data. *Neural Computation*, 15 (8):1991 – 2011, 2003.
- R. A. Choudrey and S. J. Roberts. Flexible Bayesian independent component analysis for blind source separation. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, pages 90 – 95, San Diego, USA, 2001.
- A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley & sons, 2002.

- A. Cichocki, R. E. Bogner, L. Moszczynski, and K. Pope. Modified Herault-Jutten algorithms for blind separation of sources. *Digital Signal Processing*, 7: 80 – 93, 1997.
- A. Cichocki and R. Unbehauen. Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Trans. on Circuits and Systems*, 43(11):894 – 906, 1996.
- A. Cichocki, R. Unbehauen, L. Moszczynski, and E. Rummert. A new on-line adaptive algorithm for blind separation of source signals. In *Proc. Int. Symposium on Artificial Neural Networks ISANN-94*, pages 406 – 411, Tainan, Taiwan, 1994.
- R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1 – 13, 1946.
- G. Deco and B. Schürmann. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision research*, 40:2845 – 2859, 2000.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B (Methodological)*, 39(1):1 – 38, 1977.
- D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society ser. B*, 57:301 – 337, 1995.
- S. C. Douglas, A. Cichocki, and S. Amari. A bias removal technique for blind source separation with noisy measurements. *Electronics Letters*, 1998.
- B. Everitt, editor. *An Introduction to Latent Variable Models*. Chapman and Hall, London, 1984.
- R. M. Everson and S. J. Roberts. Blind source separation for non-stationary mixing. *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, 26:15 – 24, 2000.
- FastICA. The FastICA MATLAB package. 1998. Available at <http://www.cis.hut.fi/projects/ica/fastica/>.
- P. Földiák. Forming sparse representations by local anti-hebbian learning. *Biological Cybernetics*, 64:165 – 170, 1990.
- M. Funaro, E. Oja, and H. Valpola. Independent component analysis for artefact separation in astrophysical images. *Neural networks*, 16(3 – 4):469 – 478, 2003.

- M. S. Gazzaniga, editor. *The New Cognitive Neurosciences*. A Bradford book/MIT Press, 2nd edition, 2000.
- A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, Florida, 1995.
- M. Girolami, editor. *Advances in Independent Component Analysis*. Springer-Verlag, 2000.
- R. L. Gorsuch. *Factor Analysis*. Lawrence Earlbaum Associates, Hillsdale, NJ, 2nd edition, 1983.
- M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:413 – 497, 1993.
- M. Hämäläinen and R. Ilmoniemi. Interpreting magnetic fields of the brain: Minimum norm estimates. *Med. Biol. Eng. Comp.*, 32:35 – 42, 1994.
- L. K. Hansen, J. Larsen, and T. Kolenda. Blind detection of independent dynamic components. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'01)*, pages 3197 – 3200, Seattle, Washington, USA, 2001.
- S. Haykin. *Neural Networks – A Comprehensive Foundation, 2nd ed.* Prentice-Hall, 1999.
- S. Haykin, editor. *Unsupervised Adaptive Filtering, Vol. 2: Blind Deconvolution*. Wiley, 2000.
- C. W. Hesse and C. J. James. An efficient time-frequency approach to blind source separation based on wavelets. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'04)*, pages 1048 – 1055, Granada, Spain, 2004.
- J. Himberg and A. Hyvärinen. Icasso: software for investigating the reliability of ica estimates by clustering and visualization. In *Proc. 2003 IEEE workshop on neural networks for signal processing (NNSP'2003)*, pages 259 – 268, Toulouse, France, 2003.
- G. E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pages 5 – 13, Santa Cruz, CA, USA, 1993.

- P. A.d.F.R Højen-Sørensen, O. Winther, and L. K. Hansen. Mean-field approaches to independent component analysis. *Neural Computation*, 14:889 – 918, 2002.
- K. Holzinger and H Harman. *Factor analysis: a synthesis of factorial methods*. The University of Chicago Press, Chicago, Illinois, 3rd edition, 1951.
- P. Horst. *Factor analysis of data matrices*. Holt, Rinehart and Winston, inc., New York–Chicago–San Francisco–Toronto–London, 1965.
- H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational Psychology*, 24:417 – 441, 498 – 520, 1933.
- P. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191 – 210, 2000.
- J. Hurri, A. Hyvärinen, J. Karhunen, and E. Oja. Image feature extraction using independent component analysis. In *Proc. NORSIG'96*, pages 475 – 478, Espoo, Finland, 1996.
- A. Hyvärinen. Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22:49 – 67, 1998a.
- A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing 10 (Proc. NIPS'98)*, pages 273 – 279. MIT Press, 1998b.
- A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626 – 634, 1999a.
- A. Hyvärinen. Sparse code shrinkage: Denoising by maximum likelihood estimation. *Neural Computation*, 12(3):429 – 439, 1999b.
- A. Hyvärinen. Complexity pursuit: Separating interesting components from time-series. *Neural Computation*, 13(4):883 – 898, 2001.
- A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1525 – 1558, 2001a.
- A. Hyvärinen and J. Hurri. Blind separation of sources that have spatiotemporal variance dependencies. *Signal Processing*, 84(2):247 – 254, 2004.
- A. Hyvärinen and M. Inki. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 17:139 – 152, 2002.

- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, 2001b.
- A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483 – 1492, 1997.
- A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429 – 439, 1999.
- A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, 20(7):1237 – 1252, 2003.
- ICA00. Second int. workshop on independent component analysis and blind signal separation, 2000. Helsinki, Finland.
- ICA01. Third int. conference on independent component analysis and blind signal separation, 2001. San Diego, United States.
- ICA03. Fourth int. conference on independent component analysis and blind signal separation, 2003. Nara, Japan.
- ICA04. Fifth int. conference on independent component analysis and blind signal separation, 2004, Sept. Granada, Spain.
- ICA99. First int. workshop on independent component analysis and blind signal separation, 1999. Aussois, France.
- S.-I. Ito, Y. Mitsukura, M Fukumi, and N. Akamatsu. A feature extraction of the EEG using the factor analysis and neural networks. In V. Palade, R. J. Howlett, and L. C. Jain, editors, *Proc. of the 7th Int. Conf. on Knowledge-Based Intelligent Information & Engineering Systems*, pages 609 – 616. LNAI, Springer-Verlag, Berlin Heidelberg, 2003.
- O. Jahn and A. Cichocki. Identification and elimination of artifacts from MEG signals using efficient independent components analysis. In *Proc. of the 11th Int. Conf. on Biomagnetism (BIOMAG-98)*, Sendai, Japan, 1998.
- O. Jensen and S. Vanni. A new method to identify multiple sources of oscillatory activity from magnetoencephalographic data. *Neuroimage*, 15:568 – 574, 2002.
- H. Jokeit and S. Makeig. Different event-related patterns of gamma-band power in brain waves of fast- and slow-reacting subjects. *Proc Natl Acad Sci, USA*, 91(14):6339 – 6343, 1994.

- M. Jordan, editor. *Learning in Graphical Models*. The MIT Press, Cambridge, MA, USA, 1999.
- T.-P. Jung, S. Makeig, T.-W. Lee, M. McKeown, G. Brown, A. Bell, and T. J. Sejnowski. Independent component analysis of biomedical signals. In *Proc. Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'00)*, pages 633 – 644, Helsinki, Finland, 2000.
- C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1 – 10, 1991.
- J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113 – 127, 1994.
- E. Karp, H. Gävert, J. Särelä, and R. Vigário. Independent component analysis decomposition of structural MRI. In *Proc. 2nd IASTED int. conf. on biomedical engineering (BioMED2004)*, pages 83 – 88, Innsbruck, Austria, 2004.
- T. Katila. On the current multipole presentation of the primary current distributions. *Nuovo Cimento*, 2D:660 – 664, 1983.
- M. Kendall and A. Stuart. *The Advanced Theory of Statistics*. Charles Griffin & Company, 1958.
- W. Klimesch. Eeg-alpha rhythms and memory processes. *International Journal of Psychophysiology*, 26:319 – 340, 1997.
- K. H. Knuth. Bayesian source separation and localization. In A. Mohammad-Djafari, editor, *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems*, pages 147 – 158, San Diego, USA, 1998.
- V. Koivunen, M. Enescu, and E. Oja. Adaptive algorithm for real-time blind separation from noisy mixtures. *Neural Computation*, 13 (10):2339 – 2357, 2001.
- P. Kuosmanen and J. T. Astola. *Fundamentals of nonlinear digital filtering*. CRC press, 1997.
- H. Lappalainen. Ensemble learning for independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 7 – 12, Aussois, France, 1999.
- H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, pages 75 – 92. Springer-Verlag, Berlin, 2000.



- T.-W. Lee. *Independent Component Analysis: Theory and Applications*. Kluwer academic publishers, 1998.
- T.-W. Lee, M Girolami, A. J. Bell, and T. J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *An Int. J. of computers & mathematics with applications*, 31:1 – 21, 2000.
- W. J. Levelt. Spoken word production: A theory of lexical access. *Proc Natl Acad Sci, USA*, 98(23):13464 – 13471, 2001.
- M. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337 – 365, 2000.
- M. M. Loève. *Probability theory*. Van Nostrand, Princeton, 1955.
- D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.
- S. Makeig, A. Bell, T.-P. Jung, and T. Sejnowski. Independent component analysis of electroencephalographic data. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Neural Information Processing Systems 8 (Proc. NIPS'95)*, pages 145 – 151, Cambridge MA, 1996. MIT Press.
- S. R. Marder, J. M. Davis, and G. Chouinard. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the north american trials. *Journal of Clinical Psychiatry*, 58:538 – 546, 1997.
- D. Marr. *A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, 1982.
- M. McKeown, S. Makeig, S. Brown, T.-P. Jung, S. Kindermann, A. Bell, V. Iragui, and T. J. Sejnowski. Blind separation of functional magnetic resonance imaging (fMRI) data. *Human Brain Mapping*, 1998.
- M. J. McKeown, L. K. Hansen, and T. J. Sejnowski. Independent component analysis of functional fMRI: what is signal and what is noise. *Current Opinion in Neurobiology*, 13:620 – 629, 2003.
- F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A resampling approach to estimate the stability of one- and multidimensional independent components. *IEEE Trans. Biom. Eng.*, 49(12):1514 – 1525, 2002.
- J. Miskin and David J. C. MacKay. Ensemble learning for blind source separation. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*, pages 209 – 233. Cambridge University Press, 2001.

- P. P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophysical Journal*, 76:691 – 708, 1999.
- J. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:541 – 557, 1994.
- K.-R. Müller, P. Philips, and A. Ziehe. *JADE<sub>TD</sub>*: Combining higher-order statistics and temporal information for blind source separation (with noise). In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA '99)*, pages 87 – 92, Aussois, France, 1999.
- R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355 – 368. The MIT Press, Cambridge, MA, USA, 1999.
- E. Niedermeyer and F. Lopes da Silva, editors. *Electroencephalography. Basic principles, clinical applications, and related fields*. Baltimore: Williams & Wilkins, 1993.
- C. Nikias and J. Mendel. Signal processing with higher-order spectra. *IEEE Signal Processing Magazine*, pages 10 – 37, July 1993.
- E. Oja. A simplified neuron model as a principal component analyzer. *J. of Mathematical Biology*, 15:267 – 273, 1982.
- E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927 – 935, 1992.
- E. Oja. The nonlinear PCA learning rule and signal separation – mathematical analysis. Technical Report A 26, Helsinki University of Technology, Laboratory of Computer and Information Science. Submitted to a journal, 1995.
- E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25 – 46, 1997.
- E. Oja. Convergence of the symmetrical FastICA algorithm. In *Proc. 9th Int. Conf. on Neural Information Processing (ICONIP'02)*, Singapore, 2002.
- E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Math. Analysis and Applications*, 106:69 – 84, 1985.
- E. Oja, H. Ogawa, and J. Wangviwattana. Learning in nonlinear constrained Hebbian networks. In T. Kohonen et al., editor, *Artificial Neural Networks, Proc. ICANN'91*, pages 385 – 390, Espoo, Finland, 1991. North-Holland, Amsterdam.

- B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607 – 609, 1996a.
- B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333 – 340, May 1996b.
- R. K. Olsson and L. K. Hansen. Probabilistic blind deconvolution of non-stationary sources. In *Proc. EUSIPCO*, pages 1697 – 1700. Elsevier, 2004.
- A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition, 1991.
- N. Parga and E. T. Rolls. Transform invariant recognition by association in a recurrent network. *Neural Computation*, 10(6):1507 – 1525, 1998.
- L. Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, 4 (Dec):1261 – 1269, 2003.
- L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller. Unmixing hyperspectral data. In *Advances in Neural Information Processing Systems*, volume 12, pages 942 – 948. MIT Press, 2000.
- D. B. Percival and W. T. Walden. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press, Cambridge, UK, 1993.
- D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. on Signal Processing*, 49:1837 – 1848, 2001.
- J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465 – 471, 1978.
- B. Rivet, C. Jutten, and V. Vigneron. Wavelet de-noising for blind source separation in noisy mixtures. Technical report, BLInd Source Separation project (BLISS IST 1999-14190), 2003.
- B. Rivet, V. Vigneron, A. Paraschiv-Ionescu, and C. Jutten. Wavelet de-noising for blind source separation in noisy mixtures. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'04)*, pages 263 – 270, Granada, Spain, 2004.
- R. A. Robb. *Biomedical imaging, visualization and analysis*. Wiley-Liss, Inc., 2000.

- S. Roberts and R. Everson, editors. *Independent Component Analysis: Principles and Practice*. Cambridge Univ. Press, 2001.
- D. Rubin and D. Thayer. EM algorithms for factor analysis. *Psychometrika*, 47:69 – 76, 1982.
- C. Spearman. “General intelligence,” objectively determined and measured. *American Journal of Psychology*, 15:201 – 293, 1904.
- J. V. Stone. *Independent Component Analysis: A Tutorial Introduction*. MIT Press, 2004.
- J. Särelä and R. Vigário. A Bayesian approach to overlearning in ICA: a comparison study. Technical Report A70, Helsinki university of technology, Laboratory of computer and information science, 2003.
- A. Taleb and C. Jutten. Source separation of post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47:2807 – 2820, 1999.
- T. Tanaka and A. Cichocki. Subband decomposition independent component analysis and new performance criteria. In *Proc. ICASSP'2004*, volume 5, pages 541 – 544, Montreal, Canada, 2004.
- A. Tang, B. Pearlmutter, N. Malaszenko, D. Phung, and B. Reeb. Independent components of magnetoencephalography: Localization. *Neural Computation*, 14:1827 – 1858, 2002.
- L. Tong, V. Soo, R. Liu, and Y. Huang. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38:499 – 509, 1991.
- K. Torkkola. Blind separation for audio signals: are we there yet? In *Proc. Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'99)*, pages 239 – 244, Aussois, France, 1999.
- I. Tsiftsis, K. N. Fountoulakis, K. Sitzoglou, A. Papanicolaou, K. Phokas, F. Fotiou, and G. St Kaprinis. Clinical and neuroimaging correlates of abnormal short-latency somatosensory evoked potentials in elderly vascular dementia patients: A psychophysiological exploratory study. *Ann Gen Hosp Psychiatry*, 2(1):8, 2003.
- H. Valpola. Behaviourally meaningful representations from normalisation and context-guided denoising. Technical report, Artificial Intelligence Laboratory, Department of Information Technology, University of Zurich, 2004. Available at Cogprints: <http://cogprints.ecs.soton.ac.uk/archive/00003633/>.

- H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647 – 2692, 2002.
- H. Valpola and P. Pajunen. Fast algorithms for Bayesian independent component analysis. In *Proceedings of the second international workshop on independent component analysis and blind signal separation, ICA'00*, pages 233 – 238, Espoo, Finland, 2000.
- H. Valpola, T. Raiko, and J. Karhunen. Building blocks for hierarchical latent variable models. In *Proc. 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, pages 710 – 715, San Diego, USA, 2001.
- Vectorview<sub>TM</sub>. The vectorview<sub>TM</sub> by Neuromag, Ltd.: a 306 channel MEG device. For more information see: <http://www.neuromag.com/vectorview.html>, 1997.
- M. Vetterli and J. Kovacevic. *Wavelets and subband coding*. Prentice-Hall, 1995.
- R. Vigário. Extraction of ocular artifacts from EEG using independent component analysis. *Electroenceph. clin. Neurophysiol.*, 103(3):395 – 404, 1997.
- R. Vigário. Dipole modeling in FastICA decomposition of evoked responses. In *Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'00)*, Helsinki, Finland, 2000.
- R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 229 – 235. MIT Press, Cambridge MA, 1997a.
- R. Vigário, J. Särelä, V. Jousmäki, and E. Oja. Independent component analysis in decomposition of auditory and somatosensory evoked fields. In *Proc. Int. workshop on Independent Component Analysis and Blind Source Separation of Signals (ICA'99)*, Aussois, France, 1999.
- R. Vigário, J. Särelä, and E. Oja. The MEG artefacts data. 1997b. Available at [http://www.cis.hut.fi/projects/ica/eegmeg/MEG\\_art.cgi](http://www.cis.hut.fi/projects/ica/eegmeg/MEG_art.cgi).
- R. Vigário, J. Särelä, and E. Oja. The rhythmic MEG data. 1997c. Available at [http://www.cis.hut.fi/projects/ica/eegmeg/MEG\\_rhythm.cgi](http://www.cis.hut.fi/projects/ica/eegmeg/MEG_rhythm.cgi).
- R. Vigário, J. Särelä, and E. Oja. Independent component analysis in wave decomposition of auditory evoked fields. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.

- V. Vigneron, A. Paraschiv-Ionescu, A. Azancot, O. Sibony, and C. Jutten. Fetal electrocardiogram extraction based on non-stationary ICA and wavelet denoising. In *Proceedings of ISSPA 2003*, Paris (France), July 2003.
- C. S. Wallace. Classification by minimum-message-length inference. In S. G. Aki, F. Fiala, and W. W. Koczkodaj, editors, *Advances in Computing and Information – ICCI '90*, volume 468 of *Lecture Notes in Computer Science*, pages 72 – 81. Springer, Berlin, 1990.
- C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 49(3):240 – 265, 1987.
- J. H. Wilkinson. *The algebraic eigenvalue problem*. Monographs on numerical analysis. Clarendon press, London, 1965.
- S. Winter, H. Sawada, S. Araki, and S. Makino. Overcomplete BSS for convolutive mixtures based on hierarchical clustering. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'04)*, pages 652 – 660, Granada, Spain, 2004.
- L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14:715 – 770, 2002.
- L. Xu, E. Oja, and C. Suen. Modified Hebbian learning for curve and surface fitting. *Neural Networks*, 5:441 – 457, 1992.
- J. Ylipaavalniemi and R. Vigário. Analysis of auditory fMRI recordings via ICA: A study on consistency. In *Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN 2004)*, Budabest, Hungary, 2004.
- 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 (Dec):1319 – 1338, 2003.
- A. Ziehe and K.-R. Müller. TDSEP — an effective algorithm for blind separation using time structure. In *Proc. int. conf. at neural networks (ICANN'98)*, pages 675 – 680, Skövde, Sweden, 1998.
- J. E. Zimmerman, P. Thiene, and J. T. Harding. Design and operation of stable rf-biased superconducting point-contact quantum devices and a note on the properties of perfectly clean metal contacts. *Journal of Applied Physics*, 41: 1572 – 1580, 1970.