HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Computer Science and Engineering

Independent component analysis of magnetoencephalographic signals

Jaakko Särelä

The thesis belongs to the special subject of information science

Supervisor Professor Erkki Oja
Instructor MSc. Ricardo Vigário

Espoo, Finland 6th January 1999

HELSINKI UNIVERSITY OF TECHNOLOGY **ABSTRACT OF THE MASTER'S THESIS**

Author and name of the thesis:

Jaakko Särelä

Independent component analysis of magnetoencephalographic signals

Date: 6th January 1999 **Number of pages:** 66

Department: Department of Engineering Physics and Mathematics

Professorship: Tik-61 Computer and Information Science

Supervisor: Professor Erkki Oja

Instructor: Master of Science (Physics) Ricardo Vigário

Independent component analysis (ICA) solves the problem of finding unknown independent sources from their linear mixtures. It can be used e.g. in blind source separation and in projection pursuit to find independent features from the data. FastICA is one of the algorithms that solves this problem. It has been shown to be superior in speed to other existing robust algorithms and still giving roughly the same accuracy.

Magnetoencephalography (MEG) is a sophisticated non-invasive measuring technique sensitive to the magnetic activity in human brain. It measures the changes in magnetic flux outside the human head. The flux is caused by the electrical activity of synchronised neurons. MEG has a very high time resolution, even up to one millisecond, and a good spatial resolution, since many (even hundreds of) sensors can be combined in the same device, giving a resolution close to 1-3 mm.

The MEG data can be quite high-dimensional and complex, thus needing sophisticated analysis techniques. This thesis applies FastICA on MEG recordings, showing that this implementation of ICA reproduces many conventional results without the burden of approximative models. It also presents a technique, shown to be good for extraction of artefacts, such as eye movements and heart beating from MEG recordings.

Keywords: independent component analysis, blind source separation, projection pursuit, magnetoencephalography, decomposition

TEKNILLINEN KORKEAKOULU **DIPLOMITYÖN TIIVISTELMÄ**

Tekijä ja työn nimi:

Jaakko Särelä
Independent component analysis of magnetoencephalographic signals
(Magnetoenkefalografisten signaalien analysointi riippumattomien komponenttien menetelmällä)

Päivämäärä: 6. tammikuuta 1999 **Sivumäärä:** 66

Osasto: Teknillisen fysiikan ja matematiikan osasto
Professuuri: Tik-61 Informaatiotekniikka

Valvoja: Professori Erkki Oja
Ohjaaja: Filosofian maisteri (MSc. Physics) Ricardo Vigário

Riippumattomien komponenttien menetelmä (ICA) ratkaisee ongelman, jossa halutaan löytää tuntemattomat riippumattomat lähteet niiden lineaarisista sekoitteista. Sitä voidaan käyttää mm. sokeaan lähteen erotteluun ja riippumattomien piirteiden löytämiseen. FastICA on eräs algoritmeista, joka ratkaisee tämän ongelman. Se on osoittautunut nopeudeltaan ylivoimaiseksi muihin yleispäteviin ICA-algoritmeihin verrattaessa ja kuitenkin suunnilleen yhtä tarkaksi.

Magnetoenkefalografia (MEG) on kehittynyt mittausmenetelmä aivojen magneettisen aktiivisuuden mittaamiseen. Se on täysin ei-invasiivinen ja mittaa magneettivuon muutoksia kallon pinnalla, joita aiheuttavat synkronissa laukovien neuronien sähköinen aktiivisuus. MEG:llä on erittäin hyvä ajallinen resoluutio, jopa millisekunnin luokkaa, ja sen paikannusresoluutiokin on varsin hyvä, koska jopa useita satoja sensoreita voidaan yhdistää samaan mittalaitteeseen.

Koska MEG-data on suuridimensioista ja hyvin monimutkaista, sen analysoimiseen tarvitaan kehittyneitä menetelmiä. Tämä diplomityö soveltaa FastICA:a MEG-dataan, osoittaen, että monet perinteiset tulokset voidaan toistaa ICA:lla ilman aproksimoivaa mallinnusta. Lisäksi työssä esitetään menetelmä häiriöiden, kuten silmien liikkeiden ja sydämen lyöntien, poistamiseen MEG-datasta.

Avainsanat: riippumattomien komponenttien menetelmä, sokea lähteen erottaminen, magnetoenkefalografia, hajotelma

Preface

Independent component analysis (ICA) has recently been one of the main research areas in the Laboratory of Computer and Information Science at Helsinki University of Technology. The laboratory is specialised in neural solutions and is, because of that, also interested in brain research. One of the research groups, where Mr Ricardo Vigário is the principal researcher, applies ICA to brain research in collaboration with the Brain Research Unit in the Low Temperature Laboratory at Helsinki University of Technology. I was able to join this group, when I started my Master's thesis in early summer of 1997.

I would like to thank the Laboratory of Computer and Information Science as well as the Neural Network Research Centre for the extremely pleasant working environment I have enjoyed from the first day I have worked in here in summer 1994. I would especially like to thank prof. Oja for taking me into his research group and for supervising this thesis and prof. Kohonen for providing excellent computer facilities. Great thanks are also due to my instructor Mr Ricardo Vigário: our tight collaboration, of which this thesis is a remarkable example, has really been rewarding and I am greatly in debt for all the valuable information I have received from him. I would also like to thank prof. Riitta Hari, Dr Matti Hämäläinen and Dr Veikko Jousmäki for collaboration as well as valuable advises concerning magnetoencephalography and brain research. Furthermore I would like to thank all the personnel in the laboratory and research center, in particular Dr. Aapo Hyvärinen for invaluable advises and all the fun, Mr Jarmo Hurri for such a clear presentation of ICA in his master's thesis, which has been of great help to me, and the lunch gang especially for all the great time at lunch hours. Last, but not least, I would like to thank my wife Katja for all the interdisciplinary, highly stimulating conversations and general support.

In Espoo, 20th January 1999

Jaakko Särelä

Contents

Abbreviations and Notations

BSS	blind source separation
CAT or CT	computer axial tomography
ECD	equivalent current dipole
ECG	electrocardiography
EEG	electroencephalography
EVD	eigenvalue decomposition
fMRI	functional magnetic resonance imaging
ICA	independent component analysis
MEG	magnetoencephalography
MRI	magnetic resonance imaging
PCA	principal component analysis
PET	positron emission tomography
PP	projection pursuit
SPECT	single photon emission computerised tomography
SQUID	superconducting quantum interference device

$i, \dots, n$	natural numbers
$\alpha, \beta, \gamma, \dots$	scalars
λ, λ_i	eigenvalues
κ_i	ith cumulant
$\mathbf{a}, \dots, \mathbf{z}$	vectors
$a_i, \dots, z_i$	ith component of the vectors $\mathbf{a}, \dots, \mathbf{z}$ respectively
$\mathbf{e}, \mathbf{e}_i$	eigenvectors
x, y	random variables
$\mathbf{x}, \mathbf{y}$	random vectors
x_i, y_i	random variables or components of random vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$	matrices
$\mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i \dots$	ith column vector of the matrices
$\mathbf{s}$	a column vector of sources
$\mathbf{s}(t)$	a column vector of the time series of the sources
s_i	ith source
$\mathbf{x}$	a column vector of measurements
$\mathbf{x}(t)$	a column vector of the measurement time series
x_i	ith measurement
$\mathbf{A}$	ICA mixing matrix
$\mathbf{a}_i$	ith row vector of the mixing matrix $\mathbf{A}$, giving the ith measurement
$\mathbf{W}$	ICA filtering matrix
$\mathbf{w}_i$	ith row vector of the filtering matrix $\mathbf{W}$, separating the ith independent source
$\mathbf{v}$	a column vector of whitened measurements $\mathbf{x}$
$\mathbf{B}$	whitened ICA mixing matrix or magnetic field
$\mathbf{E}$	a matrix of eigenvectors or electric field
$\mathbf{D}$	a diagonal matrix of eigenvalues
$\mathbf{V}$	whitening matrix
m	the number of measurements
n	the number of sources
f, g, h	functions
f_x	density function of random variable x
f_{xy}	the joint density function of random variables x and y
$E\{\cdot\}$	expected value
$Var\{\cdot\}$	variance
$Cov\{\cdot\}$	covariance matrix

Chapter 1

Introduction

The basic principles underlying cognitive processes have always intrigued the human mind. Already for centuries it has been evident that the brain has an utterly significant role in those processes. That has made the brain one of the most central subjects in the scientific research. In the latter part of this century, brain research has grown exponentially due to exceptional progress in various fields of science: quantum physics, medicine, cognitive psychology, and computer science, just to name a few. Quantum physics has provided researchers with new measuring devices and clinical medicine has been able to use them to produce better diagnostics etc. Cognitive psychology has been able to link these new results into former psychophysical results. The challenge of information technology and information science is to provide these sophisticated measuring techniques with sophisticated analysis tools.

Magnetoencephalography (MEG) is one such measuring technique. The measuring device itself is based on such areas in modern physics as superconductors and quantum physics. It measures the magnetic activity in brain. This data has a very complex structure and often poor signal to noise ratio. The background noise, due to external sources outside the body, is rather small since the measurements are done in a magnetically shielded room, but e.g. heart produces signals that may be many orders of magnitude stronger than any of the brain signals. Even without any bodily noise, the signals corresponding to specific events may have very low variances in comparison to the total variance of the data. This makes the signal analysis difficult. A common way to strengthen the desired signal in respect to the others, is to repeat the experimental conditions in order to accumulate several responses of the same type. These responses are then averaged. This of course can not be done if the event is unique, such as an epileptic seizure, or if it is non-stationary such as any event containing learning, or if there are no exact information about the moment of the onset of the response.

One way to reduce the noise without detailed knowledge about the onset of the studied response is to use principal component analysis (PCA). PCA is a way to express high-dimensional data with fewer dimensions, yet preserving maximal variance. This leads the projected signals to be decorrelated. However, the decorrelated signals do not necessarily correspond to physically distinct sources.

Independent component analysis (ICA) is a recently developed method for performing a linear transformation, which ensures statistical independence between the sources. During the year 1996, the study of independent component analysis increased rapidly in volume as well as in density in the Laboratory of Computer and Information Science. It was also during that year, when post graduate student Mr Vigário started to apply ICA in medical signal analysis, since independence seems a plausible and appealing attribute for distinct brain sources.

Among the world's leading research groups in magnetoencephalography is the Brain Research Unit of the Low Temperature Laboratory in the Helsinki University of Technology. In the late 1996 a collaboration was started between the two laboratories to continue the research, now to apply ICA in MEG signal analysis. Three persons participated in this collaboration from the Low Temperature Laboratory: the head of the Brain Research Unit, prof. Riitta Hari, Dr Matti Hämäläinen and Dr Veikko Jousmäki.

It was also in the late 1996, when I got interested in this project. I have always been interested in the process of thinking and brain research in general and I started my minor subject studies in cognitive science in the University of Helsinki during the same year. Full devotion to the project I gave from the summer of 1997, when I started to work for the project to make my Master's thesis.

The thesis consists of two parts: The first part, covers the mathematical framework of ICA, presenting the problem in chapter 2 together with a classical way for decorrelating the sources, namely *whitening* (in section 3.1) or *principal component analysis* (in section 5.1.1). In the rest of the chapter 3, one solution to the ICA problem will be presented step-by-step: After decorrelation of the data through whitening, we consider the *central limit theorem* in section 3.2 which states that the distribution of a sum of random variables tends to Gaussian, when the number of summed variables increases. This principle leads us to the fourth order cumulant *kurtosis* in section 3.2 and the generalised non-Gaussianity measures or *contrast functions* in section 3.3. These contrast functions are then optimised using *Newton iteration* leading to the *FastICA-algorithm* presented in section 3.4. Other algorithms for solving ICA are presented shortly in chapter 4. There are as well some discussion about the confluences between different algorithms. In the 5th and final chapter of the first part, we deal with some real-life problems one could encounter, when using ICA in real-life applications. The appendix A, at the end, gives a short introduction to the information theory needed in the first part of the thesis.

The second part applies FastICA to MEG. It starts with a short introduction to the *cognitive neurosciences* in general (chapter 6) and to *auditive* and *tactile* senses in more detail (section 6.2), since they are the two modalities of perception studied in the experimental part. After that, we present the magnetoencephalography as a high-dimensional brain imaging technique and construct the bridge needed to connect the general mathematical concept ICA to the specific physical concept of MEG in chapter 7. After that results of applying ICA to MEG in four studies are reported. The studies consist of: 1.

The extraction of artefacts in MEG data; 2. The non-averaged independent component analysis of auditory evoked fields; 3. The decomposition of averaged auditory evoked fields into independent components; and finally 4. The decomposition of auditory and somatosensory evoked fields into independent components and localisation of the equivalent current dipoles corresponding to those components. The last chapter summarises the results obtained in the experimental part and discusses some paths for future research to establish ICA as one of the standards in medical signal analysis. Finally, appendix B describes shortly the medical terms used in this thesis.

Part I

Independent component analysis: a mathematical framework

Chapter 2

Problem statement

Consider n time series of T samples $s_1(t),\ s_2(t),\ \cdots,\ s_n(t)$, where $s_i(t)$ presents the ith time series and $t = 1,\ 2,\ \cdots,\ T$. Let us think, that we are interested in knowing these time series, but we can not observe them directly. Instead, we can observe m weighted sums of them. Usually it is assumed that $m \geq n$. These observations are noted by $x_1(t),\ x_2(t),\ \cdots x_m(t)$, where $x_i(t)$ is calculated by weighting each of the s'es with a constant and summing them:

$$x_i(t) = \sum_{k=1}^{n} a_{ik} s_k(t), \tag{2.1}$$

where a_{ik} is the weighting constant. This can be also presented as a product of two vectors:

$$x_i(t) = \mathbf{a}_i^T \mathbf{s}(t), \tag{2.2}$$

where $\mathbf{a}_i = [a_{i1}\, a_{i2} \cdots\ a_{in}]^T$, is a column vector consisting of the weighting constants of the ith observed time series and $\mathbf{s}(t) = [s_1(t)\, s_2(t) \cdots s_n(t)]^T$ is a column vector of the sources. All the $\mathbf{a}_i$'s can be combined in the same matrix $\mathbf{A}$. Then

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2.3}$$

gives all the observed signals $\mathbf{x}(t)$. The weighting matrix $\mathbf{A}$ is usually called the *mixing matrix*, since it mixes the sources $\mathbf{s}(t)$. Finding the unknown sources $\mathbf{s}(t)$ without knowing anything about the mixing matrix $\mathbf{A}$ is called blind source separation (BSS) [51, 13].

Example 1 *A very common everyday life's example of a BSS-problem is the so called* cocktail party problem*: Imagine a cocktail party in a hall, where there are lots of speakers talking at the same time. If there would be at most n speakers and we could record the sounds of the hall with at least n microphones, we should be able to separate each speaker from these recordings, with some limitations on the acoustics of the hall and background noise etc.*

The BSS problem can also be stated in reverse order, which is more suitable in some cases:

$$\mathbf{s}(t) = \mathbf{W}\mathbf{x}(t), \tag{2.4}$$

where $\mathbf{W}$ is called the *filtering matrix* revealing the sources.

Projection pursuit (PP) [29, 30, 39, 48, 93, 22] is very similar to BSS, but it does not assume that there actually are sources behind the measurements. It assumes that using a mapping $\mathbf{s}(t) = \mathbf{f}(\mathbf{x}(t))$, some interesting properties of the measurements $\mathbf{x}(t)$ are revealed. In general, $\mathbf{f}$ could be any mapping, linear or non-linear, but often only linear ones are considered because only they can be analysed in full detail.

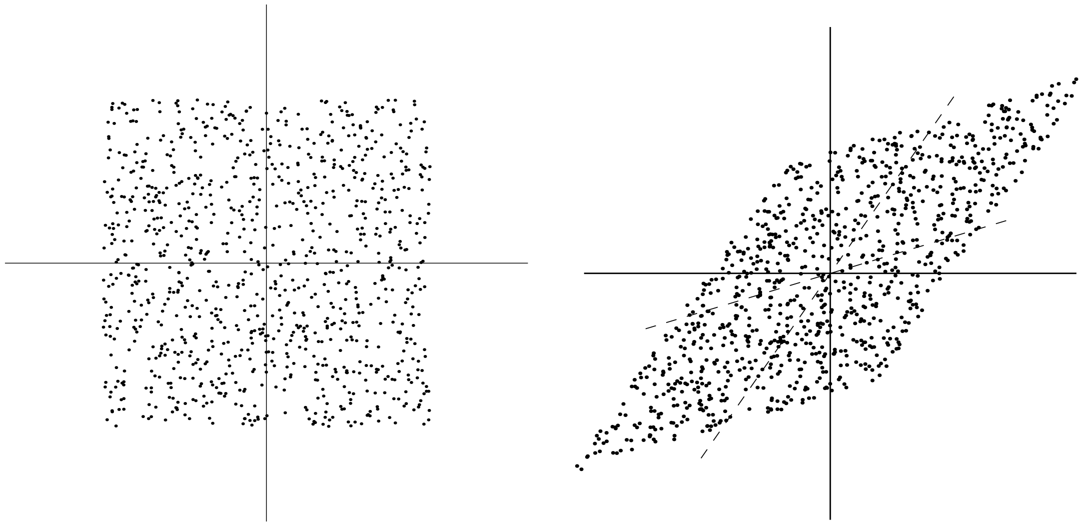
Example 2 *One situation, where one could use PP, is process monitoring: We have a very complex process, which we observe through many variables. Projection pursuit could be used to describe the process. Another application might be finding of interesting factors from stock market changes.*

The BSS and PP problems can be solved by seeing the sources $\mathbf{s}$ and the measurements $\mathbf{x}$ as instances of random variables, and making some assumption on the random variables[1]. If it is assumed that the sources $\mathbf{s}$ are statistically *independent*[2] of each others, the BSS and PP become a problem called independent component analysis (ICA, the idea was first presented in [50, 51]; the formulation in (2.3) was introduced in [21]). There are also other methods for solving BSS or PP, but this thesis concentrates on independent component analysis.

The sources $\mathbf{s}$ and the mixing matrix $\mathbf{A}$ can be solved only up to *permutation* and *scaling*, since if $\mathbf{s}$ is a vector of independent random variables then $\mathrm{diag}(\mathbf{d})\mathbf{P}\mathbf{s}$, where $\mathrm{diag}(\mathbf{d})$ is a diagonal scaling matrix and $\mathbf{P}$ a permutation matrix, is also a vector of independent random variables, which makes it impossible to distinguish between different permutations and scalings. Additionally, if the sources do not have zero mean, the $\mathbf{x}$ has to be zero-meaned and the mean has to be estimated separately.

[1] From now on s or s_i refers to one source. It can either be the random variable or the time series from it. If distinction is needed we use s and $s(t)$. $\mathbf{s}$ refers to a column vector of the sources, which is either a vector of random variables or a vector of time series. Again $\mathbf{s}(t)$ is used for distinguishing the time series from the random variables

[2] See Appendix A for a short introduction to some statistical concepts such as independence.



(a) (b)

Figure 2.1: *Two uniformly distributed random variables (a) and their mixtures (b). This example is presented by Comon in [21]*

The definition in (2.3) is the simplest definition of the ICA problem. Two more general definitions are given in [41], one for a noisy ICA and the other for more general data analysis, where a projection is searched giving as independent sources as possible, according to some measure of independence.

Example 3 *Let us consider a simple theoretic example of ICA problem as in [21]: Let there be two mutually independent uniformly distributed random variables s_1 and s_2 producing a vector $\mathbf{s} = [s_1\, s_2]^T$. Finite realisation of the $\mathbf{s}$ is shown in Fig. 2.1a. $\mathbf{x}$ in the fig. 2.1b is obtained by multiplying with $\mathbf{A}$ according to (2.3). The problem is now to find back the independent directions. In this case they are quite trivial and are marked with shorter dashed lines in fig. 2.1b.*

If in the example, the random variables $\mathbf{x}$ are known, we can calculate $\mathbf{W}$ and $\mathbf{s}$ using the fact that if two random variables s_1 and s_2 are independent their joint density becomes $f_{s_1 s_2}(s_1, s_2) = f_{s_1}(s_1) f_{s_2}(s_2)$. The problem can then be seen as a minimisation of the distance between the joint density $f_{\mathbf{s}_1\mathbf{s}_2}$ and the marginal densities $f_{\mathbf{s}_1}$ and $f_{\mathbf{s}_2}$ of the source variables or in another words $d(f_{\mathbf{s}_1\mathbf{s}_2}, f_{\mathbf{s}_1} f_{\mathbf{s}_2})$, where d is some proper distance function. Usually, however, we do not know $\mathbf{x}$ or $f_{\mathbf{x}}$ explicitly. Instead we have some measurement of the linear process, say $\mathbf{x} = \mathbf{x}(t)$. From that we should estimate the distribution of $\mathbf{x}$ i.e. $\mathbf{f_x}$. Estimation of the $\mathbf{f_x}$ however, is a very difficult problem, especially when the $\mathbf{x}$ is high-dimensional. Therefore other methods will have to be developed to solve the ICA-problem.

In the next chapter the ICA problem is solved without estimating the density functions $f_{\mathbf{s}_1\mathbf{s}_2}$, $f_{\mathbf{s}_1}$ and $f_{\mathbf{s}_2}$ using another kind of optimisation scheme.

Chapter 3

A solution to ICA

In this chapter, we redefine the ICA problem as an optimisation problem, giving the independent components as its solutions. The objective functions used in the optimisation are called *contrast functions.* But first, the data is decorrelated using a classical linear projection method, *whitening*, because decorrelation is a necessary condition for independence. Hence, it will bring us a step closer to our ultimate objective.

3.1 Whitening

A zero-mean random vector $\mathbf{v}$ is said to be *white* if its covariance matrix $C = E\{\mathbf{v}\mathbf{v}^T\} = \mathbf{I}$, which means all the components are *decorrelated* and have unit variance. The process leading from $\mathbf{x}$ to this is called *whitening* and can be achieved in several ways, the simplest way being perhaps the *eigenvalue decomposition* (EVD) of the covariance matrix. Then the whitened data $\mathbf{v}$ is given by

$$\mathbf{v} = \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x}, \tag{3.1}$$

where $\mathbf{E} = [\mathbf{e}_1\, \mathbf{e}_2\, \cdots \mathbf{e}_l]$ consists of the eigenvectors of the covariance matrix, $cov\{\mathbf{x}\}$ and $\mathbf{D} = diag(\lambda_1\ \cdots\ \lambda_l)$ is a diagonal matrix of corresponding eigenvalues, l being the number of non-zero eigenvalues. Finally, $\mathbf{v}$ is called the *whitened data vector.* The normalisation to unit variance ie. $D^{-1/2}$, and the projection into the decorrelated basis i.e. $\mathbf{E}^T$ can be combined into a *whitening matrix* $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T$. With this matrix, we can present the *whitened ICA problem*:

$$\mathbf{v} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s} = \mathbf{B}\mathbf{s}, \tag{3.2}$$

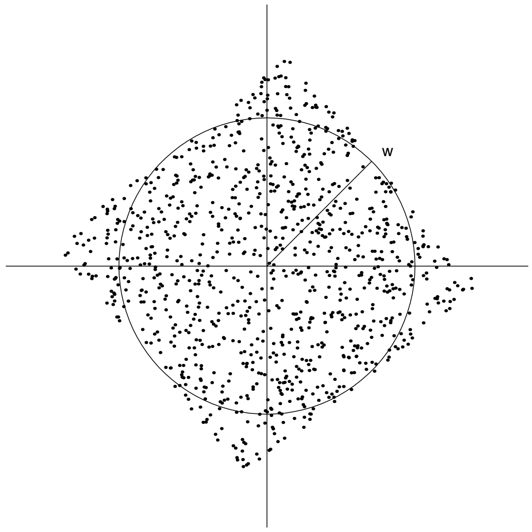


Figure 3.1: *Whitened data*

where the matrix $\mathbf{B}$ is called the *whitened mixing matrix.* The whitened inverse form is defined somewhat differently from (2.4) for further purposes:

$$\mathbf{s} = \mathbf{W}\mathbf{v} = \mathbf{W}\mathbf{V}\mathbf{x}, \tag{3.3}$$

This whitened ICA problem is solved in the next section. The original mixing matrix $\mathbf{A}$ can then be calculated from $\mathbf{W}$ using dewhitening matrix $\mathbf{V}_d = \mathbf{V}^{-1}$. If $\mathbf{V}$ is not square, the pseudoinverse is calculated so that $\mathbf{V}_d\mathbf{V} = \mathbf{I}$.

Example 4 *To continue our example, we whiten the* mixtures *in the previous example. The whitened data is shown in Fig. 3.1.*

Note that if $\mathbf{v}$ is decorrelated, then any *orthonormal* transformation of it, say $\mathbf{O}\mathbf{v}$ is also decorrelated, since

$$E\{\mathbf{O}\mathbf{v}(\mathbf{O}\mathbf{v})^T\} = \mathbf{O}E\{\mathbf{v}\mathbf{v}^T\}\mathbf{O}^T = \mathbf{O}\mathbf{I}\mathbf{O}^T = \mathbf{I}.$$

Even any *orthogonal* transformation is decorrelated, though then $\mathbf{O}\mathbf{O}^T \neq \mathbf{I}$, but rather it is a diagonal matrix. Additionally, $\mathbf{O}\mathbf{v}$ can be white only if $\mathbf{O}$ is an orthonormal transformation. So we can rotate the axes, which corresponds to an orthonormal transformation, to find the independent directions $\mathbf{W}$. One criterion, for finding the proper orthonormal transformation, can be motivated by the so called *central limit theorem.*

3.2 Central limit theorem and non-Gaussianity

The *central limit theorem* (see e.g. [85]) states that the distribution of (weighted) sum of *independent* random variables approaches a normal distribution when the number of variables approaches infinity. Let us assume that this holds not only asymptotically but monotonically, i.e. a (weighted) sum of $n+1$ independent variables is more Gaussian than the (weighted) sum of n independent variables. Then, ultimately, the distribution of the least Gaussian independent variable should be less Gaussian than any (weighted) sum of the independent variables. We can, thus, solve (3.2) (or more precisely its whitened counterpart of the inverse form (2.4)) by finding the least Gaussian distributions constructable by linear combinations of $\mathbf{x}$. This of course limits $\mathbf{s}$ to be non-Gaussian[1].

Cumulants

Cumulants (see e.g. [58]) present one way to measure non-Gaussianity. They are defined as follows:

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

$$\log \hat{f}(t) = \kappa_1(it) + \frac{\kappa_2(it)^2}{2} + \dots + \frac{\kappa_r(it)^r}{r!} + \dots, \tag{3.4}$$

the Taylor coefficients κ_r define the cumulants (of the distribution) of x. The first two cumulants are the *mean* and the *variance*

$$\kappa_1 = E\{x\} = 0, \tag{3.5}$$

$$\kappa_2 = E\{x^2\}. \tag{3.6}$$

They define perfectly a Gaussian distribution. The rest of cumulants are zero for Gaussian distributions, but may be non-zeroes for non-Gaussian distributions. For that reason they can be used as non-Gaussianity measures. The third cumulant, defined as

$$\kappa_3 = E\{x^3\}, \tag{3.7}$$

[1]Solving one Gaussian source among non-Gaussians is possible: if e.g. the sources are solved one-by-one, the Gaussian source is what is left over, when all the other $n-1$ sources have been solved.

is called *skewness*. 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 the *kurtosis* [58, 70], and is defined

$$\text{kurt}\,(x) = \kappa_4 = E\{x^4\} - 3(E\{x^2\})^2. \tag{3.8}$$

For whitened random variable v, the kurtosis becomes $\text{kurt}\,(v) = E\{v^4\} - 3$, since then $E\{v^2\} = 1$.

The higher than fourth order cumulants are not named, since they become too complex to analyse. Ideally, all the cumulants should be used to measure non-Gaussianity, but that would lead to impossible calculations. As it happens, in many cases it is enough to use kurtosis as the non-Gaussianity measure for ICA[2].

3.3 Generalising the non-Gaussianity measures

Non-normality measures can be generalised using the scheme in [40]: Let G be twice differentiable scalar function to be defined, x a zero mean random variable, normalised to unit variance and ν a zero mean Gaussian variable, with unit variance also. Then one class of contrast functions can be defined as

$$J_G(x) = |E\{G(x)\} - E\{G(\nu)\}|^p. \tag{3.9}$$

G should not be a polynomial of second order, since then J_G would always be equal to zero. Furthermore, the modulus is taken since for some functions G, and some non-Gaussian distributions $G(x) < G(\nu)$, which would lead to minimisation problem whereas for other distributions $G(x) > G(\nu)$ leading to maximisation. So by taking the modulus we can restrict the problem to finding the maximum of J_G. The exponent $p \in [1, 2]$ usually.

Note that taking $G(u) = u^4$ means using of *kurtosis* for whitened data. In [40] it is shown that for whitened data, kurtosis is a proper contrast function, i.e. its maximum gives the most kurtotic or *supergaussian* independent source and its minimum gives the least kurtotic or *subgaussian* independent source. Another important contrast function can be derived for $G(u) = ln\, cosh(u)$, which is shown in [43] to have good behaviour in terms of robustness and asymptotic variance. In [40] some other contrast functions are also introduced.

[2] See the end of the next section for elaboration.

3.4 Fixed point algorithms for ICA

Consider the whitened data $\mathbf{v}$ in example 4, shown in Fig. 3.1. Let us begin by solving the ICA problem for one independent source. Our task is to find a direction $\mathbf{w}$, which gives one row of the filtering matrix $\mathbf{W}$, in which $J_G(\mathbf{w}^T\mathbf{v})$ has its maximum. If our contrast function is a proper contrast function, $\mathbf{w}^T\mathbf{v}$ would be one independent component. The direction $\mathbf{w}$ could be found e.g. by gradient descent over J_G(see e.g. [10]). This is, however, often rather slow and more suitable methods have been developed by various people in the field. In the next chapter some of those methods are listed with brief explanations, but first the algorithm used in this work is presented.

This method for ICA, developed jointly by Hyvärinen and Oja, was published originally in [46, 40]. The optimisation problem is to maximise

$$J_G(\mathbf{w}^T\mathbf{v}) = E\{G(\mathbf{w}^T\mathbf{v})\} - E\{G(\nu)\}, \tag{3.10}$$

in which we have for simplicity assumed that p equals to 1 and that the component with greatest absolute value of J_G is positive. The optimisation is under constraint

$$h(\mathbf{w}) = ||\mathbf{w}||^2 - 1 = 0. \tag{3.11}$$

In the particular case of kurtosis J_G becomes

$$J_G(\mathbf{w}^T\mathbf{v}) = E\{(\mathbf{w}^T\mathbf{v})^4\} - 3||\mathbf{w}||^4. \tag{3.12}$$

The necessary conditions for optimum according to Kuhn-Tucker [64] imply

$$E\{\mathbf{v}g(\mathbf{w}^T\mathbf{v})\} - \beta\mathbf{w} = 0, \tag{3.13}$$

where $\beta = E\{\mathbf{w}_0^T\mathbf{v}g(\mathbf{w}_0^T\mathbf{v})\}$, g is used to notate the derivative of G and $\mathbf{w}_0$ is the value of $\mathbf{w}$ in the optimum. The equality is solved using a *Newton iteration method* [64] described by Hyvärinen in [42]. The Jacobian of the left hand of (3.13), denoted by F, is given by

$$\mathrm{JF}(\mathbf{w}) = E\{\mathbf{v}\mathbf{v}^T g'(\mathbf{w}^T\mathbf{v})\} - \beta\mathbf{I} \tag{3.14}$$

To simplify the inversion of this matrix, we decide to approximate the first term on the right in (3.14). Since the data is sphered, a reasonable approximation seems to be $E\{\mathbf{v}\mathbf{v}^T g'(\mathbf{w}^T\mathbf{v})\} = E\{\mathbf{v}\mathbf{v}^T\}E\{g'(\mathbf{w}^T\mathbf{v})\} = E\{g'(\mathbf{w}^T\mathbf{v})\}\mathbf{I}$. Thus the Jacobian matrix becomes diagonal, and can easily be inverted. β can be approximated using the current value of $\mathbf{w}$ instead of $\mathbf{w}_0$. Thus we obtain the following approximative Newton iteration:

$$\begin{aligned} \mathbf{w}^+ = \mathbf{w} - [E\{\mathbf{v}g(\mathbf{w}^T\mathbf{v})\} - \beta\mathbf{w}]/[E\{g'(\mathbf{w}^T\mathbf{v})\} - \beta] \\ \mathbf{w}^* = \mathbf{w}^+/\|\mathbf{w}^+\|, \end{aligned} \tag{3.15}$$

where $\mathbf{w}^*$ denotes the new value of $\mathbf{w}$, $\beta = E\{\mathbf{w}^T\mathbf{v}g(\mathbf{w}^T\mathbf{v})\}$ and the normalization has been added to improve the stability. This algorithm can be further simplified by multiplying both sides of the first equation in (3.15) by $\beta - E\{g'(\mathbf{w}^T\mathbf{v})\}$ giving

$$\mathbf{w}^+ = \frac{E\{\mathbf{v}g(\mathbf{w}^T\mathbf{v})\} - E\{g'(\mathbf{w}^T\mathbf{v})\}\mathbf{w}}{\beta - E\{g'(\mathbf{w}^T\mathbf{v})\}} \tag{3.16}$$

$$\mathbf{w}^* = \mathbf{w}^+/\|\mathbf{w}^+\|. \tag{3.17}$$

The denominator, $\beta - E\{g'(\mathbf{w}^T\mathbf{v})\}$, in (3.16) is scalar i.e. a normalisation. It can be omitted, because the w^+ is normalised explicitly in (3.17). This gives the following *fixed point algorithm*[3]

$$\begin{aligned} \mathbf{w}^+ = E\{\mathbf{v}g(\mathbf{w}^T\mathbf{v})\} - E\{g'(\mathbf{w}^T\mathbf{v})\}\mathbf{w} \\ \mathbf{w}^* = \mathbf{w}^+/\|\mathbf{w}^+\| \end{aligned} \tag{3.18}$$

One common convergence problem in Newton optimisation methods is an oscillation around a solution. This can be avoided by adding a learning rate to (3.15), giving the *stabilised fixed point algorithm*

$$\begin{aligned} \mathbf{w}^+ = \mathbf{w} - \mu[E\{\mathbf{v}g(\mathbf{w}^T\mathbf{v})\} - \beta\mathbf{w}]/[E\{g'(\mathbf{w}^T\mathbf{v})\} - \beta] \\ \mathbf{w}^* = \mathbf{w}^+/\|\mathbf{w}^+\| \end{aligned} \tag{3.19}$$

where $\beta = E\{\mathbf{w}^T\mathbf{v}g(\mathbf{w}^T\mathbf{v})\}$ as above, and μ is a step size parameter that may change with the iteration count. If the oscillation is around two points, which is quite often the case, it is usually enough to set the $\mu = 0.5$ for one iteration, when the oscillation is encountered and otherwise to use $\mu = 1.0$, ie original learning rule in (3.18). If

[3] The name fixed point comes from historical reasons, since the algorithm was first derived using fixed point iteration method [64].

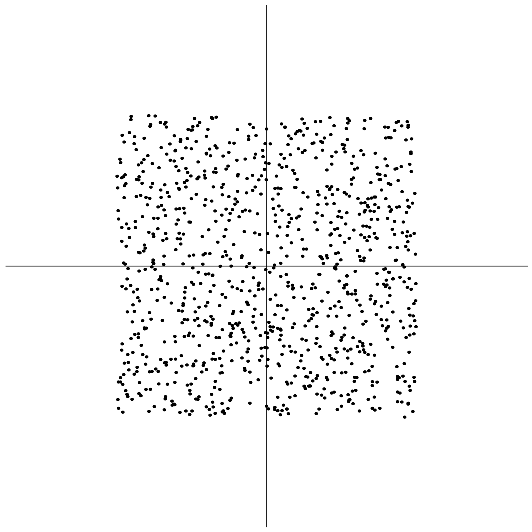


Figure 3.2: *Independent basis*

the oscillation is between more solutions, this solution may not work. Note that the stabilised fixed point algorithm is still a *proper* ICA-algorithm, because the optimum of (3.19) is also the optimum of (3.15) since the update is only multiplied with a constant.

To extract several independent components, we could simply run several times the one unit fixed point algorithm. But then we could[4] hit the same component over and over again, neglecting the others. This can be avoided by keeping the component $\mathbf{w}_k$ that is being estimated, decorrelated from the components already extracted. i.e. $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{k-1}$. This can be done using the Gram-Schmidt orthogonalisation:

$$\mathbf{w}_k = \mathbf{w}_k - \mathbf{W}\mathbf{W}^T\mathbf{w}_k, \tag{3.20}$$

after each iteration. The $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_l)$ is the whitened filtering matrix and $\mathbf{w}_i = 0, \text{for}\, i > k$ i.e. if the component i has not been calculated yet.

Example 5 *To conclude our basic example of ICA involving two uniformly distributed independent sources, we use the fixed point learning rule in (3.18) to solve the independent components from the whitened data in example 4. The result is shown in Fig. 3.2. By very close comparison of Figs. 2.1a and 3.2 it can be seen, that the the estimated sources agree with the original sources up to a sign change in the first independent component.*

It is also possible to construct a fixed point algorithm for extraction of all components at the same time i.e. *symmetrically.* We proceed by running n deflational fixed point

[4] We even should, if the global optimum is reached.

algorithms parallely. At each step of iteration, all the outputs $\mathbf{w}_1^T\mathbf{v}, ...\mathbf{w}_l^T\mathbf{v}$ have to be decorrelated in order to prevent them to converge to same solution. This can be done using a classical method involving matrix square roots by letting

$$\mathbf{W} = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1/2}, \tag{3.21}$$

or using some other methods (two others are presented in [42]).

The fixed point algorithms presented here are collected into a package called FastICA [28], which is a public domain software. The name of the package will be used instead of the fixed point algorithm from now on in this thesis.

Example 6 *Now let us consider a multidimensional ICA problem with independent sources not having the same distributions. Fig. 3.3, which is adopted with minimal changes from [98], presents step-by-step the flow of the algorithm from mixtures to IC's.*

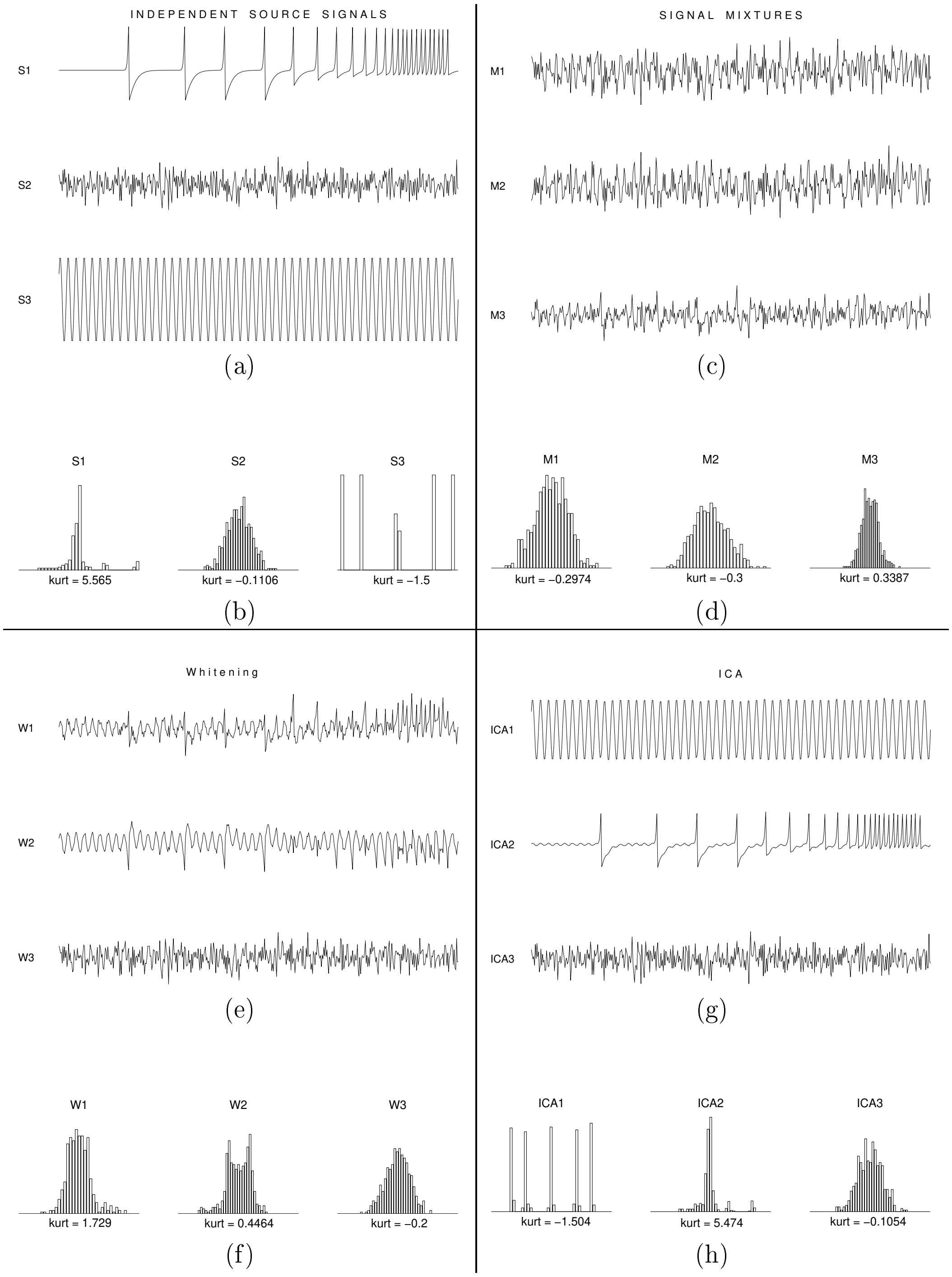


Figure 3.3: *Examples of independent source signals with positive, zero and negative kurtosis values (a); their amplitude distributions (b), and the respective value for the kurtosis below this last figure. The mixtures of the previous signals (c) have near to zero kurtosis and close to Gaussian distributions (d). From whitened (e) and (f) to ICA (g) and (h), the resulting time series have clearer negative, positive or zero kurtosis values, close to those of the initial independent source signals. Adopted from [98].*

Chapter 4

Some other methods for solving ICA

When thinking about other methods of solving ICA, we may, as suggested in [41], separate the *objective function*, i.e. the criteria used to obtain the independent basis from the *optimisation algorithm*, i.e. the computational way so solve the optimisation of the objective function. This can be presented as an equation scheme:

$$\textit{ICA method} = \textit{Objective function} + \textit{Optimisation algorithm}.$$

The objective function defines the statistical properties, such as consistency, asymptotic variance and robustness; and the optimisation algorithm defines the algorithmic properties, such as convergence speed, memory requirements and numerical stability.

In some cases it is easy to separate the Objective function from the Optimisation algorithm, as in the FastICA, where

$$\textit{FastICA} = \textit{Contrast function } J_G + \textit{Newton iteration method,}$$

in other cases they can be quite indistinguishable. Keeping this in mind, we try to introduce other main ICA methods, separating the contrast functions from the optimisation when possible.

4.1 Jutten-Hérault algorithm

The pioneering algorithm in ICA is the one by Jutten and Hérault [51]. It is based on canceling the non-linear cross-correlations between the estimated independent signals

$E\{g_1(\hat{s_i})g_2(\hat{s_j})\}$, where g_1 and g_2 are some suitably chosen non-linearities. If the sources are independent, these cross-correlations are zero. The update rule for the algorithm is

$$\begin{aligned} \Delta\mathbf{W}_{ij} \propto g_1(y_i)g_2(y_j), \text{for } i \neq j, \\ \Delta\mathbf{W}_{ij} = 0, \text{for } i = j, \end{aligned} \tag{4.1}$$

where the y_i are computed at each iteration according to $\mathbf{y} = (\mathbf{I}+\mathbf{W})^{-1}\mathbf{x}$. The algorithm could be described with a scheme

$$\begin{aligned} \textit{Jutten-Hérault} = \quad & \textit{Non-linear cross-correlation between the sources} \\ & + \textit{Recurrent network update algorithm,} \end{aligned}$$

though the objective function can not be defined explicitly. The Jutten-Hérault algorithm has been a fruitful basis for other algorithms (see e.g. [20, 19, 17]). In [59, 15] an algorithm called EASI was introduced:

$$\Delta\mathbf{W} \propto (\mathbf{I} - \mathbf{y}\mathbf{y}^T - g(\mathbf{y})\mathbf{y}^T + \mathbf{y}g(\mathbf{y}^T))\mathbf{W}. \tag{4.2}$$

4.2 The bigradient algorithm

In this algorithm by Wang and Karhunen [101, 102], the contrast function consists of two parts: the sum of kurtoses and a penalising term for orthogonalisation. This of course works only if all the sources are super-Gaussian (or sub-Gaussian), since in general $|\alpha_1 + \alpha_2 + ...| \neq \sum_i |\alpha_i|$. Then the contrast function is

$$J(B) = \underbrace{J_\Sigma(B)}_{sum\,of\,kurtoses} + \rho \underbrace{||\mathbf{I} - \mathbf{B}^T\mathbf{B}||_F^2}_{orthogonalisation}, \tag{4.3}$$

where $\rho \in \mathbb{R}$ is a free parameter adjusting the strictness of orthogonalisation and $||\mathbf{X}||_F$ denotes the *Frobenius norm* of matrix $\mathbf{X}$. This is then minimised with a gradient descent giving

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \mu_k E\{\mathbf{v}_k(\mathbf{v}_k^T\mathbf{B}_k)^3\} + \gamma_k[\mathbf{B}_k\mathbf{I} - \mathbf{B}_k^T\mathbf{B}_k], \tag{4.4}$$

where $\gamma_k = \mu_k\rho_k$, and the power of 3 is applied termwise. Instead of kurtosis, we can of course use the generalised contrast function in (3.9). Then (4.4) takes a form

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \mu_k E\{\mathbf{v}_k g(\mathbf{v}_k^T \mathbf{B}_k)\} + \gamma_k [\mathbf{B}_k \mathbf{I} - \mathbf{B}_k^T \mathbf{B}_k]. \tag{4.5}$$

To summarise, we present the scheme of the bigradient algorithm:

Bigradient $=$ *Sum of kurtoses with penalty for orthogonalisation* $+$ *Gradient descent.*

4.3 Non-linear PCA

Non-linear extensions of the former neural PCA-algorithms were developed in [76]. PCA stands for principal component analysis and is a way to explain a multidimensional random variable with less dimensions and still preserve the maximum amount of variance possible. PCA is discussed in more detail in section 5.1.1.

Oja has shown in [71] that a neural formulation of the so called principle of *Hebbian learning* [37] leads to PCA. The Hebbian rule states that

> when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells, such that A's efficiency, as one of the cells firing B, is increased.

In mathematical notation this statement means that the update of the weight between two axon cells or *neurons* is a function of their cross-correlation, therefore the final symmetric version of the neural PCA-network is obtained by the following learning rule

$$\Delta \mathbf{w}_i \propto \underbrace{y_i \mathbf{x}}_{\textit{Hebbian term}} - \underbrace{\mathbf{y}_i \sum_{j=1}^{n} \mathbf{y}_j \mathbf{w}_j}_{\textit{Orthogonalising feedback}} . \tag{4.6}$$

Non-linear versions of this rule were developed by replacing the $\mathbf{y}_i$ by a non-linear function $g(\mathbf{y}_i)$. This was done originally to make the learning more robust against outliers [53, 77, 78], but later it was found out [54, 74] that it performs ICA if the g is chosen properly. This gives a scheme

Non-linear PCA $=$ *Non-linear PCA criterion* $+$ *Gradient descent.*

4.4 Maximisation of the entropy

Information theory (see appendix A.2) provides perhaps the most natural environment to the definition of ICA-criteria. An example of an algorithm from that class is that by Bell and Sejnowski [7], where a simple gradient ascent was used to maximise the total entropy of the network outputs:

$$\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} - 2\tanh(\mathbf{W}\mathbf{x})\mathbf{x}^T. \tag{4.7}$$

The tanh-function is applied separately on each element of the $\mathbf{Wx}$-vector. This, however, converges very slowly. Better working algorithm is the one obtained by *natural gradient descent* [4, 5] on the same objective function:

$$\Delta \mathbf{W} \propto (\mathbf{I} - 2\tanh(\mathbf{y})\mathbf{y}^T)\mathbf{W} \tag{4.8}$$

with $\mathbf{y} = \mathbf{W}\mathbf{x}$. It converges much faster and is better conditioned. To summarise

Network entropy maximisation = Total entropy of the outputs + (Natural) gradient descent.

The algorithms presented in this chapter are far from exhaustive list of all the algorithms used to solve ICA. As a tribute to the other people in the field, we will briefly name them together with a representative reference. At the end, some interesting similarities are noted between different solutions.

- Jutten and Hérault [51]
- Oja, Ogawa and Wangviwattana [76]
- Pham, Garrat and Jutten [88].
- Karhunen and Joutsensalo [53]
- Comon [21]
- Cichocki, Kasprzak and Amari [18]
- Delfosse and Loubaton [24]
- Deco and Brauer [23]
- Moreau and Macchi [67]

- Pajunen, Hyvärinen and Karhunen [83]
- Amari, Cichocki and Yang [4]
- Cardoso and Laheld [15]
- Cichocki and Unbehauen [16]
- Karhunen, Oja, Wang, Vigário and Joutsensalo [54]
- Karhunen and Pajunen [55]
- Girolami and Fyfe [32]
- Pearlmutter and Parra [86]
- Pajunen [82]
- Lappalainen [60]

Since independence is practically an information theoretic concept, it could be argued that every algorithm performing ICA should have its roots in information theory. Indeed, not surprisingly, e.g. FastICA can be derived from the concept of *mutual information* (see (A.8) for definition) by making some approximations [42]. It has also been shown e.g. in [14, 87] that under certain assumptions the maximisation of differential entropy in [7] is equivalent to likelihood maximisation in [88]. Furthermore it has been shown that also non-linear PCA is equivalent to likelihood maximisation in [56, 75]. The field is still waiting for a framework, where every algorithm is derived from a common basis, e.g. mutual information, since one should exist.

Chapter 5

Real-life problems in applying ICA

5.1 Over- and under-complete bases

The FastICA and the other algorithms in the previous chapter perform well, when the number of mixtures equals the number of sources ($m = n$) i.e. when the mixing matrix $\mathbf{A}$ is square. What happens, when there are more mixtures than sources ($m > n$, under-complete basis, or overdefined problem) or when there are less mixtures than sources ($m < n$, over-complete basis, or underdefined problem)?

5.1.1 Under-complete basis

Intuitively it should be no problem to have more mixtures than sources. If we e.g. consider the cocktail party problem presented in example 1, it just means that we can measure with more microphones than we have speakers. Rather, this is an asset, if the data has some noise. How will the algorithms perform then?

Example 7 *Consider the example 6 slightly modified: Instead of taking as many mixtures as there are sources* (4)*, take as many mixtures as there are samples* (500)*. In order to avoid zero eigenvalues, add some amount of noise (even very small) in each mixture. Consider the FastICA with kurtosis as an example*[1]*. The ICA equation in (2.4) becomes*

$$\begin{aligned} \mathbf{s}(t) &= \mathbf{W}\tilde{\mathbf{x}}(t), \\ \tilde{\mathbf{x}}(t) &= \mathbf{x}(t) + \mathbf{n}(t), \textit{and } var(\mathbf{n}) \ll var(\mathbf{x}) \end{aligned} \tag{5.1}$$

[1]The results are effectively the same with other algorithms and contrast functions as shown in [47]

where in the time series $\mathbf{x}(t)$ *there are as many mixtures as there are samples. This allows building of* any $\mathbf{s}$ *from* any $\mathbf{x}$ *by changing the values of* $\mathbf{W}$.

The kurtosis of a signal, with non-zero variance, is maximised, when the signal is an impulse i.e. it has some fixed value at one point and is constant everywhere else. To partially prove this intuitive result, we show that the kurtosis of an impulse function is larger than the kurtosis of a signal with two peaks.

Proof Let $x_1(t) = [x_1(1)\, x_1(2)\, \cdots\, x_1(T)], T \in \mathbb{N}$ be an impulse signal, with zero mean and unit variance; and $x_2(t)$ a signal with two peaks, as well with zero mean and unit variance, defined as

$$
\begin{aligned}
x_1(t) &= \begin{cases} \sqrt{T-1} & \text{, if } t = t_1 \\ -\frac{1}{\sqrt{T-1}} & \text{, otherwise} \end{cases} \\
x_2(t) &= \begin{cases} \sqrt{\frac{T-2}{2}} & \text{, if } t = t_2 \neq t_3 \text{ or } t = t_3 \neq t_2 \\ -\sqrt{\frac{2}{T-2}} & \text{, otherwise} \end{cases}
\end{aligned}
\tag{5.2}
$$

Using (3.8) we get

$$
\begin{aligned}
\mathrm{kurt}(x_1) - \mathrm{kurt}(x_2) &= E\{x_1^4\} - 3\underbrace{E^2\{x_1^2\}}_{=1} - (E\{x_2^4\} - 3\underbrace{E^2\{x_2^2\}}_{=1}) \\
&= E\{x_1^4\} - E\{x_2^4\}
\end{aligned}
\tag{5.3}
$$

Implementing (5.2) and calculating sample expectations, we get

$$
\begin{aligned}
&= \frac{1}{T}\{\sqrt{T-1}^4 + (T-1)(-\frac{1}{\sqrt{T-1}})^4 - (2\sqrt{\frac{T-2}{2}}^4 + (T-2)(-\sqrt{\frac{2}{T-2}})^4)\} \\
&= \frac{1}{T}\{(T-1)^2 + \frac{T-1}{(T-1)^2} - \frac{2(T-2)^2}{4} - \frac{4(T-2)}{(T-2)^2}\} \\
&= \frac{1}{T}\{(T-1)^2 - \frac{(T-2)^2}{2} + \frac{1}{T-1} - \frac{4}{T-2}\}.
\end{aligned}
\tag{5.4}
$$

The last equation becomes smaller if $T-1$ is replaced by $T-2$, since $1^2 > \frac{1}{T-2}$, giving

$$
> \frac{1}{T}\{\frac{(T-2)^2}{2} - \frac{3}{T-2}\} > 0, \text{ for all } T > 3. \tag{5.5}
$$

For $T = 3$, ie. when the signals only have three points, the densities are symmetric versions of each others, that is $f_{x_1}(x) = f_{x_2}(-x)$. This concludes our partial proof.

Example continued *The FastICA-algorithm maximises the kurtosis of the sources. Furthermore any kind of a source can be built using (5.1). Hence, we will only get spike-like artefactual solutions rather than the real sources!*

This phenomenon, presented in its extreme form in the previous example, is not unlike to a common problem in neural network learning called *over-fitting* or *over-learning* [10, 36]. The problem can arise even if the number of sources would be smaller than the number of samples, and even if it would seem much smaller, in particular if the data is low pass filtered. To read more about these real-life cases, see [47].

One solution to this problem is to reduce the dimension of the data to increase the relative data length compared to its dimension. The number of sources can be estimated quite reliably (see [103, 61, 100, 6]), but this leaves us still with a problem: How to determine the subspace the data should be projected into? The classical solution to this problem is the Principal Component Analysis (PCA).

Principal Component Analysis

The objective of principal component analysis, sometimes also referred as the (discrete) Karhunen-Loève transform or the Hotelling transform, is to explain a multidimensional random variable with fewer dimensions, retaining the maximum amount of variance possible. There are many ways to do this, ranging from neural (these were already discussed in section 4.3) to pure statistical (for an introduction see e.g. [71, 72, 73]). Interestingly, but not surprisingly, one way to do PCA is to make use of the eigenvectors of the covariance matrix with largest eigenvalues. This of course because of the fact that the eigenvalues *are* the variances of the random variable in the directions of the eigenvectors.

By reducing the dimension of the mixed data x using PCA, we can reduce the dimensionality of possible solutions thus, hopefully, excluding the spike-like artefactual solutions. And we still retain the maximum possible variance.

5.1.2 Over-Complete basis

The problem of ICA in an over-complete basis is much harder to solve, but it is possible under some additional assumptions. Recently some algorithms have been introduced (e.g. [81, 84, 44, 62]). In the case of this thesis, we assume that either there are no more independent sources than sensors, or even if there would be, the most independent sources are still estimated quite accurately, and they are the most interesting ones.

5.2 Correlated sources

If the sources are correlated, they are not independent. But in the light of the flexible third definition of ICA in chapter 2 on page 7, we should be able to construct some scheme for finding such an interesting transformation, that would reveal the sources. One way, perhaps crude, but useful in practice, to achieve this is to use the deflational FastICA, without the restriction of decorrelation. This can be achieved by not decorrelating the sources after one solution is found. Then, of course, one might hit the same solution over and over again, but we may hope that using different initial conditions will do the trick.

5.3 Noisy ICA

All the existing algorithms presented in section 4, including FastICA, start to perform poorly, if there is noise present[2]. Recently some algorithms have been introduced for solving this problem too. (See e.g. [45, 44, 26] for recent papers and [69, 79, 8, 25, 80] for older ones. For discrete-valued sources, geometrically motivated algorithms [89, 81] are also possible.) We adopt a simpler, though of course not so robust procedure: the noise will be reduced by PCA, so that the estimation of the solutions becomes possible.

[2]Noise here means measuring errors in sensors as well as non-linearities in the mixing process.

Part II

Brain research, Magnetoencephalography and ICA

Chapter 6

Brain research

This chapter of the principles of brain research begins our experimental part of the thesis. First there will be an introduction to central issues of contemporary brain research and then we discuss in more detail the sensory systems studied in the last three experiments. In the appendix B there is a small glossary of brain research terms for those not familiar with the field.

Brain research is, of course, such a vast savannah of research that it is not possible, in the scope of this kind of methodological thesis, to cover it very densely. This in mind, we divide the brain research into two, admittingly overlapping and even non-exhaustive, fields:

1. Clinical brain research, and
2. Cognitive brain research.

Clinical brain research comprises all the research that aims at understanding better the anatomy and the functioning of the brain in order to cure or to prevent diseases or abnormal behaviour, such as mental disorders. Acknowledging the utter importance of clinical brain research, we anyhow, in our experiments, concentrate on the latter field of our division, the cognitive brain research. This does not, in anyway, mean that ICA has nothing to give to clinical brain research, on the contrary: many of the problems we hope to solve in the framework of cognitive brain research are present as such in clinical brain research.

Cognitive brain research can be divided further into 11 different subcategories as done in [27]:

1. molecular and cellular plasticity,
2. neural and psychological development,

3. sensory systems,
4. strategies and planning: motor systems,
5. attention,
6. memory,
7. language,
8. thought and imagery,
9. emotion,
10. evolutionary perspectives, and
11. consciousness

Molecular and cellular plasticity define the basics of the chemistry of the brain functions. It constitutes the basic building-block mechanisms on which perception, memory and cognition in general are based. *Neural and psychological development* deal with the questions on how the brain is developed and organised. This field has a close connection to the first subcategory.

The third subcategory, *sensory systems* studies perception and is discussed in more detail in section 6.2, since it is the field, in which our experiments are concentrated. The fourth subcategory, the *motor systems*, concentrates on studying the planning and strategies of movement and finding the neuronal basis for them. *Attention* is closely related to the two former subcategories since it decides what elements of sensorial information come into consciousness and cause some action possibly containing movement.

The sixth subcategory, *language*, might at first seem separate from the former three subcategories, but closer inspection shows otherwise: without movement of lips and tongue we can not speak and without sensory systems and attention mechanism we would not know who to speak to. Language is a very essential part of human cognition since it is the most important way for us to communicate between each other. For that fact, language is very much studied in cognitive neurosciences.

Thinking is perhaps the one subject that has puzzled human beings most, already for several thousands of years. The central subject in research of thought has without a doubt been reasoning and inference. The relation between thinking and language is highly controversial among philosophers and cognitive neuroscientists: some say that there could not be the second without the first, but others claim otherwise. Within this thesis we can not take part into this discussion, as interesting as it is.

Emotion has been very central research field around the last turn of the century and continued flourishing until the midcentury. After that, however, the brain research strongly concentrated on other parts of cognitive brain research – partly for trends and funding reasons, partly because emotion has been a very difficult subject to investigate. For some decades motions were not even included in *cognitive* processes. Nowadays, nevertheless, there has been a growing interest in the problem of emotion and its neuronal foundation.

Evolutionary perspectives discuss what evolutionary biology could give to cognitive neurosciences due to genetical involvement in the building of neural architecture. It also explores the possible differences in human versus other mammalian brain structures due to different evolutionary backgrounds and can thus help to understand, what inferences can and can not be made from other mammalian studies.

The last subcategory concentrates on possibly the toughest subject to cognitive neurosciences: the *consciousness*. Even the term is quite vague. Anyhow, the novel measuring techniques, combined with research done on brain-damaged subjects have given very interesting insights to the human consciousness.

For readers with more interest in the cognitive neurosciences, we most sincerely recommend the book editored by Gazzaniga [27]: it gives a very thorough view of the subject.

6.1 Brain researching techniques

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. By studying people with brain-functional disorders in autopsy have given us knowledge about brain parts essential to a specific function. 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 has brought many new techniques for investigating the anatomy and functioning of the human and animal brains. Of these techniques the oldest one, *electroencephalography* (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 potential is measured by electrodes on the scalp. Since action potential of a single neuron is extremely weak only simultaneous action potentials of many neurons can be measured in EEG. Furthermore, the neuronal activity has to be synchronous, otherwise different phased activities cancel each other.

Magnetoencephalography (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 cortex. The next chapter covers MEG in more detail.

Another older measuring technique is the *one-cell recording*, where electric potential of a single cell is measured by an electrode in a head of a needle. Since insertion of the needle always kill some neurons, even thousands, this technique has been used much in animal studies, but not so much in human studies. Nowadays, even the animal studies are more uncommon, because of the increased concern of the animal rights.

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

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

Magnetic resonance imaging (MRI) and *functional magnetic resonance imaging* (fMRI) are brain imaging techniques that take advantage of nuclear magnetic resonance (NMR): 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 turn parallel 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, thus the blood flow in the brain can be monitored, giving information on the activity distribution in the brain.

The newest of the modern brain imaging techniques is the transcranial magnetic stimulation (TMS). In TMS, the cortex is stimulated by strong magnetic field pulses, which induce a post synaptical flow of current leading to the excitation of neurons.

6.2 Sensory systems

Perception or sensory systems[1] are the most studied subjects in the field of brain research; this is, certainly because of the importance of sensory information. Indeed, these are the *only* sources of information from the outer world into the brain. It would be hard to imagine an evolving brain without any outer input.

Another reason, why there are so many studies on perception, is perhaps the fact that it is reasonably easy to construct experiments with simple sensorial stimuli. Responses caused by such stimuli are called event related responses or evoked responses [63] and have been perhaps the most rewarding experiment paradigm in sensory research[2]. In evoked response studies, an event or stimulus, in a more psychological terminology, is shown to the subject, evoking a response in his or her brain. By repeating the stimulus (with some variation, perhaps), we can then extract the particular response to that stimulus from other ongoing brain activity.

In the next two sections, there is a rough introduction to two sensory systems used in the experiments reported in chapter 8 (for a more advanced introduction see e.g. [52]).

6.2.1 Hearing

Hearing needs perhaps the most preprocessing of all the senses: The path between the outer ear receiving changes in air pressure and the first neurons in auditory cortex is very complex and refined. It contains the tiniest bones in human body: *hammer, anvil and stirrup.* The structure amplifies the sound to be convenient for neuronal processing. In the cochlea, in the inner ear, there are low-frequency (up to 100 Hz) selective neurons, which fire in pace with the frequency of the incoming sound. Since the neurons can not fire quicker than 100 times per second, there has to be some other method for processing the higher frequency sounds. There are two competing theories for doing this: the *volley principle* [90] and the *maximum displacement principle* [107]. The former, which is possibly used within frequencies 60 to 5000 Hz, is based on many neurons firing in different phases, but not quickly enough, producing the same effect as one neuron firing in pace with the stimulus. The second theory comes from the fact that different frequencies produce different traveling waves inside the basilar membrane of the cochlea. That fact is probably used for all the frequencies over 60 Hz. This neuronal organisation causes the spatially adjacent neurons to be receptive to adjacent frequencies. This phenomenon is called *tonotopy.*

In addition to tonotopy or frequency maps as it could be called, other kinds of maps have been found, such as neurons receptive to the duration of a sound or the direction

[1] We concentrate on the latter, since in the former, we should provide some answer to the question about the relation of mental perception and neural basis. The answer to this question is irrelevant in the scope of this thesis, as its objective is to show that ICA is a valuable tool in MEG signal analysis.

[2] The experiment paradigm is called in EEG evoked potentials (EP) and in MEG evoked fields (EF), to discriminate from EEG.

Receptor	Location	Responds to
Free nerve ending	Around the base of the hairs and elsewhere in skin	Pain, warmth, cold
Hair-follicle receptors	Hair-covered skin	Movement of hairs
Meissner corpuscle	Hairless areas	Sudden displacement of skin; low-frequency vibration (flutter)
Pacinian corpuscle	Both hairy and hairless skin	Sudden displacement of skin; high-frequency vibration
Merkel's disks	Both hairy and hairless skin	Indentation of skin
Ruffini ending	Both hairy and hairless skin	Stretch of skin
Krause end bulbs	Hairless areas, perhaps including genitals; maybe some hairy areas	Uncertain

Table 6.1: *Different mechanoreceptors and their functions*

or the location of the sound.

6.2.2 Somesthetic sensing

There are seven types of cutaneous mechanoreceptors: Free nerve endings, Hair-follicle receptors, Meissner's corpuscles, Pacinian corpuscles (PC), Merkel disks, Ruffini endings, and Krause end bulbs. Each of these convey tactile information, having spesific receptive fields, contributing to the somesthesis in a unique way. The information is conveyed into the brain by afferent fibers, which can be over a meter long, making the mechanoreceptors the longest cells in the human body. The differences between the cells are collected into the table 6.1.

Tactile stimulus is usually applied to the hands of the subject, since they have a large smooth area with many mechanoreceptors. For example, the Pacinian corpuscles have highest density on the glabrous skin of the hand. High frequency vibrotactile stimuli elicit replicable somatosensory evoked fields (and EP's), the strongest response being between 45 and 60 ms. These responses originate in the primary somatosensory cortex called SI, contralateral to the stimuli as shown by MEG studies (e.g. [33]); approximately the same areas are activated by electrical stimulation.

Chapter 7

Magnetoencephalography and ICA

In this chapter MEG is discussed in more detail than the other brain researching techniques depicted in the chapter before. After that it is explained, how ICA can be used to analyse MEG data.

7.1 Basics of magnetoencephalography

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 or SPECT), nor exposes the studied subject to any radiation (like X-rays as CT does) or to strong magnetic fields (as in MRI and fMRI). It only *measures* magnetic flux through the scalp. That makes it very safe and useful in clinical medicine as well as in brain imaging research. The measured magnetic flux is mostly due to synchronous neuronal activity in the brain, mainly at the cortex level. The sensors are made of superconducting SQUID magnetometers [106]. Novel techniques have made it possible to combine several SQUIDs in a same measuring device, producing a helmet having even 306 sensors [95].

Figure 7.1 presents a simplified picture of one MEG measuring device. The measuring SQUID's are situated in a dewar, cooled by liquid helium. The upper right figure depicts the type of sensor used in the experiments of this thesis. It consists of two orthogonal sensors, sensitive to the components of the flux in the plane tangented to the head. The sensors are double-loop pick-up coils also referred to as gradiometers. They make the measurements to be "near-sighted" [38].

The temporal resolution of MEG is very high, close to 1 ms and the spatial resolution is quite good too, ca. 2-3mm in favorable circumstances [38]. The exceptional temporal accuracy makes it worthwhile to find sophisticated analysis tools for it, such as ICA.

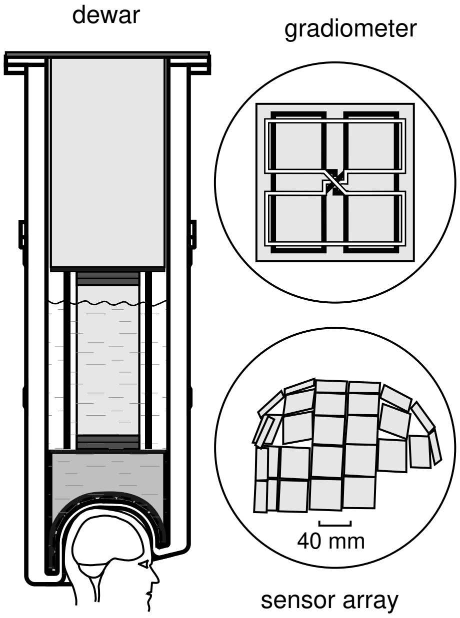


Figure 7.1: *A simplified picture of a MEG system using gradiometers adopted from [38].*

Another important reason to study MEG is the fact that it is completely non-invasive unlike all the other modern brain imaging techniques aside from EEG.

Electromagnetism

Electromagnetic fields can be fully described by the four Maxwell equations and the continuity equation:

$$\nabla\cdot\ \mathbf{E} = \rho/\epsilon_o, \quad (7.1)$$

$$\nabla\cdot\ \mathbf{B} = 0, \quad (7.2)$$

$$\nabla\times\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad (7.3)$$

$$\nabla\times\mathbf{B} = \mu_o(\mathbf{J} + \epsilon_o\frac{\partial \mathbf{E}}{\partial t}), \quad (7.4)$$

$$\nabla\times\ \mathbf{J} = -\frac{\partial \rho}{\partial t}, \quad (7.5)$$

where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields respectively; and ϵ_o and μ_o are the electric permittivity and magnetic permeability of the vacuum; ρ is the charge density, and $\mathbf{J}$ the current density; and finally $\nabla = [\frac{\partial}{\partial x}\frac{\partial}{\partial y}\frac{\partial}{\partial z}]$ is the partial differential operator called Nabla or Hamilton operator.

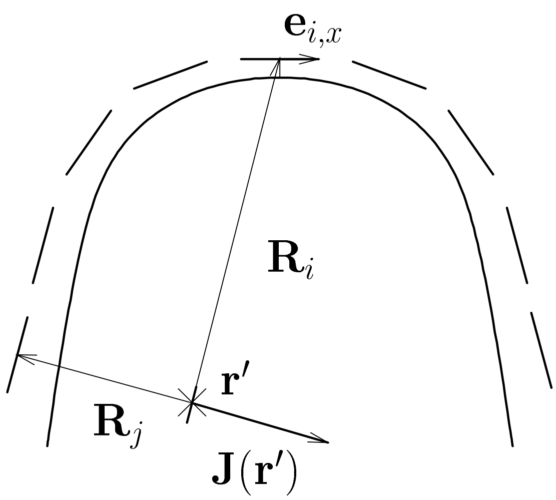


Figure 7.2: *The current density* $\mathbf{J}$*, located at* $\mathbf{r}'$*, produces a magnetic field in the sensor surface.* $\mathbf{e}_{i,x}$ *is the direction vector of one of the two sensors in the* i*th sensor couple.*

In a passive nonmagnetic medium, $\mathbf{J}$ can be calculated as the sum of the ohmic volume current and the polarisation current, i.e.

$$\mathbf{J} = \sigma\mathbf{E} + \frac{\partial \mathbf{P}}{\partial t}, \tag{7.6}$$

where σ is the conductivity of the material and $\mathbf{P} = (\epsilon - \epsilon_o)\mathbf{E}$ is the polarisation and ϵ is the electric permittivity of the material. Then (7.4) becomes

$$\nabla \times \mathbf{B} = \mu_o(\sigma\mathbf{E} + \epsilon\frac{\partial \mathbf{E}}{\partial t}). \tag{7.7}$$

As shown in [38], $|\epsilon\frac{\partial \mathbf{E}}{\partial t}| \ll |\sigma\mathbf{E}|$, if the $\mathbf{E}$ does not change rapidly. Hence the term $\frac{\partial \mathbf{E}}{\partial t}$ can be ignored as a source term. Furthermore it is shown in [38] that the $\frac{\partial \mathbf{B}}{\partial t}$ in (7.3) can be ignored as a source term too, giving the diameter of the head. These two approximations lead to the so called quasistatic approximation.

In the figure 7.2 the calculation of the magnetic field at locations $\mathbf{r}_i$ and $\mathbf{r}_j$ is depicted. The $\mathbf{R} = \mathbf{r} - \mathbf{r}'$ is the distance vector between the point in source region, which originates the magnetic field, and the location at which the magnetic field is calculated. The primed symbols refer to quantities in source region.

Under the quasistatic approximation the magnetic field can be calculated from the current density $\mathbf{J}$ of the source region using

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_o}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}') \times \mathbf{R}}{R^3} dv', \tag{7.8}$$

MEG measures only sources that lie tangentially to the sensors. This is easily seen from (7.8), since for sources radial to the sensors $\mathbf{J}(\mathbf{r}') \times \mathbf{R} = 0$. This is the case for the jth sensor couple in figure 7.2. At the location of the ith sensor couple, the distance vector $\mathbf{R}_i$ is perpendicular to the $\mathbf{J}(\mathbf{r}')$ and the cross product in (7.8) is maximum. However, the two sensors in $\mathbf{r}_i$ don't see the magnetic field from $\mathbf{r}'$ fully, since the sensors are not perpendicular to the $\mathbf{R}_i$. Instead the sensors measure $\mathbf{B}(\mathbf{r}_i) \cdot \mathbf{e}_{i,x}$ and $\mathbf{B}(\mathbf{r}_i) \cdot \mathbf{e}_{i,y}$, where $\mathbf{e}_{i,x}$ and $\mathbf{e}_{i,y}$ are the direction vectors of the sensors. If the head is assumed to be a spherically symmetric or a horizontally layered conductor, which is a rather realistic approximation, the radial sources produce no magnetic field outside the head whatsoever, no matter in what angle it is measured by the sensor array.

The so called *forward problem* of magnetoencephalography means the calculation of the magnetic field outside the head from a known magnetic field distribution inside the head originating it. In (7.8) the forward problem has been solved assuming that the *current distribution* inside the head is known. The *inverse problem*, i.e. the calculation of the magnetic field inside the brain from the magnetic flux outside the brain, is generally an ill posed problem i.e. for each magnetic flux distribution outside the brain there are an infinite number of magnetic field distributions that may originate it. The inverse problem can, however, be solved using some assumptions on the field distribution of the sources.

Equivalent current dipole model

One possible, and widely used, assumption is to model the sources with *(multiple) equivalent current dipoles* (ECD) [11] (for multiple dipoles see [35, 91, 68]). An electric current dipole is defined as a current density $\mathbf{Q}$ in one single location $\mathbf{r}_Q$, or using Dirac's delta function:

$$\mathbf{J}_p(\mathbf{r}) = Q\delta(\mathbf{r} - \mathbf{r}_Q). \tag{7.9}$$

Since only tangential components of the sources are measured by MEG sensors, the current dipole is fully defined by five parameters: the location is defined by the coordinate triplet (x, y, z) and the current density by two parameters, the dipole strength q and the orientation θ in the tangential plane.

Readers with more interest in the electromagnetism of MEG are advised to read [38] for the details of MEG and [3] for the physics of electromagnetism in general.

7.2 Applying ICA in MEG

The basic assumption, when applying ICA to MEG, is that inside the brain there are independent sources, which are mixed via a linear process and measured by SQUIDs on the sensor helmet, outside the scalp. The mixing, or calculation of the $\mathbf{x}(t)$ from the known $\mathbf{s}(t)$ via $\mathbf{A}$ in (2.3) is equivalent to the forward problem of MEG. Solving of the IC's from the equation (2.4) i.e. the calculation of the sources $\mathbf{s}(t)$ from the measurements, $\mathbf{x}(t)$ via the filtering matrix $\mathbf{W}$ is equivalent to the inverse problem, respectively.

Figure 7.3 presents the course of the overall algorithm: The magnetic field on the scalp is measured by the MEG device on the left, giving the measured $\mathbf{x}(t)$. This is then pre-whitened to decorrelate the data, to reduce noise and to avoid the artefactual bumps explained in section 5.1.1. This gives us the whitened data, $\mathbf{v}(t)$. The independent components are then solved using FastICA. Two of the independent components are shown in the bottom of the figure.

As a result of the ICA, both the mixing matrix $\mathbf{A}$ and the source signals $\mathbf{s}$ are useful: From $\mathbf{A}$ we can, using rather simple interpolation techniques, calculate the magnetic flux through the scalp, the column vector $\mathbf{a}_i$ corresponding to the field pattern of the ith source. In the figure, the magnetic flux on the scalp is depicted as a field pattern, where blue colour correspond to magnetic field coming out of the head and the red colour to the magnetic field going into the head. The intensity corresponds to the absolute value of the magnetic field. The field pattern can further be used to solve the inverse problem, under certain assumptions. The $\mathbf{s}(t)$, of course, gives the activation of the (independent) brain sources in time.

It is important to notice that ICA, as it is, does not make any assumptions on the shape of the source distribution: the sources can be either localised or distributed. In section 8.3 however, ECD is used as a source model enabling the localisation of the neural sources originating the independent components.

Another important thing one should remember about the ICA algorithm used is that it does not take into account any time correlation. In other words, the data could be in any order e.g. shuffled or sorted by some specific channel. This implies that dependent events with a time shift will appear independent. There are, of course, other ICA algorithms than FastICA (such as the one in [105], the idea was originally proposed by Tong in [94]) that would take these time-correlations into account, but as will be demonstrated in chapter 8, time insensitivity can be seen as an asset in some cases, not a drawback.

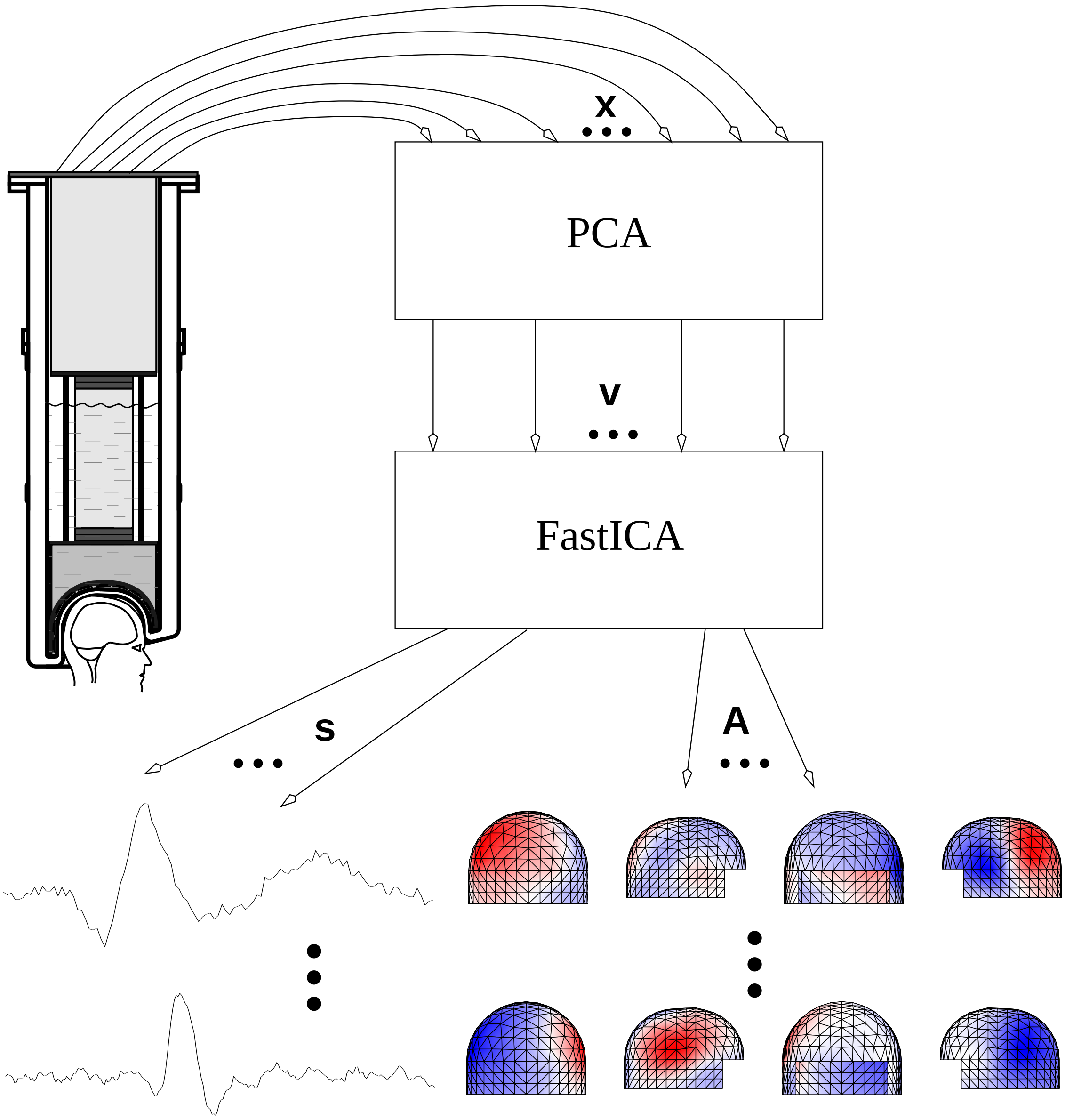


Figure 7.3: *A schematic flow of the independent component analysis of MEG.*

Chapter 8

Experiments

All MEG experiments have been recorded in a magnetically shielded room with a 122-channel whole-scalp Neuromag-122 neuromagnetometer. The device measures the magnetic flux at 61 locations, using two gradiometers, making the measurements "near-sighted" as stated in the previous chapter.

In the first experiment, FastICA is applied to extract artefacts, such as eye blinks, heart beating, etc., from MEG (and EEG) data. The second experiment is a non-averaged data analysis of auditory evoked fields (AEF) and the third decomposes the averaged AEF into components. The final experiment combines auditory and tactile stimuli.

The independent component analysis has been carried out using the matlab package for FastICA [28]. Some studies have been made with a previous, non-published version of the package.

8.1 Artefacts

In MEG (as well as in EEG) there are usually undesirable artefacts such as eye blinks, eye saccades, heart beating etc. Commonly, the parts with lots of artefactual activity are *rejected* altogether. This could, however lead to neglecting some important phenomena totally, should it normally produce some artefact. Indeed, one such case was noticed in the experiment of auditory evoked fields discussed in 8.2. Some other methods for dealing with artefacts are discussed in [99, 12, 33, 9].

Though not completely independent from brain activity, artefacts could be independent enough to allow the use of ICA-algorithms to extract them from the other data. ICA has been used for extraction of ocular artefacts in EEG [98]. The following results in MEG, were published in [99].

In the study, the subject was asked to blink and make horizontal eye movements. Fur-

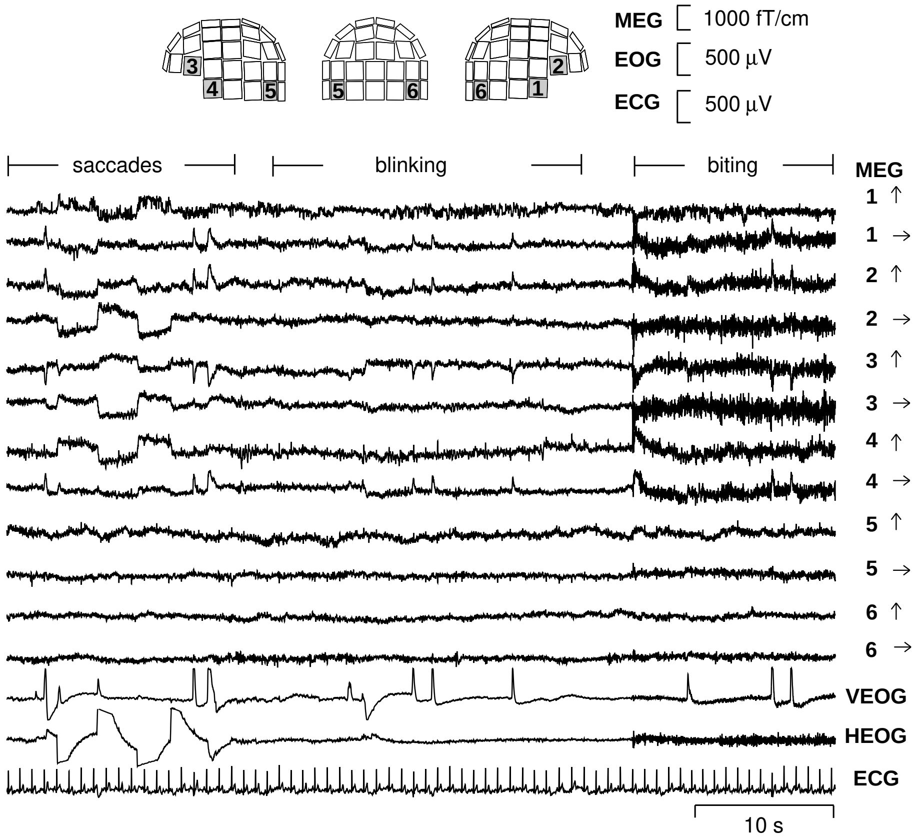


Figure 8.1: *Samples of MEG channels with artefacts including blinks, saccades, jaw biting, digital clock and a breathing artefact. For each one of the 6 positions, both orthogonal sensors (↑→) are plotted. The positions of the six sensors are shown in the three projection of the helmet on top. Adopted from [99].*

thermore, to produce myographic activity, the subject was to bite his teeth strongly for some time. Another type of artefact was a digital clock placed in the shielded room, a meter to the right of the helmet. One final artefact was produced by putting a metal piece on subject's navel to produce breathing artefact.

Fig. 8.1 shows some of the measured MEG channels as well as vertical and horisontal electro-oculograms (VEOG and HEOG) and an electrocardiogram (ECG), which were recorded to ease the identification of independent components. The blinks and the saccades as well as the biting-artefact are clearly visible already in the raw data. One can also see some marks of the heart beating and the digital clock.

Fig. 8.2 shows 9 of the independent components found. The three helmet pictures, above each of the signals, portray the field patterns i.e. the magnetic flux through the scalp, measured by the sensors, that would be caused by the corresponding IC. As explained already in section 7.2, the ith column of the mixing matrix $\mathbf{A}$ gives the value of the magnetic flux at the sensor locations, showed by dotted squares. The corresponding field pattern can be calculated from it by interpolating. The field patterns are depicted by equipotential curves in solid lines corresponding to flux going inwards and dashed lines corresponding to flux coming outwards. In the analysis, showed in the figure and taken straight from [99], the VEOG, HEOG and ECG were used as additional mixtures to ICA, but the present author reproduced all those results without them.

The first two IC's correspond clearly to the jaw biting and have practically no sign of the other artefacts. The two components seem to correspond, on the grounds of their field patterns, to different muscular groups (the first a lower and the second an upper one). The third and the fifth IC's correspond to the ocular artefacts, the third presenting the blinks and other vertical movements, and the fifth the saccades. IC4 clearly represents the cardiac artefact.

In the paper, the sixth and seventh IC's were suggested to be breathing artefacts. However, further analysis has shown them to be artefacts originating from the ill behaviour of the algorithm. This phenomenon, which leads to spiky or bumpy, hill-like signals has been discussed in section 5.1.1.

Because of the bumpy artefacts, it was not possible to extract the digital watch without preprocessing the data. In the paper it was noticed that high-pass filtering the data brought the watch artefact up. It is shown as the IC8. The ninth independent component corresponds to a broken sensor having much higher root mean squared noise than the others.

8.2 Auditory Evoked Fields

MEG has been used extensively to study the sensory systems. In many such experiments some simple stimulus is presented to the subjects repeatedly, perhaps varying some aspect of the stimulus. In case of auditory stimulus this could be e.g. frequency, duration or loudness. In some experiments, subjects are advised to perform some task depending of the stimulus characteristics. One could for example hear syllables, and be asked to press a button each time one hears syllable 'ta'. These kind of experiments are used to study the participation of different brain areas in particular tasks and to study the features of that participation, such as the latency between activation and the response as well as the duration of the response.

To study the use of FastICA in the analysis of sensory systems, a series of experiments using auditory stimulus was carried out. The setup details of the experiments are in table 8.1 and a draft portraying the experiment in Fig. 8.3. The tones applied to the subject's ears were of 1 kHz frequency and duration of 100 ms. They were used because they produce clear responses and cause negligible onsetting/offsetting -overlapping[1]. The detailed results of the analysis of the experiments #2 through #4 are published

[1]If the duration of the tone would be larger there would be an overlap between the preceding offset response and the succeeding onset response.

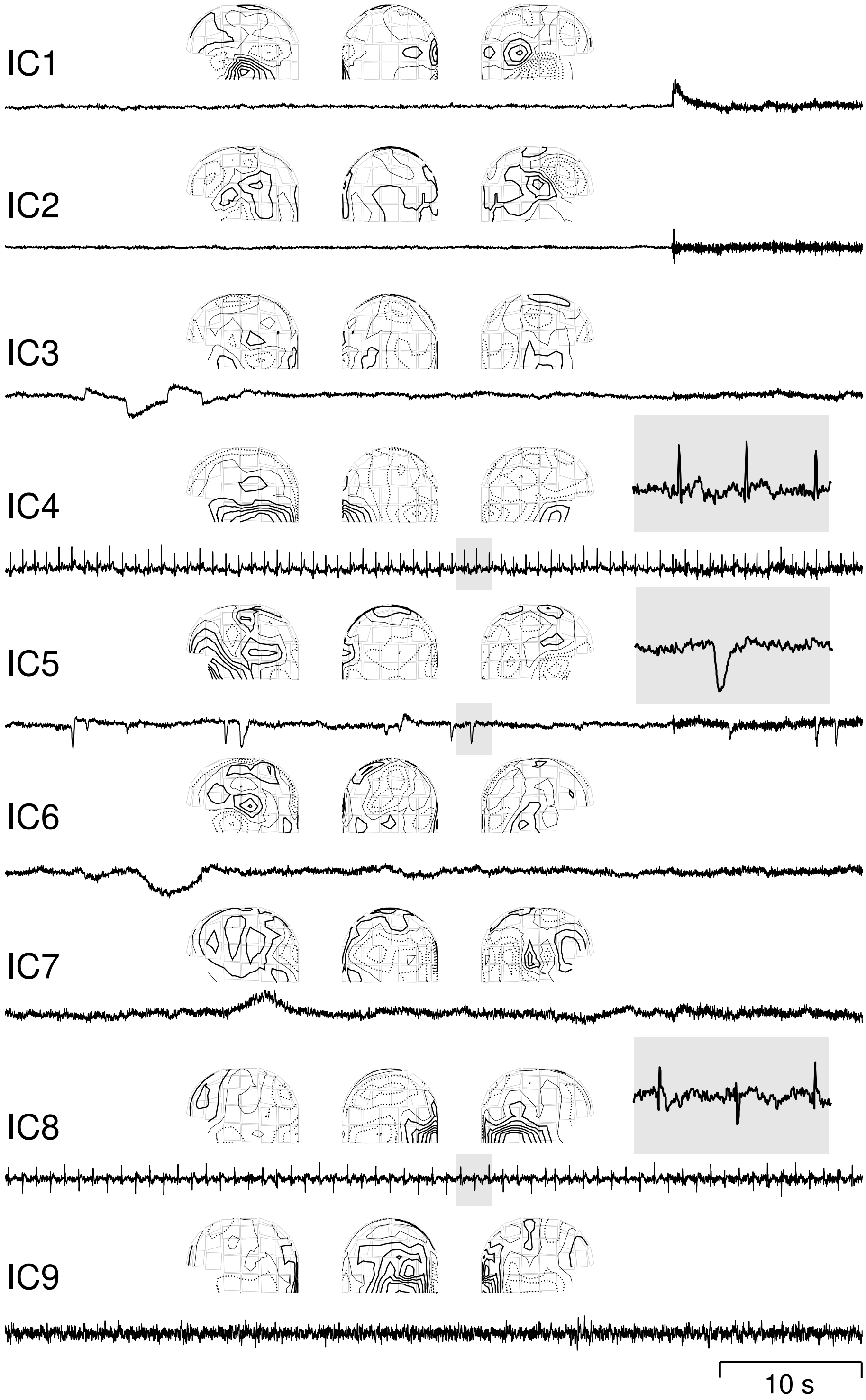


Figure 8.2: *Nine independent components found from the second minute of the data. In the field patterns the solid line corresponds to equipotential curves of the magnetic flux coming out of the head and the dotted line the flux inwards. Adopted from [99].*

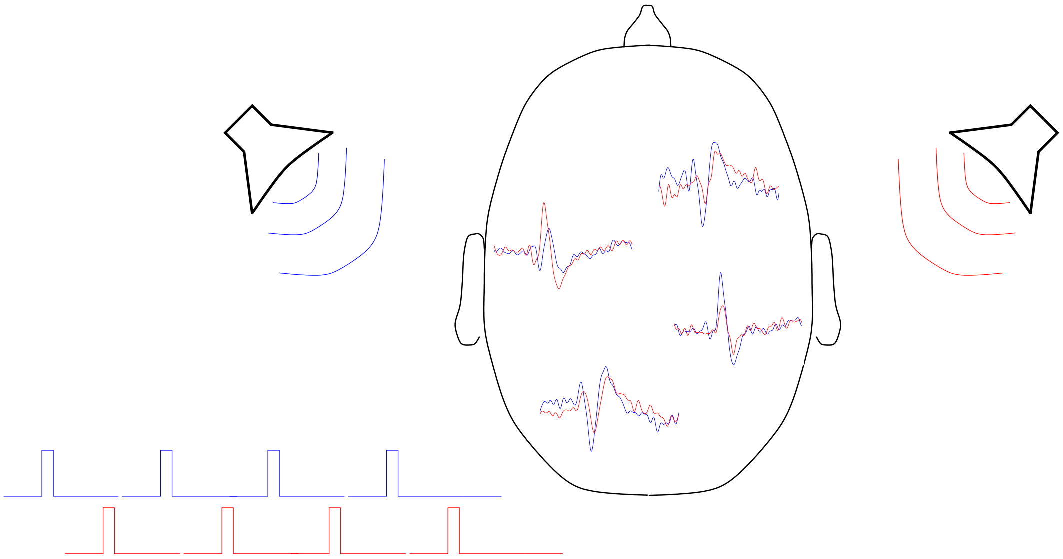


Figure 8.3: *Experiment draft of Auditory Evoked Fields. Stimuli are presented alternating to each ear. The signals inside the head represent the averaged responses measured on the scalp.*

in [92] and [97].

experiment	no. of stimuli l/r	duration ms	freq kHz	ISI ca. s
#1	18/18	100	1.0	5
#2	100/98	100	1.0	1
#3	200/200	100	1.0	1
#4	200/200	100	1.0	2

Table 8.1: *setup details of the AEF experiments. The second column is the number of stimuli presented to each ear separated by "/". The last column is the inter stimulus interval (ISI), which is the time between succeeding stimuli.*

Non-averaged case

The classical way of extracting the common response to repeated stimuli is to average the signal around the stimuli. This, of course, assumes the response to be stable. One case where the response evidently can not remain stable is learning. Since practically all experiments have some form of learning, should it be even as simple as habituation or excitation, this could be quite a problem. For more criticism of the averaging approach see [31]. For this reason the experiment #2 was used to test the power of FastICA on plain raw data without averaging. The following results are presented in [92].

The Figs. 8.4 and 8.5 present one IC found: The former shows the component averaged around the right ear stimuli after the ICA and the latter presents the field pattern corresponding to the IC. From the field pattern it can be seen that there is a strong dipolar structure on the left hemisphere (contralateral to the stimulation). However, there is also rather strong activation on the right hemisphere (ipsilateral).

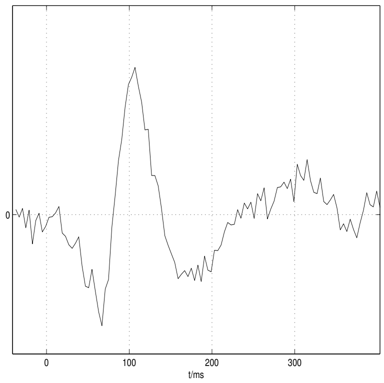


Figure 8.4: *Averaged independent component (times from -20ms to 400ms). Adapted from [92]*

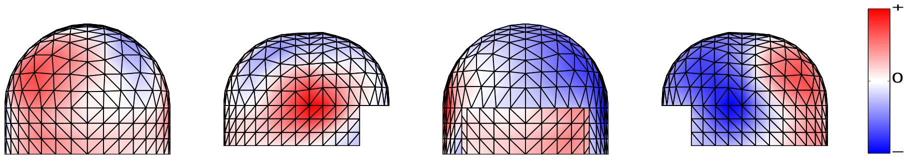


Figure 8.5: *The field pattern corresponding to the IC shown above. Adapted from [92]*

The signal to noise ratio (SNR) of the IC component is demonstrated in Figs. 8.6 and 8.7. The former presents the correlation between the total average and the averages of n first responses. Similarly such correlations are shown to all the MEG channels. It is clearly evident that the cumulative average of IC converges much quicker to its final value than the raw MEG channels do . Furthermore, the first row of the Fig. 8.7 shows that the final average of the IC has even lower SNR than the final averages of the MEG channels. The rest of the figure demonstrates the convergence speed of the IC.

The stationarity of the evoked response was also tested by comparing successive averages of n samples ($n \in [10, 20, 30]$ were tested), but no substantial differences were noticed. However, in the case experiment #1 such differences were noticeable. The main non-stationarity was the latency of the main peak around the expected 100 milliseconds (N100 or N1m). Sadly there were so few stimuli that no pattern could be found.

Although the Figs. 8.4 through 8.7 clearly demonstrate the power of FastICA in the case of non-averaged AEF-data, principal component analysis provided components with comparable physiological significance as well as SNR. The PCA even discriminated between the contra- and ipsilateral responses, which ICA did not do. So, *as such*, ICA does not bring anything more to non-averaged event related studies. We, however, emphasise the *as such* part, since this analysis was quite preliminary.

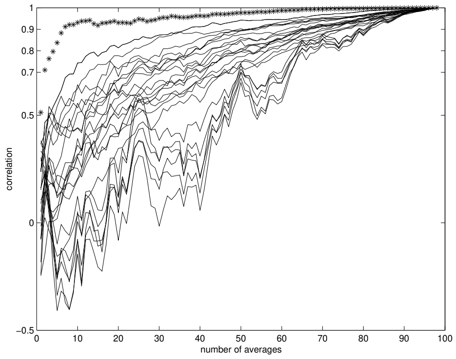


Figure 8.6: *Speed of convergence of the cumulative average of the IC (•) and of some MEG signals (-). Adapted from [92]*

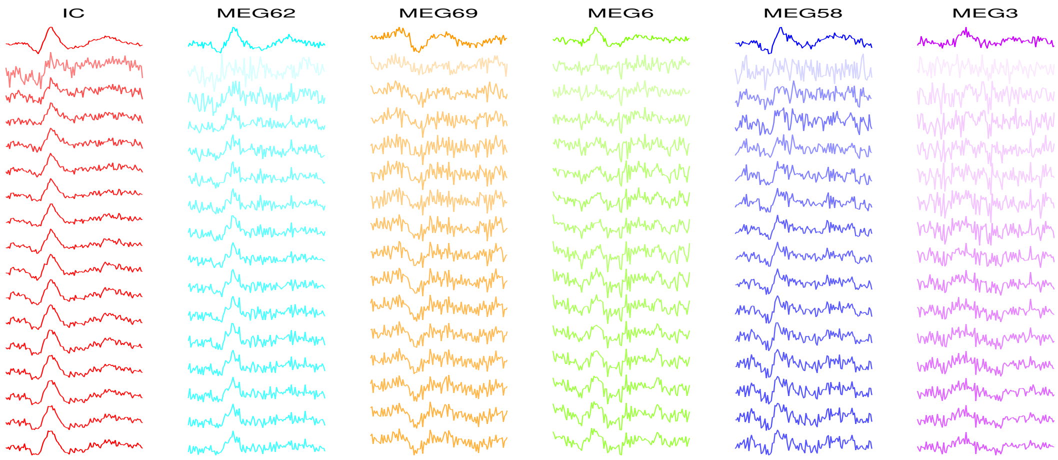


Figure 8.7: *Global (100 responses) and cumulative (1–15) averages of the IC and five of the most representative MEG channels. The intensity (as in HSI colour model) of each curve is the correlation between it and the total average. The time scale, as in Fig. 8.4, range from -20ms to 400ms. Adapted from [92]*

Averaged case

The response to an auditory stimulus has many aspects, consider for example recognition of speech. Even quite simple (pure, sinusoidal) tones evoke quite complex responses including, besides the recognition of the frequency, duration and loudness, cognitive tasks such as deciding whether the sound needs attention or even some intermodal action, redirection of visual attention or movement of body etc. For this reason it is desirable to be able to decompose the response into parts, which may have physiological relevance.

One way to achieve this is to construct a model for each of the components, i.e. making some a priori hypothesis. Better way, in our understanding, would be to have some

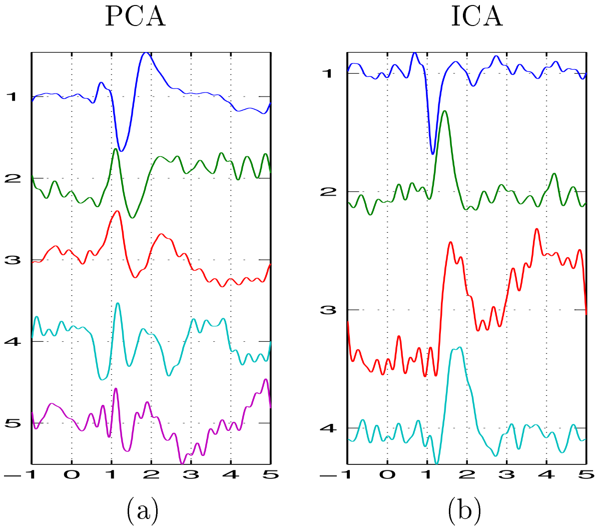


Figure 8.8: *Principal (a) and independent (b) components for the 1s ISI.Each tick in (a) and (b) corresponds to 100ms, going from 100ms before stimulation onset to 500ms after. Adapted from [97]*

general, flexible, statistical model that would lead to physiological components. For this PCA has been used [1], but it has been criticised for its unphysiological results [104]. This matter is still controversial (read [66, 2] for the ongoing debate). Anyhow, since ICA can be seen as an extension of PCA, it is worthwhile to study the use of it for the decomposition of evoked responses[2].

To do this the data of experiment #3 was averaged around the right ear stimuli[3] giving 594 mixture samples. The data was also low-pass filtered, with a cutoff frequency of 45 Hz, to eliminate the 50 Hz noise due to external power source.

Fig. 8.8 shows the independent component decomposition (b) and Fig. 8.9 their field patterns. Evidently the first and the third independent components (IC1 and IC3) correspond to the main response peak around 100ms. From the field patterns 1 and 3, we can see that one represents dominance (even a clear dipole) on the contralateral side, and the other on the ipsilateral side. Indeed, as expected [34], IC3, corresponding to the response ipsilateral to the stimulus has longer latency than IC1, which corresponds to the contralateral response.

Yet another demonstration of the capability of the FastICA to discriminate between contra- and ipsilateral responses is shown in Fig. 8.10. In the figure, the four independent components are regressed on two MEG channels (60 and 10) lying near the ears. From the figure it becomes evident that IC1 (blue line) explains almost totally the N1m

[2]Note that if we would use an algorithm that takes the time dependencies into account, we would not be able to do the decomposition, since certainly the components will be time dependent.

[3]Left ear stimulation was also analysed, giving comparable results. The experiment #4, with 2s ISI also gave comparable results. Neither of the results are presented here.

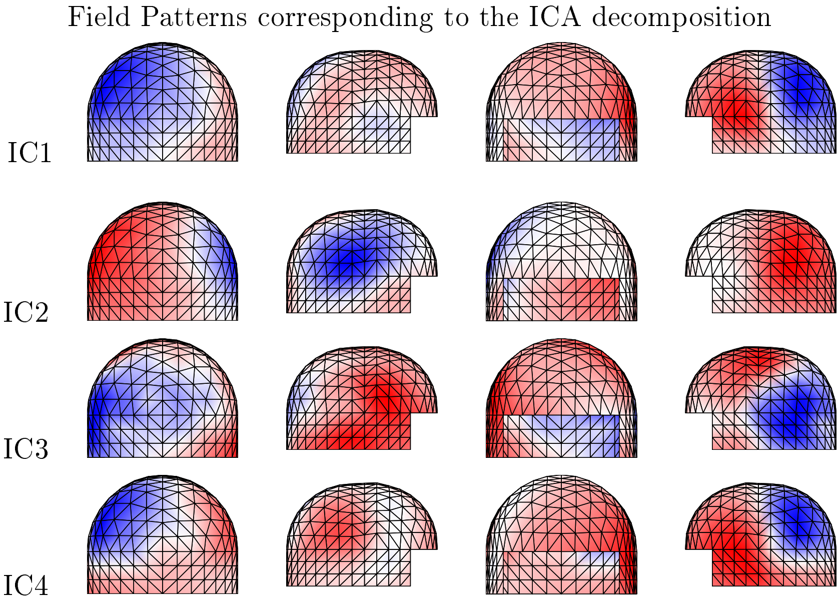


Figure 8.9: *Field patterns associated to the ICA decomposition in previous figure. Adapted from [97]*

of MEG60, but has very little to do with the counterpart in MEG10; A similar consideration can be made for IC3, and its approximation of MEG10. Finally, IC4 explains the longer response around 180ms on the contralateral side. No component of that kind was found on the ipsilateral side, though. During the experiments one more component, though not presented here, was found. It presents a clearly dipolar structure in frontal areas, like the eye saccades in artefact experiments. One possible reason for this is that,every now and then, the beep caused an eye movement.

In the case of non-averaged evoked field study it was noted that PCA, already, gave comparable results to ICA. Is this also true for averaged decomposition? Clearly not, as can be seen from Fig. 8.8a: the principal components are much more complex

To summarise the results in this section we note that the decomposition of auditory evoked fields give field patterns similar to patterns arousing from electrical dipoles. This supports the hypothesis of equivalent current dipoles as a source model. It is once more stressed that when performing ICA, we did not make *any* assumption on the distribution of the sources, other than independency[4].

In the next section yet another experiment on auditory evoked responses was performed using ECD model for source localisation. This was done to further compare the results from FastICA to classical results. The experiment also included somatosensory stimulation.

[4]Statistically this is, of course, a rather strong constraint, but the knowledge used is free from physiological models as it would not be e.g. in the case of applying ECD straight to averaged data.

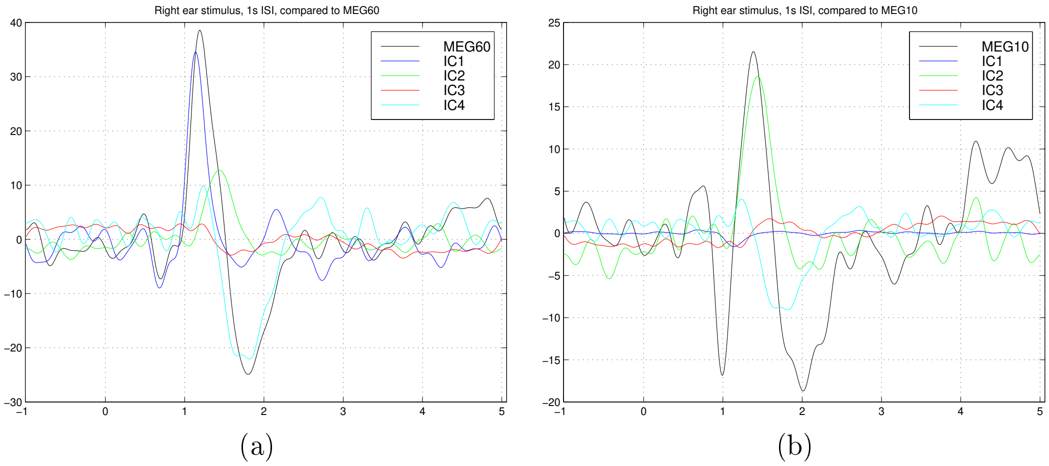


Figure 8.10: *Example of 2 MEG signals, MEG60 recorded on the left temporal area (a), and MEG10 on the right (b). The independent components found are plotted after scaling to the respective MEG recording. Adapted from [97]*

8.3 Rudolph Evoked Fields

The results of this experiment, including both auditory and somatosensory stimulation on 6 healthy subjects, are reported in [96] and in [49] (ICA and conventional analysis respectively). Vibrotactile stimuli, generated with a bass-reflex loudspeaker and a tube delivering the tone into the magnetically shielded room, were coupled to a balloon[5] which was lightly touched by the subject's hands. Both auditory and somatosensory stimulation was elicited in the same experiment, because of the high sound pressure (ca. 60 dB SPL) of the stimulator. The averaged responses and magnification of the strongest responses of the subject S4 are shown in Fig. 8.11. The subject S4 was also chosen as the case subject for ICA.

Fig. 8.12 presents the result of the ICA study and Fig. 8.13 the corresponding field patterns. The former shows the principal (a) and independent (b) decomposition of the averaged response. The latter figure shows the field patterns using equipotential flux curves as in Fig. 8.2. Since the previous study with AEF supported dipolar patterns for decomposition of auditory responses, we use ECD to localise the sources in IC1 and IC2, which present the clearest dipolar structure.

We localised the sources corresponding to the first two independent components, finding the primary auditory responses on the supratemporal area, corresponding to IC1, and the primary somatosensory responses, corresponding to IC2. The equivalent current dipoles are superimposed in Fig. 8.13. The localisations can be compared to conventional source localisations from [49] by comparing respectively the Figs. 8.14 and 8.15, where the dipoles have been superimposed on MRI-slides. The conventional and the

[5]The name Rudolph comes after Santa Clause's red-nosed reindeer, because that is what the balloon looks like — a red nose.

ICA-based procedures agree in localisations, given the spatial resolution of MEG.

The goodness-of-fit [57] of the model based on the first 2 independent components was of the same order of magnitude as the one in [49] (exceeding 90% during the peak activations of SI and auditory cortices), slightly surpassing it when other independent components were included in the estimations.

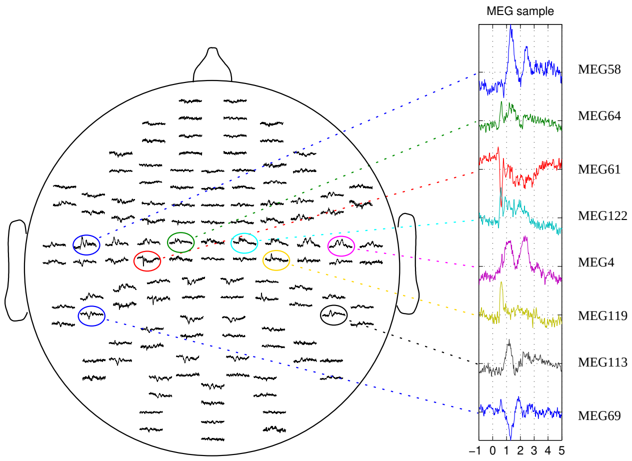


Figure 8.11: *Recorded MEG, in a case of simultaneous somatosensory and auditory event related study. A sample of the 122 averages is enlarged on the right side. Each tick represents 100 ms. Adapted from [96]*

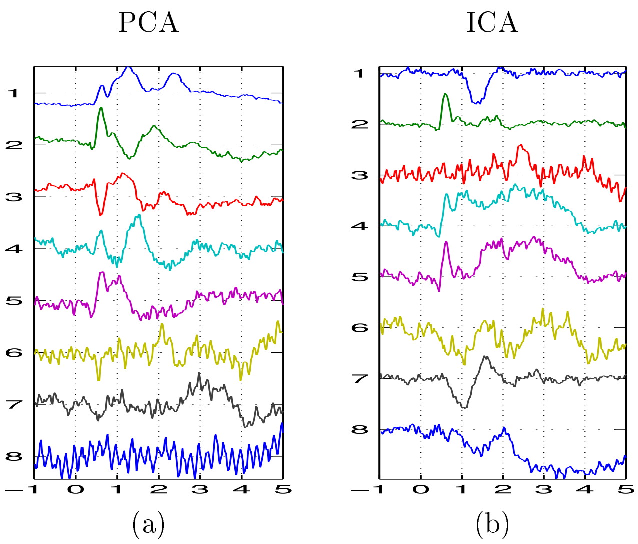


Figure 8.12: *Principal (a) and independent (b) component analysis for the combined SEF and AEF study. Adapted from [96]*

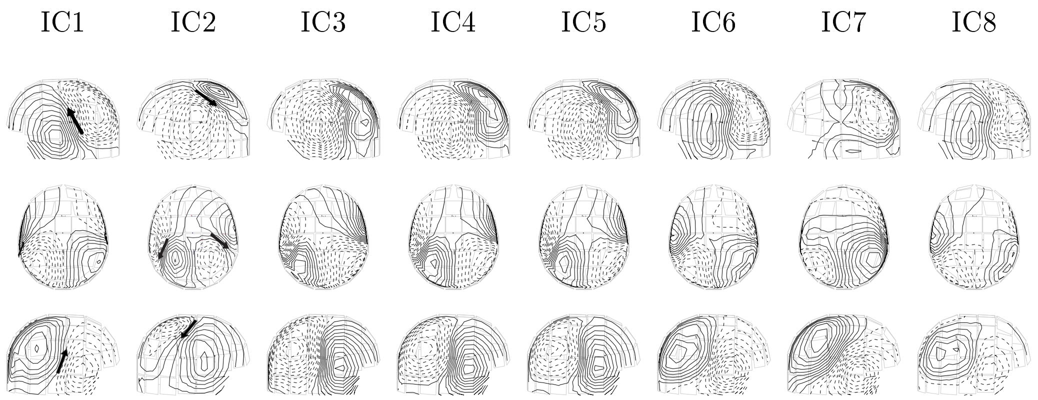
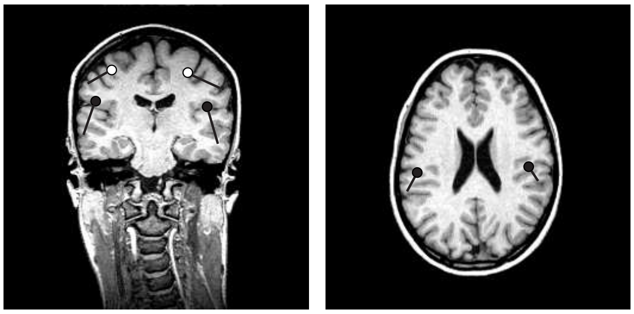


Figure 8.13: *Field patterns corresponding to each independent component. The arrows in IC1 and IC2 correspond to the equivalent current dipole generating the pattern. Note that only the orientation is relevant, as the IC's are defined up to a permutation and a scaling constant (that may be negative). Adapted from [96]*



(a) (b)

Figure 8.14: *Localization of IC1 (black circles, corresponding to the auditory cortex activation) and IC2 (white circles, corresponding to the SI cortex activation). The bars illustrate the orientation of the source net current. Adapted from [96]*

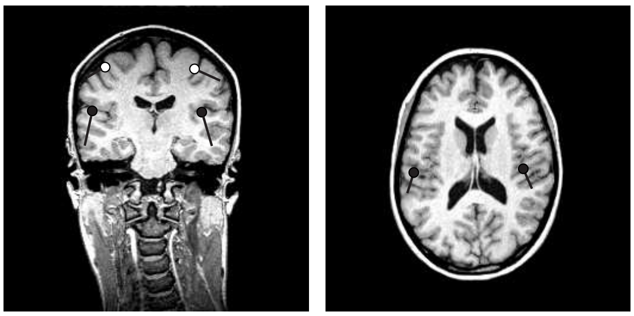


Figure 8.15: *Conventional source localization from the same study as above (adapted from [49]).*

Chapter 9

Summary and conclusion

In this thesis we applied *independent component analysis* on *magnetoencephalographic* signals. MEG, which is a sophisticated, completely non-invasive technique for measuring brain activity, provides accurate information of the rapid changes in magnetic flux through the scalp due to the synchronous neuronal activity. ICA is a novel signal analysis method for blind source separation and projection pursuit. In MEG-analysis both the goals are desirable, which makes ICA a natural tool for MEG-analysis.

It was shown how ICA was applied successfully in *artefact removal* from MEG data (section 8.1). The first study where the author of this thesis contributed is the experiment studying the *auditory responses* to simple sinusoidal stimuli (in section 8.2). The ICA seemed promising for detecting the non-stationarities in successive responses, when applied to non-averaged data (section 8.2), but a wider series of experiments would be needed with many subjects to verify the fact. In decomposition of *evoked fields* ICA showed its strength (section 8.2): the response was decomposed into dipolar sources, distinguishing e.g. between the ipsi- and contralateral main responses. The strength was further demonstrated in *Rudolph evoked field* -study (section 8.3), where actual *equivalent current dipole* -model (ECD) was applied to the independent components found. The localisations agreed with conventional ones.

In the light of the experimental results, this thesis has shown that ICA is very useful in the analysis of MEG signals, but there is still much work to be done in the area. For example the facts already known about the structure of the brain could be utilised far more better, which means that our approach has been blinder than needed. On the other hand, a blind approach can strengthen the belief in conventional results: It was shown that ICA leads to dipolar patterns, thus promoting the ECD model. The next step could be to incorporate the ECD model into the ICA. Additionally, the ECD model inside ICA could solve the problem of the artefactual bumps presented in section 5.1.1 and in full detail in [47], since ECD model dramatically reduces the set of possible mixing matrices. This model could also be used to evolve further the non-averaged data analysis. Also some other possibilities should be considered, keeping single-trial responses as the ultimate goal.

The use of ICA should also be considered in other brain imaging techniques, such as functional magnetic resonance imaging (in this something has already been done [65]), positron emission tomography, computer axial tomography as well as to continue the already started exploration of ICA in electroencephalography. Finally, the use of ICA in clinical applications should be explored to establish it as one of the standard tools for medical signal analysis.

Appendix A

Review of some statistical concepts

A.1 Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are called *independent* if and only if

$$P(\mathcal{A}\mathcal{B}) = \mathcal{P}(\mathcal{A})\mathcal{P}(\mathcal{B}). \tag{A.1}$$

From the conditional probability $P(\mathcal{B}|\mathcal{A})$ given by

$$P(\mathcal{B}|\mathcal{A}) = \frac{\mathcal{P}(\mathcal{A}\mathcal{B})}{\mathcal{P}(\mathcal{A})}, \tag{A.2}$$

provided that $P(\mathcal{A}) \neq 0$, we see that if $\mathcal{A}$ and $\mathcal{B}$ are independent then $P(\mathcal{B} \mid \mathcal{A}) = \mathcal{P}(\mathcal{B})$. By analogy $P(\mathcal{B}|\mathcal{A}) = \mathcal{P}(\mathcal{B})$. These definitions can be straightforwardly applied to *random variables* instead of events. Let x and y be independent random variables, then it holds

$$P(x \in A, y \in B) = P(x \in A)P(y \in B), \tag{A.3}$$

where $A, B \subset \mathcal{R}$. This relation can be formulated for the densities of x and y respectively:

$$f_{xy}(x, y) = f_x(x)f_y(y). \tag{A.4}$$

The conditional density

$$f_y(y|x) = \frac{f_{xy}(x,y)}{f_x(x)} \tag{A.5}$$

becomes $f_y(y|x) = f_y(y)$ if x and y are independent, $f_x(x|y) = f_x(x)$ respectively. This means that no knowledge of any form about the value of x (or y) gives us information about the value of y (or x).

A.2 Information theory

In this section we derive the basic information theoretic concepts for continuous variables. These results are widely used in chapter 4.

Perhaps the most important information theoretic concept is the *differential entropy*. It is defined, for a continuous variable x, as

$$H(x) = -\int_{-\infty}^{\infty} f_x(x) \log f_x(x)\, dx. \tag{A.6}$$

When the *entropy* of a discrete random variable measures the uncertainty of the outcome, the differential entropy measures the uncertainty of x.

It is very natural to ask: how does the uncertainty of x change, when we observe something about another random variable y. The answer to this question is the *conditional differential entropy*, defined as

$$H(x|y) = -\int_{-\infty}^{\infty} f_y(y) \int_{-\infty}^{\infty} f_x(x|y) \log f_x(x|y)\, dx\, dy. \tag{A.7}$$

If x is *independent* from y the conditional differential entropy is equal to the differential entropy of x, otherwise it is smaller (for proof of this, see [85]).

With the conditional differential entropy, we can relate the information between two variables into each other. The information in common or *the mutual information* can be calculated as

$$\begin{aligned} I(x,y) &= H(x) + H(y) - H(x,y) \\ &= H(x) - H(x|y) = H(y) - H(y|x). \end{aligned} \tag{A.8}$$

Another measure for distance between random variables is given by the *Kullback-Leibler divergence*[1]. It is defined as

$$I_{KL}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y) \log \frac{f_{xy}(x, y)}{f_x(x) f_y(y)} \, dx \, dy, \tag{A.9}$$

The previous concepts can also be defined for random vectors $\mathbf{x} = [x_1 \ \cdots \ x_n]^T$. The differential entropy of $\mathbf{x}$ is

$$H(\mathbf{x}) = -\int_{-\infty}^{\infty} f_{x_1 \cdots x_n}(x_1, \ldots, x_n) \log f_{x_1 \cdots x_n} d\mathbf{x}, \tag{A.10}$$

where $\int_{-\infty}^{\infty} d\mathbf{x}$ means integration with respect to all components. The mutual information becomes

$$\begin{aligned} I(\mathbf{x}) &= \sum_{i=1}^{n} H(x_i) - H(x_1, \ldots, x_n) \\ &= \int_{-\infty}^{\infty} f_{\mathbf{x}}(\mathbf{x}) \log \frac{f_{\mathbf{x}}(\mathbf{x})}{\prod_{i=1}^{n} f_{x_i}(x_i)} \, d\mathbf{x}. \end{aligned} \tag{A.11}$$

One more useful concept can be constructed from the definition of differential entropy, noting that for distributions with fixed mean and variance, a Gaussian random variable has the largest differential entropy [36]. Then we get a measure of non-Gaussianity by defining

$$J(\mathbf{x}) = H(\mathbf{x}_G) - H(\mathbf{x}). \tag{A.12}$$

This measure is called *negentropy*. Note that negentropy is actually part of the generalised contrast function class defined in section 3.9.

The mutual information for a random vector $\mathbf{x}$ may be also calculated from the negentropy (see [21] for proof) giving

$$I(\mathbf{x}) = J(\mathbf{x}) - \sum_{i=1}^{n} J(x_i). \tag{A.13}$$

[1] It is sometimes also referred to as Kullback-Leibler *distance*, although it is not a metric, since it is not symmetric.

This is more useful for ICA than the differential entropy, since it is invariant under linear invertible transformations.

Appendix B

Glossary of medical terms

afferent axon an axon that brings information into a neuron
anvil one of the small bones of the middle ear; also known as the incus
aphasia lack of some form of language abilities
axon a single fiber that extends from a neuron
cochlea structure in the inner ear, containing auditory receptors
contralateral on the other side of the body (can be either right or left)
cortex, cerebral cortex layer of cells on the outer surface of the forebrain
cutaneous skin
basilar membrane floor of the middler chamber in cochlea
efferent axon an axon that carries information away from a neuron
glabrous naked, without hair
hammer one of the small bones of the middle ear; also known as the malleus
hemisphere either the left or the right half of the brain
ipsilateral on the same side of the body (can be either right or left)
SI primary somatosensorial cortex, part of the brain where the primary information processing of somatosensorial input takes place
SII secondary somatosensorial cortex
somesthetic somatosensory
somatosensory cortex the portion of the cerebral cortex that responds to touch and body information
stirrup one of the small bones of the middle ear; also known as the stapes
tactile touch
tonotopy arrangement in such a way that each anatomical subdivision of a brain area corresponds to a different auditory tone, adjacent areas corresponding to adjacent tones.
vibrotactile vibrating touch

Bibliography

[1] A. Achim and S. Bouchard. Towards a dynamic topographic components model. *Electroenceph. clin. Neurophysiol.*, 103:381–385, 1997.

[2] A. Achim and W. Marcantoni. Principal component analysis of event-related potentials: misallocation of variance revisited. *Electroenceph Clin Neurophysiol*, 34(5):597–606, 1997.

[3] M. Alonso and E.J. Finn. *Fundamental University Physics, Volume II: Fields and Waves*. Addison-Wesley, 7th edition, 1975.

[4] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing 8 (Proc. NIPS'95)*, pages 757–763. MIT Press, Cambridge, MA, 1996.

[5] S.-I. Amari. Neural learning in structured parameter spaces — natural riemannian gradient. In *Advances in Neural Information Processing 9 (Proc. NIPS'96)*, pages 127–133. MIT Press, Cambridge, MA, 1997.

[6] M. S. Bartlett. A note on the multiplying factors for various chi-square approximations. *J. Roy. Stat. Soc.*, 16 ser B:296–298, 1989.

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

[8] A. Belouchrani and J.-F. Cardoso. Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation. In *Proc. NOLTA*, pages 49–53, 1995.

[9] P. Berg and M. Scherg. A multiple source approach to the correction of eye artifacts. *Electroenceph. clin. Neurophysiol.*, 90:229–241, 1994.

[10] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[11] D. Brenner, J. Lipton, L. Kaufman, and S.J. Williamson. Somatically evoked magnetic fields of the human brain. *Science*, 199:81–83, 1978.

[12] C. H. M. Brunia, J. Möcks, and M. Van den Berg-Lennsen. Correcting ocular artifacts — a comparison of several methods. *J. Psychophysiol.*, 3:1–50, 1989.

[13] J.-F. Cardoso. Source separation using higher order moments. In *Proc. ICASSP'89*, pages 2109–2112, 1989.

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

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

[16] A. Cichocki, S. I. Amari, and R. Thawonmas. Blind signal extraction using self-adaptive non-linear hebbian learning rule. In *Proc. NOLTA'96*, pages 377–380, 1996.

[17] 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.

[18] A. Cichocki, W. Kasprzak, and S. I. Amari. Multi-layer neural networks with alocal adaptive learning rule for blind separation of sources. In *Proceedings of the Research Society of Nonlinear Theory and Applications (NOLTA) '95*, pages 61–65, Las Vegas, USA, 1995.

[19] 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.

[20] 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.

[21] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[22] D. Cook, A. Buja, and J. Cabrera. Projection pursuit indexes based on orthonormal function expansions. *J. of Computational and Graphical Statistics*, 2(3):225–250, 1993.

[23] G. Deco and W. Brauer. Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures. *Neural Networks*, 8(4):525–535, 1995.

[24] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[25] A. P. Dempster, N. Laird, and D. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *J. of the Royal Statistical Society, ser. B*, 39:1–38, 1977.

[26] S.C. Douglas, A. Cichocki, , and S. Amari. A bias removal technique for blind source separation with noisy measurements. *Electronics Letters*, 1998. To appear.

[27] M.S. Gazzaniga ed. *The Cognitive Neurosciences.* A Bradford book/MIT Press, 4th edition, 1997.

[28] The FastICA MATLAB package. Available at `http://www.cis.hut.fi/projects/ica/fastica/`, 1998.

[29] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. of Computers*, c-23(9):881–890, 1974.

[30] J.H. Friedman. Exploratory projection pursuit. *J. of the American Statistical Association*, 82(397):249–266, 1987.

[31] A. S. Gevins and N. H. Morgan. Classifier-directed signal processing in brain research. *IEEE Transactions on Biomedical Engineering*, BME-33(12):1054–1068, 1986.

[32] M. Girolami and C. Fyfe. Kurtosis extrema and identification of independent components : A neural network approach. In *Proc. ICASSP'97*, pages 3329–3332, Munich, Germany, 1997.

[33] R. Hari. Magnetoencephalography as a tool of clinical neurophysiology. In Ernst Niedermeyer and Fernando Lopes da Silva, editors, *Electroencephalography. Basic principles, clinical applications, and related fields*, pages 1035–1061. Baltimore: Williams & Wilkins, 1993.

[34] R. Hari, M. Pelizzone, J.P. Mäkelä, J. Hällström, L. Leinonen, and O.V. Lounasmaa. Neuromagnetic responses of the human auditory cortex to on- and offsets of noise bursts. *Audiology*, 26:31–43, 1987.

[35] R. Hari, K. Reinikainen, E. Kaukoranta, M. Hämäläinen, R.J. Ilmoniemi, A. Penttinen, J. Salminen, and D. Teszner. Somatosensory evoked cerebral magnetic fields in si and sii in man. *Electroencephalogr. Clin. Neurophysiol.*, 57(?):254–263, 1984.

[36] S. Haykin. *Neural Networks. A Comprehensive Foundation.* Macmillan College Publishing, New York, 1994.

[37] D. O. Hebb. *The Organization of Behavior.* Wiley, New York, 1949.

[38] M. Hämäläinen, R. Hari, R.J. 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.

[39] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[40] A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, Munich, Germany, 1997.

[41] A. Hyvärinen. *Independent Component Analysis: A Neural Network Approach.* PhD thesis, Helsinki University of Technology, Department of Computer Science and Engineering, October 1997.

[42] A. Hyvärinen. Independent component analysis by minimization of mutual information. Technical Report A46, Helsinki University of Technology, Laboratory of Computer and Information Science, 1997.

[43] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Neural Networks for Signal Processing VII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pages 388–397, Amelia Island, Florida, 1997.

[44] A. Hyvärinen. Independent component analysis in the presence of noise: A maximum likelihood approach. In *Proc. I&ANN Workshop*, pages 32–38, Tenerife, Spain, 1998.

[45] A. Hyvärinen. Noisy independent component analysis by gaussian moments. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, Aussois, France, 1999. To appear.

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

[47] A. Hyvärinen, J. Särelä, and R. Vigário. Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, Aussois, France, 1999. To appear.

[48] M.C. Jones and R. Sibson. What is projection pursuit ? *J. of the Royal Statistical Society, ser. A*, 150:1–36, 1987.

[49] V. Jousmäki and R. Hari. Somatosensory evoked fields to large-area vibrotactile stimuli. *Electroenceph. clin. Neurophysiol.*, 1998. Submitted.

[50] C. Jutten. *Calcul neuromimétique et traitement du signal, analyse en composantes indépendentes*. PhD thesis, INPG, Univ. Grenoble, 1987. (in French).

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

[52] J.W Kalat. *Biological Psychology*. Brooks/Cole publishing company, 5th edition, 1995.

[53] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

[54] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8(3):486–504, 1997.

[55] J. Karhunen and P. Pajunen. Blind source separation using least-squares type adaptive algorithms. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3048–3051, Munich, Germany, 1997.

[56] J. Karhunen, P. Pajunen, and E. Oja. New results on nonlinear PCA criterion in blind source separation. *Neurocomputing*, 1998. To appear.

[57] E. Kaukoranta, M. Hämäläinen, J. Sarvas, and R. Hari. Mixed and sensory nerve stimulations activate different cytoarchitectonic areas in the human primary somatosensory cortex si. *Exp. Brain Res.*, 63(1):60–66, 1986.

[58] M. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin & Company, 1958.

[59] Beate Laheld and Jean-François Cardoso. Adaptive source separation with uniform performance. In *Proc. EUSIPCO*, pages 183–186, Edinburgh, 1994.

[60] H. Lappalainen. Ensemble learning for independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, Aussois, France, 1999. To appear.

[61] D. N. Lawley. Test of significance of the latent roots of the covariance and correlation matrices. *Biometrica*, 43:128–136, 1956.

[62] M. Lewicki and T. J. Sejnowski. Learing overcomplete representations. In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 556–562. MIT Press, 1998.

[63] F. Lopes da Silva. Event-related potentials: Methodology and quantification. In Ernst Niedermeyer and Fernando Lopes da Silva, editors, *Electroencephalography. Basic principles, clinical applications, and related fields*, pages 877–886. Baltimore: Williams & Wilkins, 1993.

[64] D. G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, 1969.

[65] Martin McKeown, Scott Makeig, S. Brown, Tzyy-Ping Jung, S. Kindermann, Anthony Bell, V. Iragui, and Terrence Sejnowski. Blind separation of functional magnetic resonance imaging (fmri) data. *Human Brain Mapping*, 1998.

[66] J. Mocks and R. Verleger. Principal component analysis of event-related potentials: a note on misallocation of variance. *Electroencephalogr Clin Neurophysiol*, 65(5):393–398, 1986.

[67] E. Moreau and O. Macchi. High-order contrasts for self-adaptive source separation. *International Journal of Adaptive Control and Signal Processing*, pages 19–46, 1996.

[68] J. Mosher, P Lewis, and R Leahy. Multidipole modeling and localization from spatio-temporal meg data. *IEEE Trans Biomed Eng*, 39:541–557, 1992.

[69] E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3617–3620, Munich, Germany, 1997.

[70] C. Nikias and J. Mendel. Signal processing with higher-order spectra. *IEEE Signal Processing Magazine*, pages 10–37, July 1993.

[71] E. Oja. *Subspace Methods of Pattern Recognition.* Research Studies, Press, Letchworth, England, 1983.

[72] E. Oja. Neural networks, principal components, and subspaces. *Int. J. on Neural Systems*, 1:61–68, 1989.

[73] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927–935, 1992.

[74] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[75] E. Oja. Nonlinear pca criterion and maximum likelihood in independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, Aussois, France, 1999. To appear.

[76] 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.

[77] E. Oja and L.-Y. Wang. Neural fitting: Robustness by anti-hebbian learning. *Neurocomputing*, 12:155–170, 1996.

[78] E. Oja and L.-Y. Wang. Robust fitting by nonlinear neural units. *Neural Networks*, 9:435–444, 1996.

[79] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–340, May 1996.

[80] J. J. K. O'Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Springer, 1996.

[81] P. Pajunen. Blind separation of binary sources with less sensors than sources. In *Proc. ICNN'97*, Houston, Texas, 1997.

[82] P. Pajunen. Blind source separation using algorithmic information theory. In *Proc. I&ANN*, pages 26–31, 1998.

[83] P. Pajunen, A. Hyvärinen, and J. Karhunen. Nonlinear blind source separation by self-organizing maps. In *Proc. Int. Conf. on Neural Information Processing*, pages 1207–1210, Hong Kong, 1996.

[84] P. Pajunen and J. Karhunen. Least-squares methods for blind source separation based on nonlinear PCA. *Int J Neur Syst*, 1998. To appear.

[85] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 3rd edition, 1991.

[86] B. Pearlmutter and L. Parra. Maximum likelihood blind source separation: A context-sensitive gereralization of ica. In *Advances in Neural Information Processing Systems 9 (NIPS'97)*, 1997.

[87] B.A. Pearlmutter and L.C. Parra. A context-sensitive generalization of ICA. In *Proc. ICONIP'96*, pages 151–157, Hong Kong, 1996.

[88] D.-T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.

[89] C. G. Puntonet, A. Prieto, C. Jutten, M. Rodriguez-Alvarez, and J. Ortega. Separation of sources: A geometry-based procedure for reconstruction of n-valued signals. *Signal Processing*, 46:267–284, 1995.

[90] J.E. Rose, J.F. Brugge, and D.J. Anderson. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of Neurophysiology*, 30(7):769–793, 1967.

[91] M. Scherg and D. von Cramon. Two bilateral sources of the late aep as identified by a spatiotemporal dipole model. *Electroencephalogr Clin Neurophysiol*, 62:232–244, 1985.

[92] J. Särelä, R. Vigário, V Jousmäki, R Hari, and E. Oja. ICA for the extraction of auditory evoked fields. In *4th International Conference on Functional Mapping of the Human Brain (HBM'98)*, Montreal, Canada, 1998.

[93] J. Sun. Some practical aspects of exploratory projection pursuit. *SIAM J. of Sci. Comput.*, 14:68–80, 1993.

[94] L. Tong, V. Soo, R. Liu, and Y. Huang. Amuse: a new blind identification algorithm. In *Proc. ISCAS*, New Orleans, USA, 1990.

[95] The vectorview^{TM} by neuromag, ltd.: a new 306 channel meg device. For more information see: http://www.neuromag.com/vectorview.html.

[96] 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. To appear.

[97] 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.

[98] R.N. Vigário. Extraction of ocular artifacts from eeg using independent component analysis. *Electroenceph. clin. Neurophysiol.*, 103(3):395–404, 1997.

[99] R.N. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographics recordings. In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 229–235. MIT Press, Cambridge MA, 1997.

[100] H. Wang and M. Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Trans. on ASSP*, 33:823–831, 1985.

[101] L. Wang, J. Karhunen, and E. Oja. A bigradient optimization approach for robust PCA, MCA, and source separation. In *Proc. IEEE Int. Conf on Neural Networks '95*, pages 1684–1689, Perth, Australia, 1995.

[102] L.-Y. Wang and J. Karhunen. A unified neural bigradient algorithm for robust PCA and MCA. *Int. J. of Neural Systems*, 7(1):53–67, 1996.

[103] M. Wax and T. Kailath. Detection of signals by information-theoretic criteria. *IEEE Trans. on ASSP*, 33:387–392, 1985.

[104] C.C. Wood and G. McCarthy. Principal component analysis of event-related potentials: simulation studies demonstrate misallocation of variance across components. *Electroenceph Clin Neurophysiol*, 59(3):249–260, 1984.

[105] A. Ziehe, K-R. Müller, G. Nolte, B-M. Mackert, and G. Cuiro. Artifact reduction in magnetoneurography based on time-delayed second order correlations. Technical Report 31, GMD - Forschungszentrum Informationstechnik GmbH, 1998.

[106] 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.

[107] J.J. Zwislocki. Sound analysis in the ear: A history of discoveries. *American Scientist*, 69:184–192, 1981.