Independent Component Analysis of Magnetoencephalographic Signals

Christos Papathanassiou

Submitted for the Degree of
Doctor of Philosophy
from the University of Surrey

Centre for Vision, Speech and Signal Processing
School of Electronics and Physical Sciences
University of Surrey
Guildford, Surrey, GU2 7XH, UK

October 2003

ALL MISSING PAGES ARE BLANK

IN

ORIGINAL

Summary

Magnetoencephalography (MEG) is a non-invasive brain imaging technique which allows instant tracking of changes in brain activity. However, it is affected by strong artefact signals generated by the heart or the eye blinking.

The blind source separation problem is typically encountered in MEG studies when a set of unknown signals, originating from different sources inside or outside the brain, is mixed with an also unknown mixing matrix during their recording. Independent component analysis (ICA) is a recently developed technique which aims to estimate the original sources given only the observed mixtures.

ICA can decompose the observed data into the original biological sources. However, ICA suffers from a major intrinsic ambiguity. In particular, it cannot determine the order of extraction of the source signals. Thus, if there are numerous source signals hidden in lengthy MEG recordings, the extraction of the biological signal of interest can be an extremely prolonged procedure.

In this thesis, a modification of the ordinary ICA is introduced in order to cope with this ambiguity. In case there is prior knowledge concerning one of the original signals, this information is exploited by adding a penalty/constraint term to the standard ICA quality function in order to favour the extraction of that particular signal. Our approach requires no reference signal, but the knowledge of some statistical property of one of the original sources, namely its autocorrelation function. Our algorithm is validated with simulated data for which the mixing matrix is known, and is also applied to real MEG data to remove artefact signals.

Finally, it is demonstrated how ICA can simplify the ill-posed problem of localising the sources/dipoles in the cortex (inverse problem). The advantage of ICA lies in using non-averaged trials. In addition, there is no need to know in advance the number of dipoles.

Key words: Magnetoencephalography, Independent Component Analysis, Artefact Rejection, Inverse Problem, Simulated Annealing.

http://www.ee.surrey.ac.uk

e-mail: c.papathanassiou@eim.surrey.ac.uk

Acknowledgements

I would especially like to thank Dr Blair T Dickson, Dr Mark Smith, and Dr Geoff Barrett from DERA for the financial support of this project, the useful discussions, and their helpful comments. I would also like to thank BBSRC for sponsoring the fees of my course. Special thanks to Dr Line Garnero at the Pitié Salpêtrière Hospital in Paris, France who kindly provided me with the MEG data necessary to perform this study.

I would like to express my extreme gratitude to my supervisor, Prof Maria Petrou, for her constant support and continuous guidance she offered me towards the completion of this thesis.

Many thanks to my housemates, George & George, for helping to preserve my sanity.

To my family in Greece, a simple thanks is not enough to express my unlimited gratitude for the help and support of every kind that they have given me during all this time. Nobody said it would be easy.

Contents

List of Figures

List of Tables

Abbreviations

BEM	Boundary element model
BIT	Brain imaging technique
BSS	Blind source (or signal) separation
cICA	Constrained independent component analysis
CLT	Central limit theorem
CT	Computed tomography
DNA	Deoxyribonucleic acid
ECD	Equivalent current dipole
EEG	Electroencephalography
EMG	Electromyogram
EOG	Electro-oculogram
EPP	Exploratory projection pursuit
EPR	Event-related potential
EVD	Eigenvalue decomposition
FA	Factor analysis
fECG	Fetal electrocardiography
FEM	Finite element model
fMRI	Functional magnetic resonance imaging
ICA	Independent component analysis
IFA	Independent factor analysis
JADE	Joint approximate diagonalisation of eigenmatrices
KL	Kullback-Leibler divergence
MAP	Maximum a posteriori estimate
MEG	Magnetoencephalography
MLE	Maximum likelihood estimate
MNS	Minimum norm solution

MRI	Magnetic resonance imaging
MUSIC	Multiple signal classification
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PDF	Probability density function
PET	Positron emission tomography
QRS	*Cardiac pattern in ECG signals*
rMUSIC	Recursive multiple signal classification
SPECT	Single photon emission computed tomography
SQUID	Superconducting quantum interference device
SNR	Signal-to-noise ratio
SVD	Singular value decomposition

Chapter 1

Introduction

Human brain activity can be recorded using several techniques. Functional brain imaging has to fulfil two primary requirements; the activated brain areas should be localised, and the time of activation during the cognitive process should be determined.

When a participant is asked to perform a specific task, such as to pay attention to an auditory or visual stimulus, small electric currents circulate along the cortical neurons in the activated cerebral regions. These currents generate electric and magnetic fields which are recorded during electroencephalographic (EEG) and magnetoencephalographic (MEG) studies respectively. EEG and MEG are non-invasive methods which can track almost instantly changes in the brain activity. They have been used successfully to image the functional structure of the brain and study the way the brain processes signals, such as those arising from our sense of hearing, sight, touch, as well as those associated with voluntary movements of the body.

Any EEG/MEG experiment has to confront the following fundamental challenges: (a) the removal of artefact fields, and (b) the localisation of the stimulated cortical areas. In section 1.1 these two questions are presented briefly. Previous efforts in the literature concerning these issues are also provided. Section 1.2 states the scope of our work in which a novel signal processing technique called *Independent Component Analysis* is employed to overcome the abovementioned problems. The major achievements of this thesis are presented in section 1.3. Finally, an outline of the thesis is given in section 1.4.

1.1 The Motivation

1.1.1 Artefact Fields in Biomagnetism

The efficiency of both EEG and MEG is limited by a major issue. The cortical neurons in the brain have columnar arrangement allowing the re-enforcement of the fields. The total electric and magnetic fields, which are observed, are the sum of the fields produced by the individual current elements. Nevertheless, the magnetic fields are extremely weak. The environmental noise is typically many orders of magnitude stronger than the biological signals of interest. Detection and measurement of these biomagnetic signals requires magnetic shielding and sensors in special configurations [144]. The measuring device must be also of high intrinsic sensitivity. However, there are many strong artefact fields originated from the human body itself, such as heart interference, eye blinking and horizontal saccades, respiratory movements, and myographic artefacts, which cannot be suppressed by methods of room shielding. The signals produced by the brain are usually a few orders of magnitude weaker than the artefact ones. For example, the heart generates a magnetic field which is two to three orders of magnitude greater than that generated by the brain [138]. The relative ratio between the heart's signal and the brain's contribution depends on the position of the sensor. However, even when the sensors are very close to the skull, the cardiac contamination can easily outweigh the signal of interest. The use of a gradiometer only partially helps because of the heart's proximity to the sensors.

1.1.2 Previous Work on Artefact Rejection

The identification and removal of artefact signals is a complicated procedure. In practice the researchers try to minimize the source of artefacts by asking the participants to fixate their eyes to a target, refrain from blinking, or stop momentarily breathing. However, this is not easy to achieve when the participants are children or patients with neurological disorders. The crudest method of block trial averaging cancels out only random noise. Another simple method is to discard portions of the recordings containing artefacts exceeding a predetermined threshold. However, this subjective tactic demands experience. In addition, it often leads to loss of data of significant importance (especially when the interesting bit of data coincides with ocular artefacts). Moreover, the recorded portions that are kept are not guaranteed to be artefact-free.

More sophisticated methods require the existence of a reference channel for a particular artefact in order to subtract a regression portion. This stimulus-free reference channel can be an electro-oculogram (EOG) for ocular artifacts, or an electromyogram (EMG) for myographic artefacts generated from biting. However, this kind of proportional EOG subtraction distorts the EEG/MEG recording [76].

Artefact fields can also be removed by using known properties, e.g. by exploiting the known spatial structure of the field. The data can be divided into the signal and its orthogonal complement, referred to as the noise subspace. The split of a vector space into several component subspaces of lower-dimensionality can be achieved by applying *singular value decomposition* (SVD). Two matrix projection operators (signal and artefact subspace projection operators) can be formed which project onto the signal and artefact subspaces respectively. By performing a projection of the weight vector onto the signal subspace, the part of the data corresponding to artefact fields can be rejected. The essential requirement for applying signal-space projection is that the undesired fields are known up to unknown multiplicative constants [120, 151]. For example, the eye blinking artefact can be extracted from measurements without the stimulus which is supposed to induce the signal of interest. The source decomposition does not depend on the availability of source or conductivity models.

1.1.3 Source Localisation

The second major problem in MEG signal processing, equally important and complex to artefact removal, is known as *source localisation*, i.e. the spatial localisation of the activated cerebral regions. The principal sources of neuromagnetic fields are considered to be focal current sources in the cerebral cortex. These sources are often represented with a set of equivalent current dipoles (ECD). Complex distributions use several current dipoles [121].

The cortical electric currents are classified in two categories: primary and passive. Primary currents are produced by the neural activity, while passives result from the macroscopic electric field in the conducting cerebral medium. The *forward problem* is to determine the magnetic field produced by the primary currents. The *inverse problem* is to locate these primary currents, and hence the sources of brain activity; in other words, find the sources that best explain the MEG recordings.

The forward problem can be solved both numerically and analytically. Assuming that the head is a spherically symmetric conductor consisting of concentric spheres, each with uniform

isotropic conductivity, analytic solutions for the magnetic field can be found [131]. The key advantage of the spherical model is that the problem can be solved without any knowledge of the conductivity profile. By contrast, in EEG the conductivities must be specified. In addition, the model is extremely fast to compute. More realistic volume-conductor models for the head, approximated by compartments of isotropic and homogeneous conductivities, can also be employed [55]. They provide better localisation accuracy; yet they are usually computationally intensive.

On the other hand, the electromagnetic inverse problem is an ill-posed problem and cannot be solved in a unique way [57]. Even with an infinite number of sensors, there is an infinite number of dipole distributions within the brain which can produce identical external magnetic profiles. The set of equations is highly under-determined. Mathematical equations can point to brain areas where the activity might be. In most cognitive processes, many areas are activated, and thousands of sites can be considered as potential sources. However, there is no way to tell which of the possible inverse solutions is the right one. The problem becomes more intense when the number of the sensors used is smaller than that of the sources. Increased number of sensor measurements enhances the possibility of finding a better solution, as well as the computational workload. The aim is to select physiologically meaningful solutions. Several models have been developed in order to approach the ambiguous inverse problem. These models try to predict the best configuration of active areas in the brain that could best explain the recorded activity. In order to obtain a unique solution, additional assumptions and conditions about the sources must be imposed. The results depend highly on the selected estimation methods. In general, reconstruction is reliable when the number of activated sites is relatively small.

1.1.4 Previous Work on Source Localisation

An approach often used is called the *minimum norm solution* (MNS) [106]. According to this technique, the solution is estimated by minimising a weighted norm of the solution vector. The localisation problem is presented in terms of finding a least-squares fit of a set of dipoles to the recorded data. This method does not take into account temporal information, treating each time sample individually. This technique reconstructs effectively a focal dipole source under ideal noiseless conditions. When the signal is severely distorted by noise, the result is usually a very blurred, unstable reconstruction. Another drawback of this method is that

it does not introduce any anatomical or physiological constraints, and thus it may yield unrealistic solutions with significant localisation errors.

To overcome this problem, signal subspace methods commonly used in array signal processing can be employed, such as the Multiple Signal Classification algorithm (MUSIC) [115], and recursive MUSIC (rMUSIC) [113]. In these techniques, a set of dipoles (or a single dipole when applying rMUSIC) is scanned through a grid confined to a 3D head volume. At each point on this grid, the forward model for this dipole distribution is projected against a signal subspace that has been computed from the MEG data. The locations on the grid which give the best projections correspond to the dipole locations. The problem is how to choose the best projections especially when searching for peaks in 3D. Moreover, the use of a finite set of grid points during the computations and not a continuous one, together with the approximations in the model of the head and the data acquisition system, make the problem more difficult to solve.

Another method that can be employed, introduces a probabilistic model which incorporates anatomical constraints on the location of the activation sites. By restricting the position of the dipoles to a discrete space, the inverse problem can be contained. The potential sources are limited to cerebral cortex. Moreover, because each brain differs in how its cortex is folded, structural MRI scans can be used to adapt the calculations to each brain individually [110]. The probabilistic method is known as the *Bayesian method* [9, 131]. One of the key differences of this approach is that it considers temporal correlation providing temporal smoothness into the solutions. The Bayesian approach provides good results when the activation tends to be sparse and focal.

According to this method, the source configuration S is considered as a random field described by a prior probability $p(S)$. The Bayes theorem states:

$$p(S \mid X) = \frac{p(X \mid S)\, p(S)}{p(X)} \tag{1.1}$$

where $p(X \mid S)$ is the probability of the recorded data set X, conditioned on the source configuration S, and $p(S \mid X)$ is the probability of the source configuration S conditioned on the observed data X (posterior probability).

In consequence, $p(X \mid S)$ expresses the forward model, and $p(S \mid X)$ corresponds to the inverse problem. The source configuration S maximising the posterior probability, called *maximum a posteriori estimate* (MAP), can be used as the estimate of the neural sources. Given the

measurements X, we need only to maximise the product $p(X \mid S)\, p(S)$. All the probability density functions involved can be expressed as exponential energy functions:

$$p(X \mid S) \propto e^{-[U_a(X,S)]} \quad \text{and} \quad p(S) \propto e^{-[U_b(S)]} \tag{1.2}$$

where $U_a(X, S)$ and $U_b(S)$ are some given functions. Hence, the MAP estimator can be found by minimising the posterior energy $U(S, X)$:

$$U(S, X) \equiv U_a(X, S) + U_b(S) \tag{1.3}$$

The *a priori* constraints are introduced in the prior energy $U_b(S)$, which is written as:

$$U_b(S) = U_s(S) + U_t(S) \tag{1.4}$$

where $U_s(S)$ corresponds to spatial priors, and $U_t(S)$ introduces temporal ones.

1.2 Scope of this Study

A novel approach to the identification and removal of artefacts from an MEG scan is the *independent component analysis* (ICA) [29, 155]. ICA is a statistical technique which separates component signals according to measures of their non-Gaussianity. The basic assumption is that of statistical independence between brain and artefact waveforms in terms of production. Independence can be verified by the known differences in the physiological and anatomical origins of these signals. Even in event-related potential studies (EPR), where both cerebral and ocular signals have very close relation during stimulation, independence is assumed throughout the entire signals, and not only during the time when the stimulus is applied. Therefore, the strong relation during stimulus does not affect their global statistical independence. The original independent sources are assumed to be unknown, linearly mixed through propagation before arriving at the sensor array. We have only access to their weighted sums which are expressed as extracranial measurements of biomagnetic fields. ICA aims to estimate these unknown sources.

ICA cannot solve directly the inverse problem. However, it offers the means to simplify the problem in a significant way. ICA estimates the mixing matrix which is used to blend the unknown original sources at the sensor array. The elements of the mixing matrix depend mainly on the relative geometry between sources and sensors in terms of position and orientation. Therefore, ICA can provide invaluable information about the characteristics of each source, which then can be incorporated in a source localisation algorithm.

In this thesis, we will show how ICA can be employed in practical issues such as artefact rejection and source localisation in magnetoencephalography. The purpose of this study can be split into the following tasks:

1. identification and removal of *artefact signals*, such as cardiac interference and ocular artefacts, which may contaminate the MEG recordings;
2. extraction of the *biological signals of interest*, i.e signals produced in the brain due to a stimulus;
3. identification of one or more signals of interest by localising in the cortex the sources (dipoles) which generated those particular signals.

1.3 Achievements of the Thesis

In this thesis, we present a modification of the standard ICA in order to incorporate prior knowledge about one or more source signals. Our algorithm is particularly useful in real world applications, such as in biomedical studies, where some additional information about the signals (artefacts or biological signals) is often known in advance. Hence, the source separation problem is not totally blind. We can exploit this information by introducing a penalty/constraint term to the standard ICA contrast function in order to favour the extraction of a particular signal. Our approach requires no actual reference signal, but the knowledge of some statistic of a model signal, namely its autocorrelation function. However, any other statistical property can be used with the proper modification of the quality function.

Our algorithm is validated successfully with real MEG data. The data which were obtained from the Pitié Salpêtrière Hospital in Paris, France [132], are heavily contaminated with ocular and cardiac artefacts. The researchers typically use a crude method of averaging block trials in order to cancel out the artefacts and proceed with source localisation. However, the data are of such a poor quality that averaging does not result in any practical improvement. In this thesis, we show how ICA can be performed in unaveraged data in order to identify and remove these artefacts. In addition, ICA reduces the need for prolonged data acquisition which may be inconvenient or even not possible as with children, people with neurological disorders, or simply when the phenomenon under examination lasts only a small fraction of time and cannot be reproduced easily.

Finally, although ICA is used mainly for preprocessing, we attempt to apply ICA in unaveraged data in order to facilitate the localisation of sources which are modelled as dipoles in the cortex. Due to lack of essential information about the anatomy of the participant which could be provided with an MRI scan, we are forced to use a simple spherical homogeneous head model instead of a more realistic multi-compartment one. Unfortunately, this limitation is proved to be critical when localising the dipoles with real MEG data. Nevertheless, we present a complete framework in which ICA can be used in simplifying significantly the source localisation problem.

1.4 Thesis Outline

Chapter 2 presents the essentials about MEG. The elementary mechanisms, which are engaged in cellular level for the generation of biomagnetic fields, are described in brief in order to provide a better understanding of the physiological origins of the signal processing challenges confronted in MEG. The most common applications of MEG, both in research and clinical field, are also given.

Chapter 3 provides a solid theoretical background on ICA. First, it explains the details of the blind source separation problem which is often encountered in the signal processing research field. The problem can be summarised as follows: a set of unknown source signals is mixed with an also unknown mixing matrix; the goal is to estimate the original sources given only the observed mixtures. This problem is typically met in EEG/MEG studies when biological signals originating from different sources inside or outside the brain are mixed together during their recording. Therefore, decomposing the observed data into the original sources provides better understanding of the biological processes taking place in the human brain, and allows the removal of undesired artefact signals.

The ICA data model is discussed under a mathematical point of view. The key element in ICA is statistical independence which can be assessed using several measures of non-Gaussianity. The novelty of ICA is that it points towards non-Gaussian signals, whereas in the past only Gaussian signals were considered as interesting. This explains why ICA was developed recently despite being rather simple in concept. All the available measures of non-Gaussianity are presented within a unifying framework. The fundamental conditions are stated, as well as the main intrinsic ambiguities of ICA. Finally, the usefulness of ICA is

demonstrated by providing a diverse series of practical applications.

Chapter 4 presents the practical aspect of ICA. The principal ICA algorithms are compared so as to pick the most appropriate one (FastICA) for our MEG study, based on its appealing attributes in solving real world problems. Some essential preprocessing is provided which also explains the poverty of second-order statistics over higher-order statistics used in ICA. Finally, the chapter concludes with some preliminary results on simulated data in order to demonstrate the efficiency of ICA in decomposing linear mixtures of unknown original sources in noise-free and noisy environments.

Chapter 5 confronts the most annoying indeterminacy of ICA in real world applications. We propose a modification of the standard ICA in order to cope with the intrinsic ambiguity of ICA in the extraction order of the independent components. Usually this is not a serious issue. However, there are situations where this ambiguity poses severe practical obstacles. We should emphasize that the order of extraction is significant in an MEG clinical environment when processing time is an important factor. If there are numerous source signals hidden in lengthy recordings, the extraction of the signal of interest will be an extremely prolonged procedure.

Our algorithm is validated with simulated data for which the true mixing coefficients are known, and is also applied to real MEG data. We show that the incorporation of the constraint allows the extraction of the desired signal as the first independent component under diverse running conditions defined by the initialization step of the algorithm. Once this component is removed from the recordings, the remaining components may be identified in the standard way.

Chapter 6 shows how ICA can be employed to simplify the ill-posed biomagnetic inverse problem. Although ICA does not solve the inverse problem, it provides information which is exploited using a minimum norm solution technique. The main advantage of ICA is that it registers a single dipole to each independent component. Therefore, there is no need to know in advance the exact number of dipoles to be fitted. In addition, the computational workload is significantly reduced because the inverse problem is confronted for that particular source only.

Finally, chapter 7 presents the conclusions of this study. The main contributions of this thesis are clearly outlined. The limitations of our work are also stated. In addition, future research directions towards possible improvements are suggested.

Chapter 2

Magnetoencephalography

This chapter provides the basic elements of magnetoencephalography. Section 2.1 explains the biophysical mechanisms which are employed in the generation and propagation of neural activity in the human brain. A brief introduction in MEG is given in section 2.2. The strong points of MEG are presented in comparison with other brain imaging techniques. Aspects of instrumentation and applications in research and clinical environment are also provided.

2.1 Human Brain Electrophysiology

The human brain is the most complex structure we know. It is often likened to a computer, yet it is much more powerful. This section provides a brief description of the human brain electrophysiology. For a detailed study, see [14, 122, 145].

The nerve cell, also known as *neuron*, is the basic functional unit of the human brain. The role of the neuron is to process information and transmit it to other interconnected neurons in the brain. The nerve cells are able to generate complicated and adaptive behavior.

A typical neuron is shown in figure 2.1. It resembles any other human cell. The neuron consists of a cell body, called *soma*, which contains the typical cell organelles, and a nucleus composed of DNA. A single, long fibre which extends out from the cell body, is called *axon* and carries the information to other cells. All the other fibres are called *dendrites* and used to receive information from the axons of other neurons. The axon can be as long as one meter, whereas the dendrites are less than a millimeter.

Nerve cells are specialized in conducting information to one another. The nerve impulse, also

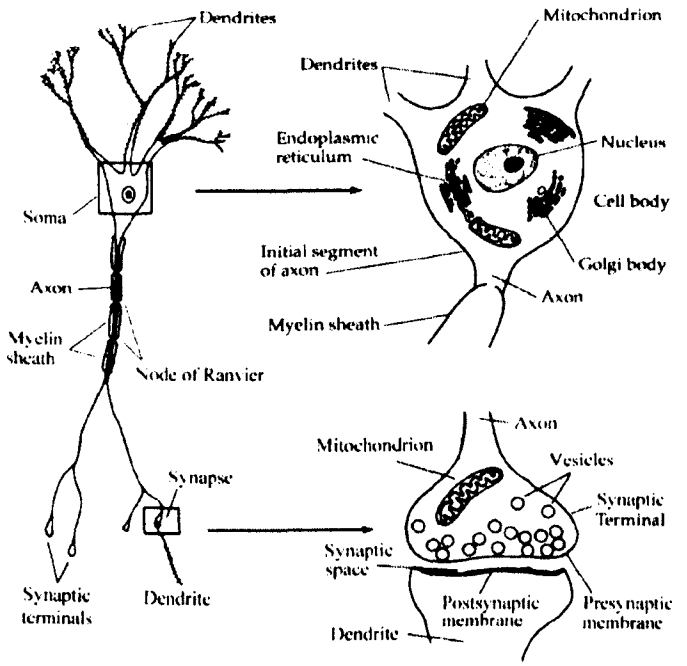


Figure 2.1: Schematic representation of a typical neuron (modified from [145])

called *action potential*, travels along the neuron axon. Rather than an actual electric current, the nerve impulse is a train of changes in the electrical potential of the axon membrane. These changes generate electrical fields that can be recorded.

At rest, the fluid inside the axon contains mostly protein molecules in negatively charged ionic form, and very little positively charged sodium ions. By contrast, the outside substance has little protein, but a considerable amount of sodium. Therefore the inside is negative relative to the outside. The voltage difference is nearly a tenth of a volt.

The axon membrane bears tiny openings, called *sodium channels*, which allow sodium ions Na^+ to pass through. These channels are voltage-gated, and normally closed at rest. However, when the nerve impulse develops, they open briefly to allow sodium ions to rush in, changing the voltage at that little region of the axon to a bit more positive than its normal resting negative value. The positive voltage shift triggers the switches on the immediately adjacent closed sodium channels and pops them open. Sodium ions move out by another mechanism in cell membrane called *ion pump*. It is always at work and operates much more slowly than the nerve impulse.

Once the nerve impulse begins, it travels all the way along the axon. The speed depends on the diameter and other properties of the axon. It ranges from about one mile per hour to 150

miles per hour. Some axons are covered with a myelin sheath, which functions as a kind of electrical insulation. Thus the influx of ions is limited to those areas between each segment of myelin, called *nodes of Ranvier*. This results in nerve impulses travelling many times faster down a myelinated axon than a naked axon of the same size. In the human brain all long axons are myelinated.

The axon branches into a large number of small fibres which have specialized terminals called *synaptic terminals*. Each of these terminals forms a functional connection to a target cell (neuron or other cell) which is called *synapse*. A typical neuron may form several thousand synaptic connections with other neurons. The narrow space between the presynaptic terminal at the very end of the axon and the postsynaptic membrane of the target cell is called *synaptic space*. The presynaptic terminal contains a large number of tiny vesicles with neurotransmitter chemicals.

When the action potential reaches the axon terminals, it sets off a different process. The synapse becomes active, and the vesicles fuse with the presynaptic membrane releasing their contents into the synaptic space. The released molecules attach themselves to receptor molecules on the outside surface of the postsynaptic membrane resulting in the activation of the target cell. Depending on the type of chemical neurotransmitter and receptor molecules, the synaptic action on the neuron excites or inhibits the neuron, increasing or decreasing its activity respectively.

Cell body and dendrites have no voltage-controlled sodium channels. The synaptic connections cause small change in the electrical potential. If there are enough of synaptic excitations to reach the threshold, the membrane of the target cell body becomes more positive, the closed voltage-controlled sodium channels at the beginning of its axon are triggered to open, and the nerve impulse travels along the target neuron. Otherwise the target neuron decides not to transmit the action potential. The whole process takes only a few thousandths of a second.

The major difference of neurons compared to other cells is that no new nerve cells are produced after the birth of its host. Still unknown why that happens, it is assumed that further division and formation of new cells is prohibited in order not to lose already formed patterns of connections. However, new synapses grow and develop all the time. These connections form the circuits and networks in the brain. The human brain is made up of about 10^{11} interconnected neurons, which form at least 10^{14} synapses.

2.2 Magnetoencephalography

2.2.1 Introduction

The first ever reported human magnetoencephalogram was taken by David Cohen in 1968 in a multilayered magnetically shielded chamber, producing evidence of weak alternating magnetic fields generated by alpha rhythm currents [27]. A few years later, Cohen measured the brain's magnetic field at higher sensitivity using for the first time a superconducting quantum interference device (SQUID) magnetometer [28].

The activation of a small region of brain tissue is associated with a primary current. The electric currents flowing in the network of neurons generate a tiny magnetic field. It has been estimated that if about 50,000 similarly oriented adjacent cortical neurons act synchronously, they produce a magnetic field which is large enough to be detected extracerebrally by a SQUID magnetometer [158]. Biomagnetic cerebral fields are exceptionally weak, typically in the range of 50-500 fT in strength [53], which is about 10^{-8} to 10^{-9} of the Earth's geomagnetic field [53], and 10^{-6} to 10^{-7} of the typical urban magnetic noise [5]. In consequence, successful recording of extracranial neuromagnetic signals requires imperatively SQUID technology and usually extensive magnetic shielding. A thorough presentation of MEG, covering multiple aspects of mathematical theory, instrumentation, and applications in human brain studies is given in [52].

2.2.2 Instrumentation of MEG

An MEG scanner is a complex system of relatively high cost, which employs superconducting technology operating at the liquid helium temperature of 4.2K. The biomagnetic fields are detected by a sensing coil which is coupled to a SQUID via a superconducting flux transformer, all submersed in liquid helium or contained within vacuum space. High temperature superconductors, such as that of liquid nitrogen at 77K, still remain too noisy for application in MEG.

When an external magnetic field is imposed, an electric current is produced in the detection coil in order to keep the total magnetic flux through the transformer constant. The secondary magnetic field generated by this current is then sensed by the SQUID. The sensing coils come in various geometries. Their simplest form is a single loop which measures effectively the total

flux through the sensing coil. However, this crude configuration requires exceptionally good shielding from environmental magnetic noise. Contemporary MEG systems make use of gradiometers in which two or more loops (pick-up and compensation loops) are oppositely wound. A gradiometer is insensitive to homogeneous magnetic fields produced by distant sources because these uniform fields impose the same but opposite flux on the pick-up and compensation coil, thus effectively cancelling out themselves. However, if the pick-up coil is close to the head and the distance to the compensation coil is large enough, the neuromagnetic field is essentially sensed by the pick-up coil only. Consequently, the gradiometer successfully discriminates uniform, ambient magnetic fields, and favours the detection of nearby brain sources.

Use of gradiometers improves environmental noise rejection, and thus reduces or even eliminates the need for expensive shielded chambers. A typical magnetically shielded room is constructed from a multilayer combination of aluminium and high permeability iron (mu metal). The aluminium layer attenuates the high frequency band of external magnetic noise by using an eddy current shielding effect. The mu metal effectively screens out low frequencies.

An up-to-date MEG system boasts well over 100 channels which record biomagnetic signals over large portions of the head simultaneously. A typical MEG installation is shown in figure 2.2.

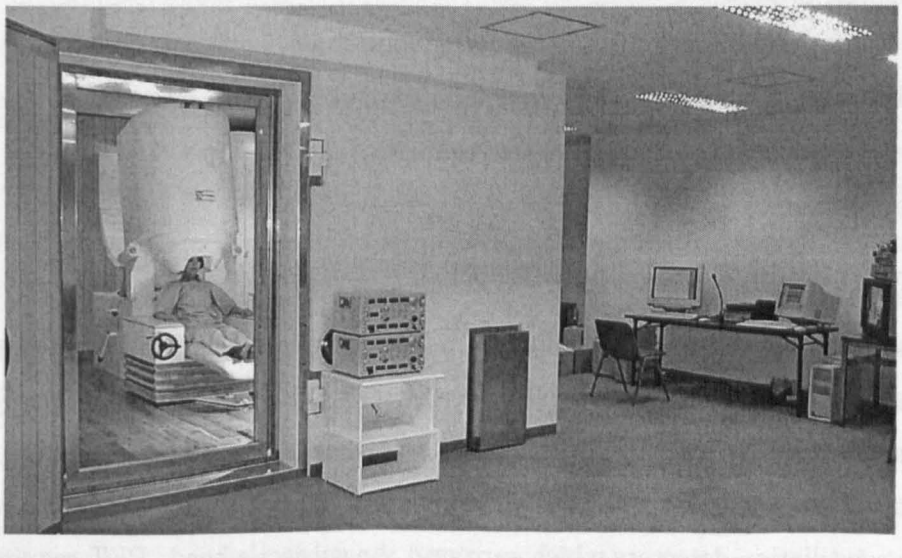


Figure 2.2: A complete installation showing an MEG scanner within a magnetically shielded room, and the acquisition station (courtesy of CTF Systems Inc. [32])

2.2.3 MEG vs Other Techniques

A plethora of brain imaging techniques (BITs) is used nowadays to provide invaluable information about the structure and the function of the human brain. Each BIT has particular strengths and certain drawbacks with respect to revealing the brain anatomy or exploring the brain physiology. This section summarises briefly the principal BITs in comparison with MEG. For a comprehensive review of the topic, see [47, 74, 117, 123, 134, 142].

Computed Tomography (CT)

CT is a widely used, yet costly, digital based, X-ray technique, which offers structural only details of the brain with exceptional spatial resolution. Images of transverse brain slices are acquired when collimated beams of X-rays pass through the head, losing energy in proportion to the density of the various tissues (skull, grey matter, white matter, and cerebrospinal fluid). However, the association of CT with radiation exposure limits the frequency of use.

Magnetic Resonance Imaging (MRI)

MRI derives from a laboratory technique known as nuclear magnetic resonance (NMR). The hydrogen atoms in the brain align themselves during their exposure to an extremely strong magnetic field generated by a superconducting magnet around the subject's head. Application of brief radiowave pulses perturbs the atoms. The re-alignment times are different between gray and white matter resulting in high contrast among brain structures. MRI provides anatomical information only. It is not suitable for subjects suffering from claustrophobia, or patients bearing pacemakers or ferromagnetic implants such as intracranial aneurysm clips. There is also a potential biological hazard, especially to pregnant women, due to the intensity of the magnetic fields to which the tissues are exposed.

Positron Emission Tomography (PET)

PET detects short-lived positron-emitting radiopharmaceuticals which are administered intravenously into the subject, and therefore involves a small amount of radiation. The radioisotope decays producing a positron. After a short distance of a few millimeters, the positron collides with an electron releasing two photons in opposite directions. These photons are detected by scintillation detectors which surround the subject's head. PET requires significant data acquisition time. Moreover, PET demands a substantial capital investment, and thus it is not widely available for clinical use. Nevertheless, PET provides truly quantitative information about the biochemical or physiological processes that take place in the brain, such

as glucose metabolism or the binding of a neurotransmitter molecule to its corresponding receptor.

Single Photon Emission Computed Tomography (SPECT)

SPECT also provides functional brain information on cerebral metabolism and blood flow. SPECT makes use of photon emitting radioisotopes, and thus also suffers from radiation exposure. The instrumentation involves a rotating gamma camera which accumulates multiple angular projections around the subject's head. However, in order to obtain sufficient image counts, the required acquisition time is increased. Moreover, SPECT suffers from photon scattering and variable attenuation due to variations in skull thickness. In consequence, the spatial and temporal resolution of SPECT are inferior to those of PET. On the other hand, SPECT is much less expensive and complex.

Functional Magnetic Resonance Imaging (fMRI)

fMRI assesses local changes in cerebral blood volume, flow and oxygenation levels. These changes depend on the neuronal activity. fMRI makes use of endogenous contrast mechanisms which associate the magnetic resonance relaxation properties of brain tissue to blood flow and blood oxygenation levels. The functional image is obtained by adding information from a stimulated brain image to a routine, non-stimulated brain MRI scan. Thus the functional image provides not only anatomical information but also details of the brain region which is involved in the stimulated state. However, fMRI may not be useful in patients who are subject to pharmacological treatments altering local cerebral blood flow.

Electroencephalography (EEG)

EEG evaluates directly the electrical activity of the brain with small metal electrodes attached on the surface of the scalp. Differential amplifiers are then used to record the time-varying changes of the electrical potential between two electrodes. The electrical potentials are produced by extracellular synaptic trans-membrane currents in neuronal dendrites. EEG records mainly electrical currents perpendicular to the skull. The skull can be considered as a passive volume conductor. Thus the transmission of electrical activity is effectively instantaneous (or more accurately, determined by the speed of light in cerebral tissue). Therefore EEG has an exceptional temporal resolution. However, EEG is vulnerable to distortions of the electrical potentials, created by the inhomogeneous tissue conductivity inside the skull. In consequence, EEG requires detailed knowledge of head geometry and conductivity for all cerebral tissues.

The cost of EEG equipment is much less than that of MEG. However, EEG requires longer preparation time in placing the electrodes on the scalp.

In summary,

- MEG is a completely non-invasive imaging technique which poses no biological risk. Other BITs expose the subject to X-rays (CT), radioactive tracers (PET, SPECT), or intense magnetic fields (MRI, fMRI).
- MEG provides invaluable functional information, while some BITs offer only structural details (CT, MRI).
- MEG makes *in vivo* direct measurements of brain electrophysiology, whereas other BITs assess quantities associated with neuronal activity, such as cerebral blood volume, blood flow, or oxygenation levels (PET, SPECT, fMRI).
- MEG is capable of almost instant tracking of neural activity, and thus provides exceptional temporal resolution, lower than a millisecond. On the other hand, the blood flow changes measured by PET, SPECT, and fMRI take several seconds to develop.
- MEG has short data acquisition times. By contrast, PET and SPECT are relatively slow in collecting the data taking up to a minute or 10 minutes respectively, while fMRI needs a few seconds. Slow data acquisition is not only time-consuming, but it can also distort the results since the cognitive process which is under examination lasts only a fraction of the data acquisition time. In order to overcome this problem, researchers perform block trials in which the subject performs a string of similar tasks causing the brain to repeat the same mental process while the data are gathered. However, block trials are not effective when an element of surprise is involved.
- MEG provides a spatial resolution comparable to that of PET and fMRI. Nevertheless, MEG cannot spot directly the source of the biomagnetic signal, but requires use of a mathematical model.
- MEG has relatively short preparation times, and can be performed by a technician with a minimum of training. PET and SPECT require specialised scientists in order to produce and administer the radiopharmaceuticals, and make the measurements.
- MEG has a high cost of equipment, yet comparable to that of CT, PET, and MRI.

A graphic comparison of spatiotemporal resolution for different functional BITs is given in figure 2.3. The graph is based on data taken from [117].

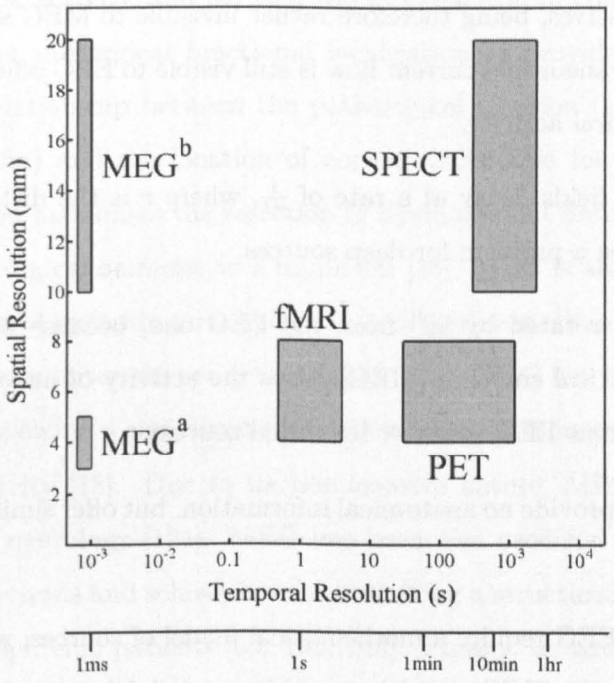


Figure 2.3: Comparison of spatiotemporal resolution of PET, SPECT, fMRI, and MEG, based on data presented in [117]. MEG^a refers to sources located in the cortex, and MEG^b to subcortical sources. MEG is rather insensitive to deep cerebral activity.

In comparison with EEG,

- MEG makes absolute measurement of magnetic fields, while EEG measures the potential difference between two electrode positions.
- The magnetic fields detected by MEG are mainly produced by high-density intracellular currents (primary currents), whereas electrical potentials are generated by volume extraneuronal currents (secondary currents). The latter depend on the conductivity properties of the head. Although the properties of the magnetic fields depend on both primary and secondary currents, the volume current contamination is usually small. When homogeneous spherical head models are used, the contribution of volume currents can be ignored. Thus, MEG is not affected by tissue conductivity anisotropy as much as EEG does. Neuromagnetic signals penetrate the scalp without significant distortion. Moreover, this relative transparency allows detection of very low or very

high frequency neuronal activity which is difficult to record electrically.

- Neurons oriented at random in deep subcortical regions produce magnetic fields which cancel out themselves, being therefore rather invisible to MEG sensors (see also figure 2.3), but net extraneuronal current flow is still visible to EEG offering more information about deep cerebral activity.

- Electromagnetic fields decay at a rate of $\frac{1}{r^2}$, where r is the distance from the source, and this might be a problem for deep sources.

- MEG pattern is rotated by 90^0 from the EEG one, because magnetic fields are orthogonal to electrical currents. MEG senses the activity of nerve cells with tangential orientation, whereas EEG is better for radial sources.

- Both techniques provide no anatomical information, but offer similar, excellent temporal resolution.

- Both MEG and EEG require a mathematical model of sources, widely known as *equivalent current dipole* (ECD) model, to achieve spatial localisation. They both suffer from an ill-posed problem, known as the *inverse problem*, in which there is an infinite number of possible source configurations which are consistent with the electromagnetic recordings.

- MEG has shorter preparation time, and uses more sensors over EEG, thus offering potentially increased spatial resolution. A typical modern MEG system has more than 100 channels, whereas high density EEG makes use of 20-30 scalp electrodes. On the other hand, EEG equipment costs a fraction of MEG.

The future of brain imaging lies in multi-modality integration [59, 35]. PET and fMRI complement MEG functional imaging. Combined together result in optimal spatiotemporal resolution. They can provide MEG with helpful physiological and anatomical constraints which are then used as an initial guess for the position of ECDs, thus reducing drastically the number of possible solutions for the ill-posed inverse problem [96].

2.2.4 Applications of MEG

By offering instantaneous assessment of cerebral activity, MEG has been used extensively in cognitive brain research, such as studies of cortical rhythms, somatosensory research,

investigation of cortical responses to visual and auditory stimuli, attention and memory examinations [97, 118]. The development of large-array, whole-cortex MEG systems improved localisation of cortical activity, and led to a recent expansion of clinical applications. Thus, MEG can be used in presurgical functional localisation to provide important information about the spatial relationship between the pathological location (such as brain tumour or vascular malformation) and the location of cortex responsible for specific brain functions [150]. This evaluation maximises the resection of dysfunctional brain tissue, while keeps the post-operative neurological damages to a minimum [38]. MEG is also employed for the non-invasive identification of epileptic cortical foci, reducing or even eliminating the need for depth electrodes in order to monitor seizure activity [103, 128]. MEG has been also used in cortical ischaemia studies, revealing a significant correlation between reduced regional cerebral blood flow and neural activity [18]. Due to its non-invasive nature, MEG has been successfully applied in pediatric neurology [124]. MEG has been also used for the study of psychiatric disorders such as psychosis and schizophrenia, indicating a structural change in the left brain hemisphere of schizophrenic patients [50, 135, 146]. Finally, an exciting clinical field is the application of MEG in fetal medicine to assess the neurological status of the fetus during pregnancy. Both fetal visual and auditory evoked responses have been effectively detected [40, 140, 160].

Chapter 3

The Theory of Independent Component Analysis

This chapter provides the essential theoretical background of Independent Component Analysis (ICA). It describes the problem in its simple noise-free form in sections 3.1 and 3.2. The fundamental assumptions of the ICA data model are stated in section 3.3. However, ICA suffers from some indeterminacies which are given in section 3.4. The key point of ICA is statistical independence which is rigorously defined in section 3.5. The most commonly used measures of independence are provided in section 3.6 within a unifying framework. Finally, some practical applications of ICA in real world problems are briefly presented in section 3.7. A thorough presentation of ICA, covering multiple aspects of mathematical theory, practical algorithms, and applications is given in [68, 69, 92].

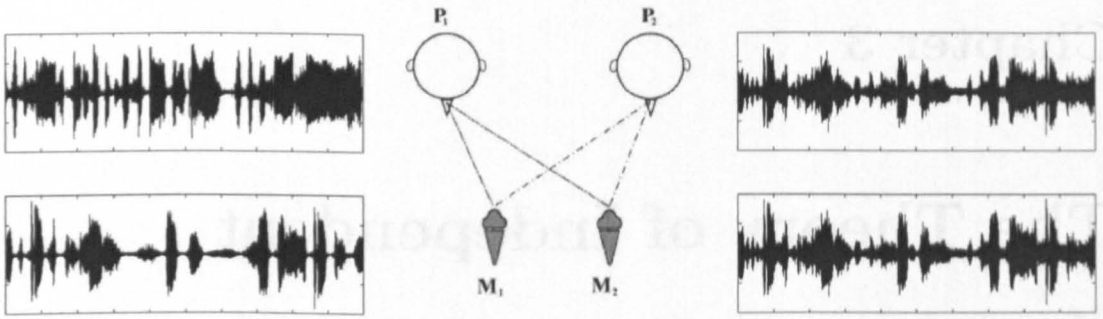
3.1 The Problem

Let us consider two persons, P_1 and P_2, in a room speaking simultaneously, and two microphones, M_1 and M_2, recording their speeches (see figure 3.1). The recorded signals, x_1 and x_2, represent a linear mixture of the original speech signals, s_1 and s_2, described by the following set of equations:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t)$$

where t is the time variable. Parameters a_{ij} depend on the distance of the microphones from the speakers. Assume also that the position of the speakers and microphones remains constant over time.



(a) original speeches (b) the problem (c) microphone recordings

Figure 3.1: Graphical representation of the *cocktail-party* problem. The speeches of two persons, P_1 and P_2, are recorded by two microphones, M_1 and M_2.

Our goal is to estimate the original speech signals, s_1 and s_2, using only the recorded ones, x_1 and x_2. This is known as the *blind source* (or *signal*) *separation* (BSS) problem. It is often called as the *cocktail-party* problem. The parameters a_{ij} are considered unknown, and therefore we must use information on the statistical properties of the signals s_i in order to estimate a_{ij}. A detailed study of the statistical principles and practical algorithms of BSS is provided in [21, 56].

3.2 Linear ICA Data Model

In the general problem, assume N unknown signals s_i $(i = 1, 2, \ldots, N)$ (often referred as *latent variables*). Consider M linear mixtures x_j $(j = 1, 2, \ldots, M)$ of the original sources s_i :

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jN}s_N, \quad \text{for all } j = 1, 2, \ldots, M \tag{3.1}$$

In order to make more explicit the dependence of the signals on time, equation 3.1 can be rewritten as

$$x_{jk} = a_{j1}s_{1k} + a_{j2}s_{2k} + \ldots + a_{jN}s_{Nk} \tag{3.2}$$

where the time variable has been incorporated into the signals x_j and s_i as the second index k ($k = 1, 2 \ldots, K$, where K is the total number of time slices).

Thus, the j^{th} mixture x_j at the k^{th} time slice is given in a compact way by

$$x_{jk} = \sum_{i=1}^{N} a_{ji} s_{ik} \tag{3.3}$$

where the parameters a_{ji} are unknown, but constant over time.

Using vector-matrix notation, the linear mixing model can be formulated for a particular time slice k as

$$\mathbf{x}_k = A\mathbf{s}_k \tag{3.4}$$

where $\mathbf{x}_k$ and $\mathbf{s}_k$ are M- and N-dimensional column vectors respectively for the k^{th} time slice:

$$\mathbf{x}_k = [x_{1k} \;\; x_{2k} \;\; \cdots \;\; x_{Mk}]^T \quad , \quad \mathbf{s}_k = [s_{1k} \;\; s_{2k} \;\; \cdots \;\; s_{Nk}]^T$$

and $A = [a_{ji}]$ is the $(M \times N)$, constant over time, unknown mixing matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{pmatrix}$$

Independent Component Analysis (ICA) is a relatively new concept which can be employed in performing blind source separation [29, 80]. ICA is a linear transformation method which intends to minimise the statistical dependence of the components of the new representation. The generative model of equation 3.4 is also known as *noise-free* ICA model because no noise term has been considered.

The aim of ICA is to estimate both unknown A and s_{ik} using known x_{jk} only. If we estimate the mixing matrix A, then it is easy to compute its inverse (or pseudo-inverse if A is not a square matrix), say W, and from that the original signals as $\hat{s}_{ik}$.

3.3 Essential Assumptions in ICA

ICA assumes the following restrictions [29]:

- The fundamental assumption of ICA is that all s_i's are mutually statistically independent at each time instant k. ICA exploits the information of independence to separate the original sources s_i. Statistical independence is defined rigorously in section 3.5.

- Another essential requirement is that the original sources are non-Gaussian. To be precise, the ICA data model is still identifiable when only one of the s_i's is Gaussian. When there are more than one Gaussian original sources, then the mixing matrix A can be estimated up to an orthogonal transformation. The underlying connection of non-Gaussianity with statistical independence is discussed thoroughly in section 3.6.

- Ideally, the number M of the observed linear mixtures x_j should be at least as large as the number N of the original sources s_i (i.e. $M \geq N$). In practice, for the sake of simplicity, many ICA algorithms assume that $M = N$ without loss of generality. If $M < N$, the problem is described as *overcomplete*, and the mixing matrix A may still be identifiable [19].

- The mixing matrix A must be of full rank and constant over time.

- Ideally, there is no sensor noise. However, original sources with low additive noise are allowed.

3.4 Ambiguities of ICA

The ICA data model suffers from the following ambiguities due to the fact that both A and $\mathbf{s}$ are unknown.

- The scale and the sign of the original sources cannot be determined. Indeed, if p_i denotes a scalar multiplier for each of the original sources s_i, then p_i can be cancelled out by dividing the corresponding i^{th} column A_i of the mixing matrix A by the same scalar p_i as follows:

$$x_{jk} = \sum_{i=1}^{N} a_{ji} s_{ik} = \sum_{i=1}^{N} \frac{a_{ji}}{p_i} p_i s_{ik} = \sum_{i=1}^{N} a'_{ji} s'_{ik}, \quad \text{for all } j = 1, 2, \ldots, M$$

where $a'_{ji} = \frac{a_{ji}}{p_i}$ is an element of the new $(M \times N)$ mixing matrix A' (assumed unknown), and $s'_{ik} = p_i s_{ik}$ is the original source s_i expressed in different scale.

A practical way to remove this ambiguity is to consider the magnitudes of the original sources s_i fixed by putting on each of them the restriction of unit variance, i.e. $E\{s_i^2\} = 1$. However, the ambiguity of the sign remains [29].

- The order of extraction of the original sources cannot be estimated. Indeed, if P is an $(N \times N)$ permutation matrix, then

$$\mathbf{x} = A\mathbf{s} = A(P^{-1}P)\mathbf{s} = (AP^{-1})(P\mathbf{s})$$

where AP^{-1} can be considered as a new mixing matrix (assumed unknown), and the elements of $P\mathbf{s}$ are the original sources s_i in different order.

3.5 Definition and Properties of Statistical Independence

Two random variables, y_1 and y_2, are considered to be statistically independent when knowledge about the value of y_1 does not yield any information on the value of y_2, and vice versa. In simple maths, y_1 and y_2 are independent if and only if their joint probability density function (PDF) $P(y_1, y_2)$ can be factorised [125]:

$$P(y_1, y_2) = P_1(y_1)P_2(y_2) \tag{3.5}$$

where $P_1(y_1) = \int P(y_1, y_2)dy_2$ and $P_2(y_2) = \int P(y_1, y_2)dy_1$ are the marginal densities of y_1 and y_2 respectively.

Similarly, if we consider N random variables y_i $(i = 1, 2, \ldots, N)$, then the variables y_i are mutually independent if and only if

$$P(\mathbf{y}) = \prod_{i=1}^{N} P_i(y_i) \tag{3.6}$$

where $P(\mathbf{y})$ is the joint density of $\mathbf{y} = [y_1 \;\; y_2 \;\; \ldots \;\; y_N]^T$, and $P_i(y_i)$ is the marginal density of y_i.

A very important property of independent variables comes as a consequence:

$$E\{f_1(y_1)f_2(y_2)\} = E\{f_1(y_1)\}E\{f_2(y_2)\} \tag{3.7}$$

for any functions f_1 and f_2, where $E\{.\}$ denotes the expectation value.

For example,

$$E\{y_1y_2\} = E\{y_1\}E\{y_2\} \tag{3.8}$$

The covariance $C_{y_1y_2}$ of y_1 and y_2 is defined as:

$$C_{y_1y_2} = E\{(y_1 - \overline{y}_1)(y_2 - \overline{y}_2)\} \tag{3.9}$$

where $\overline{y}_1$ and $\overline{y}_2$ are the means of y_1 and y_2 respectively.

Therefore,

$$C_{y_1y_2} = \int\int (y_1 - \overline{y}_1)(y_2 - \overline{y}_2)P(y_1, y_2)dy_1dy_2 = \int\int y_1y_2P(y_1, y_2)dy_1dy_2 - \overline{y}_1\overline{y}_2 \Rightarrow$$

$$C_{y_1y_2} = R_{y_1y_2} - \overline{y}_1\overline{y}_2 \tag{3.10}$$

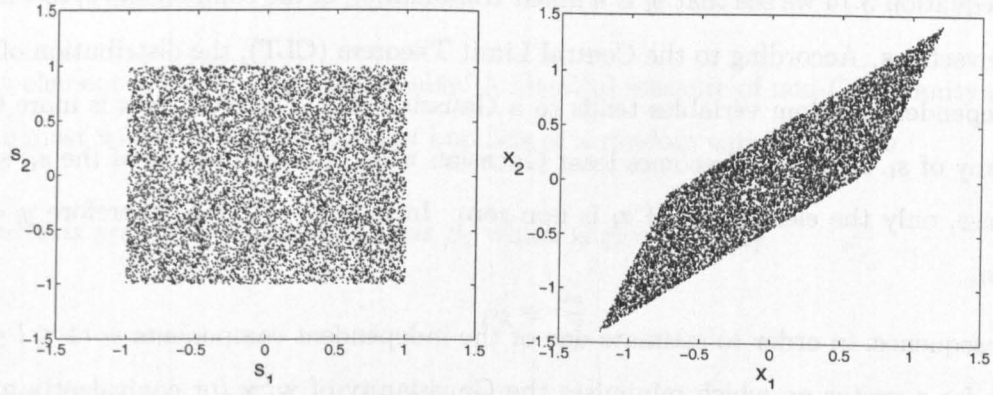
where $R_{y_1y_2} = E\{y_1y_2\}$ is the correlation between y_1 and y_2.

Two random variables, y_1 and y_2, are said to be uncorrelated if $R_{y_1y_2} = E\{y_1\}E\{y_2\}$. In consequence, for two uncorrelated variables: $C_{y_1y_2} = 0$.

From equation 3.8 we realise that two independent variables are also uncorrelated. Several ICA algorithms use this remark to simplify the estimation procedure by calculating uncorrelated estimates of the independent components. In general, uncorrelatedness is a weaker form of independence. Uncorrelatedness does not imply independence.

However, independence and uncorrelatedness are equivalent when the random variables are Gaussian [29]. Therefore, any decorrelating transformation of the variables also yields a set of independent components, and hence the mixing matrix A is not identifiable.

Consider two original sources, s_1 and s_2, which follow uniform distributions. The sources are statistically independent because the values of s_1 do not yield any information about s_2, as it can be seen from their scatter-plot in figure 3.2. However, if the original sources are linearly mixed (with a (2×2) matrix A), the mixed signals, x_1 and x_2, are now dependent. For example, if x_1 attains one of its maximum or minimum values, then the value of x_2 can be determined. In a geometrical sense, ICA can be considered as a rotation of the mixed signals such that the axis of the estimated sources have the same direction as the axis of s_1 and s_2.



(a) original sources (b) mixed signals

Figure 3.2: Scatter-plots of the original sources (s_1,s_2), and the mixed signals (x_1,x_2).

3.6 Criteria of Statistical Independence

Consider a linear combination y_l of the components x_j of the observed signal vector $\mathbf{x}$. Thus, the value of y_l for the k^{th} time slice is:

$$y_{lk} = w_{l1}x_{1k} + w_{l2}x_{2k} + \ldots + w_{lM}x_{Mk} \Rightarrow y_{lk} = \sum_{j=1}^{M} w_{lj}x_{jk} \tag{3.11}$$

where the parameters w_{lj} ($1 \leq l \leq N$ and $j = 1, 2, \ldots, M$) are constant over time.

Therefore, according to the ICA data model in equation 3.3:

$$y_{lk} = \sum_{j=1}^{M}\left(w_{lj}\sum_{i=1}^{N} a_{ji}s_{ik}\right) \Rightarrow y_{lk} = \sum_{i=1}^{N}\left[\left(\sum_{j=1}^{M} w_{lj}a_{ji}\right)s_{ik}\right] \Rightarrow y_{lk} = \sum_{i=1}^{N} z_{li}s_{ik} \tag{3.12}$$

where $z_{li} = \sum_{j=1}^{M} w_{lj}a_{ji}$. For the sake of simplicity, we drop the time index k. Hence,

$$y_l = \sum_{i=1}^{N} z_{li}s_i \tag{3.13}$$

Using vector-matrix notation:

$$y_l = \mathbf{w}_l^T\mathbf{x} \Rightarrow y_l = \mathbf{w}_l^T A\mathbf{s} \Rightarrow y_l = \mathbf{z}_l^T\mathbf{s} \tag{3.14}$$

where $\mathbf{w}_l = [w_{l1} \;\; w_{l2} \;\; \cdots \;\; w_{lM}]^T$ and $\mathbf{z}_l = [z_{l1} \;\; z_{l2} \;\; \cdots \;\; z_{lN}]^T$. Note that $\mathbf{z}_l = A^T\mathbf{w}_l$.

If $\mathbf{w}_l^T$ is the l^{th} row of the (pseudo)inverse matrix A^{-1}, then the linear combination $\mathbf{w}_l^T\mathbf{x}$ will yield the l^{th} independent component s_l. The purpose of ICA is to estimate the weighting vector $\mathbf{w}_l$, and compute the estimate $\hat{s}_l$ of each constituent component s_l in turn ($l = 1, 2, \ldots, N$).

From equation 3.14 we see that y_l is a linear combination of the components s_i of the original source vector $\mathbf{s}$. According to the Central Limit Theorem (CLT), the distribution of the sum of independent random variables tends to a Gaussian [81]. Therefore, $\mathbf{z}_l^T\mathbf{s}$ is more Gaussian than any of s_i. In fact, it becomes least Gaussian when it equals to one of the s_i, say s_l. In this case, only the element z_{ll} of $\mathbf{z}_l$ is non-zero. In fact, $z_{ll} = 1$, and therefore y_l coincides with s_l.

In consequence, in order to estimate one of the independent components s_l $(1 \leq l \leq N)$, we search for a vector $\mathbf{w}_l$ which minimises the Gaussianity of $\mathbf{w}_l^T\mathbf{x}$ (or equivalently maximises the non-Gaussianity of $\mathbf{w}_l^T\mathbf{x}$). Thus, the complete set of N independent components s_i $(i = 1, 2, \ldots, N)$ can be extracted by estimating N vectors $\mathbf{w}_i$ which lead to local maxima of non-Gaussianity for $\mathbf{w}_i^T\mathbf{x}$. Table 3.1 summarises the most commonly used measures of non-Gaussianity.

The vector $\mathbf{w}_i^T$ defines the i^{th} row of the demixing matrix W which is an estimation of A^{-1}. Thus, matrix $Q = WA$ provides a practical degree of the separation quality, and is often called *performance matrix*. Due to the ambiguities of scaling and permutation described in section 3.4, the performance matrix of a successful source separation should be close to the identity matrix after normalising and reordering.

The linear mixing model of equation 3.4 and the ICA separation of the mixtures x_j to the original sources s_i is presented graphically in figure 3.3.

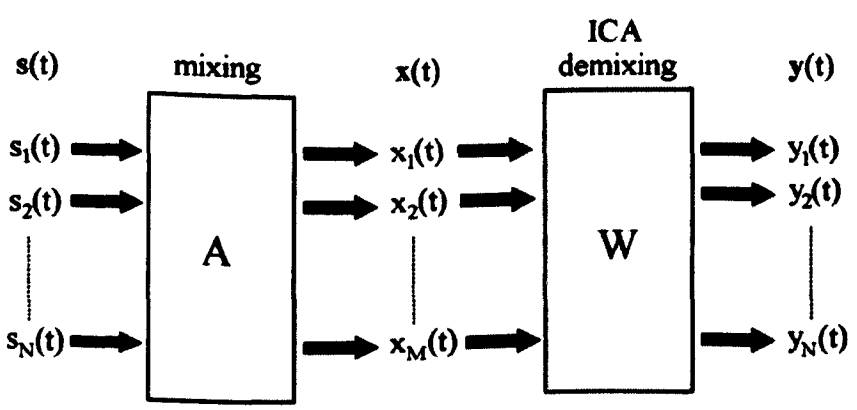


Figure 3.3: Schematic representation of mixing and unmixing processes. $\mathbf{s}(t)$: original sources, $\mathbf{x}(t)$: observed mixtures, and $\mathbf{y}(t)$: estimated source signals. A: mixing matrix, and W: unmixing matrix (estimation of A^{-1}). Matrix WA defines the performance matrix Q.

3.6.1 Kurtosis

The key element in ICA is non-Gaussianity. A classical measure of non-Gaussianity is *kurtosis*. The most widely used definitions of kurtosis of a random variable y are:

- *kurtosis proper* or *Pearson kurtosis* β_2 which is given by [82]

$$\beta_2 \equiv \frac{\mu_4}{\mu_2^2} \tag{3.15}$$

- *kurtosis excess* γ_2 which is defined as [82]

$$\gamma_2 \equiv \frac{\mu_4}{\mu_2^2} - 3 \tag{3.16}$$

  where μ_i denotes the i^{th} *central moment* of y:

$$\mu_i \equiv E\{(y-\mu)^i\} = \int (y-\mu)^i P(y) dy \tag{3.17}$$

  where μ is the mean of y (also known as the first *raw moment* of y):

$$\mu = \mu_1 \equiv E\{y\} = \int y P(y) dy \tag{3.18}$$

  The second central moment μ_2 of y is better known as the *variance* σ^2 of y.

Kurtosis is regularly associated with the *fourth order cumulant* κ_4 of y which can be expressed in terms of central moments as

$$\kappa_4 = \mu_4 - 3\mu_2^2 \tag{3.19}$$

Recall that in order to remove the ambiguity of the scale of the extracted independent components in ICA, we consider that each component has unit variance, i.e. $\mu_2 = 1$. In this case,

$$\gamma_2 = \kappa_4 = \mu_4 - 3 \tag{3.20}$$

Cumulants involving at least two different random variables are called *cross-cumulants*. For random variables y_i, y_j, y_k, y_l (assumed to be zero-mean for the sake of simplicity), the second order cross-cumulants are given by

$$k_2(y_i, y_j) = E\{y_i y_j\}$$

whereas the fourth order cross-cumulants are

$$k_4(y_i, y_j, y_k, y_l) = E\{y_i y_j y_k y_l\} - E\{y_i y_j\}E\{y_k y_l\} - E\{y_i y_k\}E\{y_j y_l\} - E\{y_i y_l\}E\{y_j y_k\}$$

Note that when $i = j = k = l$, the fourth order cross-cumulant reduces to $\kappa_4 = \mu_4 - 3\mu_2^2$ of a single variable as in equation 3.19. In addition, if y_i and y_j are uncorrelated, then $k_2(y_i, y_j) = 0$.

An appealing property of the cumulants is their multilinearity. If we consider a linear transformation $\mathbf{y} = W\mathbf{x}$ (or analytically, $y_i = \sum_{p=1}^{M} w_{ip}x_p$ with $i = 1, 2, \ldots, N$), then, for example, the fourth order cross-cumulants of the new representation are given by

$$\kappa_4(y_i, y_j, y_k, y_l) = \sum_{p,q,r,s=1}^{M} w_{ip}w_{jq}w_{kl}w_{ls}\,\kappa_4(x_p, x_q, x_r, x_s), \quad 1 \le i, j, k, l \le N \tag{3.21}$$

If $\mathbf{s}$ is the N-dimensional vector of the original independent sources, then

$$\kappa_4(s_p, s_q, s_r, s_s) = \kappa_4(s_p)\,\delta(p, q, r, s) \tag{3.22}$$

and the cross-cumulants for the ICA data model, $\mathbf{x} = A\mathbf{s}$, are

$$\kappa_4(x_i, x_j, x_k, x_l) = \sum_{u=1}^{N} a_{iu}a_{ju}a_{ku}a_{lu}\,\kappa_4(s_u) \tag{3.23}$$

where a_{ij} the $(i,j)^{\text{th}}$ element of mixing matrix A.

From now on, whenever we use kurtosis, we will imply kurtosis excess. In practice for a sample of N values of y, kurtosis can be computed as

$$\gamma_2 = \frac{\sum_{i=1}^{N}(y_i - \mu)^4}{N\sigma^4} - 3 \tag{3.24}$$

where y_i is the i^{th} value ($i = 1, 2, \ldots, N$), μ is the mean, and σ^2 is the variance of y.

Kurtosis describes the *peakedness* of a distribution relative to the normal distribution. If the distribution has a flattened shape, then it is known as *platykurtic* or *sub-Gaussian*, and its kurtosis is negative. On the other hand, if the distribution has a sharp peak near the mean with heavy tails, then it is called *leptokurtic* or *super-Gaussian*, and its kurtosis is positive. The kurtosis of a Gaussian random variable is zero, and the distribution is often referred to as *mesokurtic*. In consequence, the absolute value of kurtosis (or alternatively its square) can be used to assess non-Gaussianity.

In figure 3.4 we present a typical example of a super-Gaussian and a sub-Gausssian distribution compared with the normal distribution (all distributions in the graph are of unit variance). The Laplace distribution is leptokurtic with $\gamma_2 = 3$, whereas the continuous uniform distribution is platykurtic with $\gamma_2 = -\frac{6}{5}$.

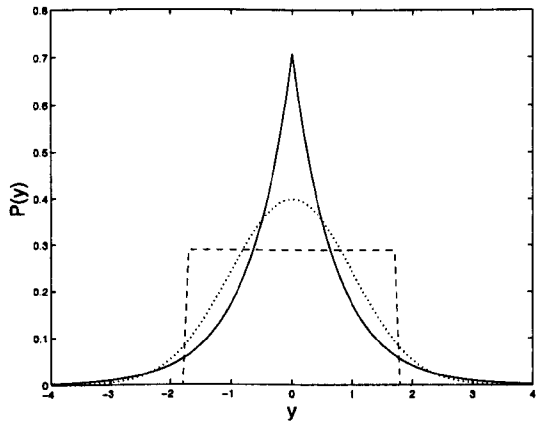


Figure 3.4: Probability density function of a Gaussian distribution (dotted line), Laplace distribution which is super-Gaussian (solid line), and continuous uniform distribution which is sub-Gaussian (dashed line). All distributions are normalised at unit variance.

The main practical drawback of kurtosis is that it can be sensitive to outliers when it is calculated from a finite sample according to equation 3.24. Therefore, in practice kurtosis may not be a very robust measure of non-Gaussianity.

3.6.2 Negentropy

The *Kullback-Leibler (KL) divergence* D_{KL} between two probability densities P and Q (also called the *relative entropy* of P with respect to Q) is defined as [30]

$$D_{KL}(P(y)\|Q(y)) \equiv \int P(y) \log \frac{P(y)}{Q(y)} dy \tag{3.25}$$

The KL divergence is a measure of the distance between two probability distributions. It is always non-negative, and equals zero only if $P = Q$.

The KL divergence is employed to define *negentropy* J of a random vector $\mathbf{y}$ as [30]

$$J(\mathbf{y}) \equiv D_{KL}(P(\mathbf{y})\|P_G(\mathbf{y}_G)) = \int P(\mathbf{y}) \log \frac{P(\mathbf{y})}{P_G(\mathbf{y}_G)} d\mathbf{y} \tag{3.26}$$

where $\mathbf{y}_G$ follows a Gaussian distribution with density P_G of the same mean and covariance matrix as $\mathbf{y}$.

Negentropy (also known as *negative entropy*) is closely related to *entropy* which is a fundamental quantity of information theory. The entropy of a random variable y_i offers a degree of disorder and lack of structure for that variable.

The entropy H (often called *Shannon differential entropy*) of a continuous random variable y with density $P(y)$ is defined as [30]

$$H(y) \equiv -\int P(y) \log P(y) dy \tag{3.27}$$

Using the definition of entropy in equation 3.27, negentropy can be expressed as

$$J(y) = H(y_G) - H(y) \tag{3.28}$$

Consequently, entropy and negentropy differ only by a constant and the sign.

It can be proven that among all random variables of equal variance, a Gaussian one has the largest entropy [30]. In consequence, negentropy is always non-negative, and zero if and only if y follows a normal distribution. Hence, negentropy can be used to assess non-Gaussianity.

However, negentropy suffers from a major limitation. It requires knowledge of the density function in order to be computed. In practice, the computational complexity of negentropy can be resolved by implementing approximations.

One way to approximate negentropy is to use higher-order cumulants as follows [77]

$$J(y) \approx \frac{1}{12}\kappa_3^2(y) + \frac{1}{48}\kappa_4^2(y) \tag{3.29}$$

where $\kappa_i(y)$ is the i^{th} order cumulant of y (y is here assumed to be of zero mean and unit variance, i.e. $\mu(y) = 0$ and $\sigma^2(y) = 1$). Nevertheless, these approximations are affected by the same problem of limited robustness as kurtosis.

An alternative way to approximate the negentropy of a random variable y (with $\mu(y) = 0$ and $\sigma^2(y) = 1$) is given by [65]:

$$J(y) \approx \sum_{j=1}^{P} c_j \left[E\{G_j(y)\} - E\{G_j(y_{G\nu})\}\right]^2 \tag{3.30}$$

where c_j are positive constants, G_j are non-quadratic functions of y, and $y_{G\nu}$ is a standardised Gaussian variable (i.e. $\mu(y_{G\nu}) = 0$ and $\sigma^2(y_{G\nu}) = 1$).

Equation (3.30) takes its simplest form when only one non-quadratic function G is used:

$$J(y) = c\left[E\{G(y)\} - E\{G(y_{G\nu})\}\right]^2 \tag{3.31}$$

Note that if we opt for $G(y) = y^4$ and $c = 1$, equation 3.31 is rewritten as

$$J(y) = \left(E\{y^4\} - E\{y_{G\nu}^4\}\right)^2 = \left(\mu_4(y) - \mu_4(y_{G\nu})\right)^2 \overset{(3.20)}{=} \left(\gamma_2(y) - \gamma_2(y_{G\nu})\right)^2 \Rightarrow$$

$$J(y) = \gamma_2^2(y) \tag{3.32}$$

Equation (3.32) shows clearly that kurtosis γ_2 can be actually incorporated into the general framework of negentropy. However, other choices of G can provide more attractive statistical properties than kurtosis, such as robustness to outliers. For example [66],

$$G(y) = \frac{1}{\alpha_1}\log\cosh(\alpha_1 y) \quad \text{or} \quad G(y) = -\frac{1}{\alpha_2}\exp(-\frac{\alpha_2 y^2}{2}) \tag{3.33}$$

where $1 \leq \alpha_1 \leq 2$, and $\alpha_2 \approx 1$ are constants.

3.6.3 Mutual Information

The KL divergence between the actual joint density $P(\mathbf{y})$ of a random vector $\mathbf{y}$ and the factorised density $\hat{P}(\mathbf{y}) = \prod_{i=1}^{N} P_i(y_i)$, where $P_i(y_i)$ is the marginal density of y_i, is used to define the *mutual information* I between $y_1, y_2, \ldots, y_N$ as follows [30]:

$$I(\mathbf{y}) \equiv D_{KL}(P(\mathbf{y})\|\hat{P}(\mathbf{y})) = -\int P(\mathbf{y}) \log \frac{\prod_{i=1}^{N} P_i(y_i)}{P(\mathbf{y})} d\mathbf{y} \tag{3.34}$$

Mutual information is by definition a degree of dependence between random variables. The smaller the mutual information, the more independent the random variables $y_1, y_2, \ldots, y_N$ are. It is always non-negative, and zero if and only if $y_1, y_2, \ldots, y_N$ are statistically independent (in other words, when their joint density function is factorised to their marginal densities).

Using definitions it is easy to show that [92]

$$I(\mathbf{y}) = J(\mathbf{y}) - \sum_{i=1}^{N} J(y_i) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{y}) \tag{3.35}$$

When $\mathbf{y} = W\mathbf{x}$ (where W non-singular matrix), the density P_y of $\mathbf{y}$ is given by [125]

$$P_y(\mathbf{y}) = \frac{P_x(W^{-1}\mathbf{y})}{|\mathbf{J}(\mathbf{x})|} \tag{3.36}$$

where $|\mathbf{J}(\mathbf{x})|$ is the absolute value of the Jacobian determinant of the transformation from $\mathbf{x}$ to $\mathbf{y}$ defined by

$$\mathbf{J}(\mathbf{x}) = \det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_N}{\partial x_1} & \cdots & \frac{\partial y_N}{\partial x_N} \end{bmatrix} = \det W \tag{3.37}$$

Therefore it is easy to prove that [92]

$$H(\mathbf{y}) = H(\mathbf{x}) + \log|\det W| \tag{3.38}$$

Thus, equation 3.35 is now written as

$$I(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{x}) - \log|\det W| \tag{3.39}$$

If, for the sake of simplicity, y_i's are considered to be uncorrelated and of unit variance, then

$$E\{\mathbf{y}\mathbf{y}^T\} = E\{W\mathbf{x}\mathbf{x}^T W^T\} = WE\{\mathbf{x}\mathbf{x}^T\}W^T = \mathbf{I}$$

Hence,

$$\det E\{\mathbf{y}\mathbf{y}^T\} = \det(WE\{\mathbf{x}\mathbf{x}^T\}W^T) = \det W \det E\{\mathbf{x}\mathbf{x}^T\} \det W^T = 1$$

Therefore, $\det W$ is a constant, and equation 3.39 is rewritten as

$$I(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) + C_1 = C_2 - \sum_{i=1}^{N} J(y_i) \tag{3.40}$$

where C_1, C_2 are constants. Equation (3.40) shows that mutual information and negentropy are closely associated. The unknown original sources can be estimated either by minimising their mutual information or equivalently by maximising the sum of their negentropies.

3.6.4 Likelihood

The recorded signals $\mathbf{x}$ can be modelled as the observed values of M random variables which are generated from $\mathbf{s}$ via the linear transformation of equation 3.4. Their joint probability density function $P(\mathbf{x})$ is dependent upon the $M \times N$ elements a_{ji} of the mixing matrix A, i.e. $P(\mathbf{x}) \equiv P(\mathbf{x}; A)$.

Then, for a fixed observation of $\mathbf{x}$, the *likelihood* L is a function of A given by [39, 141]

$$L(A) \equiv L(A; \mathbf{x}) = P(\mathbf{x}; A) \tag{3.41}$$

The likelihood $L(A; \mathbf{x})$ of a given set of data $\mathbf{x}$ is defined as the probability of obtaining that particular data set $\mathbf{x}$ for a mixing matrix A. In other words, whereas with probability, $\mathbf{x}$ are the variables and A is the constant, with likelihood A is the variable for constant $\mathbf{x}$. It is often more convenient to work with the log of the likelihood function which is called the *log-likelihood function* $\ell(A)$.

Given an $L(A)$, a *maximum likelihood estimate* (MLE) of A is an $\hat{A}$ which attains the maximum value of $L(A)$:

$$L(\hat{A}) = \arg\max_{A} L(A) \tag{3.42}$$

In practice, the MLE is obtained by maximising the log-likelihood. If $\ell(A)$ can be differentiated analytically, then the MLE satisfies the following equations, known as the *likelihood equations*:

$$\frac{\partial \ell}{\partial a_{ji}} = 0 \qquad \text{for all } i = 1, 2, \ldots, N \text{ and } j = 1, 2, \ldots, M \tag{3.43}$$

However, in many cases it is impossible to obtain a closed-form expression for the MLE. Then we have to apply numerical methods either to solve the likelihood equations, or to maximise directly the log-likelihood function.

In the ICA data model of equation 3.4, the log-likelihood function can be given in an analytical form as [129]

$$\ell(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N) = \sum_{k=1}^{K} \sum_{i=1}^{N} \log P_i(\mathbf{w}_i^T \mathbf{x}_k) + K \log |\det W| \tag{3.44}$$

where $\mathbf{w}_i^T$ is the i^{th} row of $W = A^{-1}$, P_i is the density function of the original source s_i (assumed to be known), and $\mathbf{x}_k$ is the value of the original source vector $\mathbf{x}$ for the k^{th} time slice ($k = 1, 2, \ldots, K$).

If we consider the expectation of the log-likelihood, equation 3.44 can be written as

$$\frac{1}{K} E\{\ell(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)\} = \sum_{i=1}^{N} E\{\log P_i(\mathbf{w}_i^T \mathbf{x})\} + \log |\det W| \tag{3.45}$$

If P_i is the actual density function of $\mathbf{w}_i^T \mathbf{x}$, then

$$\sum_{i=1}^{N} E\{\log P_i(\mathbf{w}_i^T \mathbf{x})\} = -\sum_{i=1}^{N} H(\mathbf{w}_i^T \mathbf{x}) \tag{3.46}$$

and equation 3.45 is rewritten as

$$\frac{1}{K} E\{\ell(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)\} = -\sum_{i=1}^{N} H(\mathbf{w}_i^T \mathbf{x}) + \log |\det W| \tag{3.47}$$

Comparing equations 3.39 and 3.47 (taking into account that $\mathbf{y} = W\mathbf{x}$ and $y_i = \mathbf{w}_i^T \mathbf{x}$), we see that mutual information and log-likelihood differ only by a constant and the sign. Therefore, maximising the log-likelihood in equation 3.44 is equivalent to minimising the mutual information of the y_i's.

A small practical limitation of maximum likelihood estimation is that the densities P_i of the original sources s_i are considered to be known. In practice, it is enough to estimate whether they are sub-Gaussian or super-Gaussian [93].

Quantity	Definition
Kurtosis	$\gamma_2(y) = \frac{\mu_4(y)}{\mu_2^2(y)} - 3$
Entropy	$H(\mathbf{y}) = -\int P(\mathbf{y}) \log P(\mathbf{y}) dy$
Negentropy	$J(\mathbf{y}) = H(\mathbf{y}_G) - H(\mathbf{y})$ $J(y) \approx \frac{1}{12}\kappa_3^2(y) - \frac{1}{48}\kappa_4^2(y)$ $J(y) \approx \sum_{j=1}^{P} c_j \left[E\{G_j(y)\} - E\{G_j(y_{G\nu})\}\right]^2$
Mutual Information	$I(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{y}) = J(\mathbf{y}) - \sum_{i=1}^{N} J(y_i)$

Table 3.1: Measures of non-Gaussianity. y: random variable, $\mathbf{y}$: random vector, P: density function, μ_i: i^{th} central moment, κ_i: i^{th} order cumulant, c_j: positive constant, G_j: non-quadratic function, $y_{G\nu}$: standardised Gaussian variable, $\mathbf{y}_G$: Gaussian vector.

3.7 Applications of ICA

ICA has been applied in numerous signal processing problems, such as in speech enhancement and speech recognition systems [127, 159], feature extraction from image data [12, 49, 60], data mining [45, 109], telecommunications [31, 136, 148], and remote sensing [25, 149, 162].

The BSS problem is frequently met in biomedical signal processing when biological signals originating from different sources are mixed together during their recording. Therefore, decomposing the observed data into the original sources provides better understanding of the biological processes taking place in the human body, allows the removal of artefact signals, and overall helps the clinicians to diagnose and treat disorders in a more efficient way.

ICA has been applied to EEG and MEG data either for identification and removal of ocular, cardiac, and myographic artefacts [78, 79, 102, 152, 153] or for direct study of brain functioning [86, 154, 155, 156]. More details about the applicability and the effectiveness of ICA in EEG/MEG problems are provided in chapter 5. An interesting application of ICA with clinical importance in prenatal screening is the extraction of the fetal electrocardiogram (fECG) from multi-lead potential recordings on mother's skin [89, 161]. ICA has been applied in electrogastrograms (EGG) to separate gastric slow waves from respiratory and motion artefacts [94, 157]. Other biomedical, non-bioelectrical applications of ICA include

fMRI [107, 108, 143] and PET [91, 126]. ICA has been proved to be useful in studying long DNA array data [95, 104]. ICA can be also used in animal data to remove cardiac interference in EEG recordings [147], or *in vivo* optical recordings of neural dynamics [100].

In fact, ICA can be employed wherever hidden factors are present. The main advantage of ICA, compared to other techniques used for data exploration, is that ICA is a non-parametric, non-application-specific approach. Instead of constructing parametric models which may be imprecise or non-robust to different sets of data, we let the data separate the underlying independent sources.

ICA has found direct application in the field of astrophysics for separation of astrophysical components superimposed in maps of the sky [101], and for artefact detection and removal, such as cosmic rays and atmospheric events, recorded in astrophysical images [44]. In seismology, ICA has been used to separate seismic signals produced by volcanic eruptions [1]. In environmental physics, ICA is employed to study global temperature time series and examine the effect of different sources, such as volcano eruptions, El Niño variations, and human contributions, on the global climate [2, 42].

An exciting application of ICA is in the world of finance. ICA has been employed to uncover the hidden patterns in financial time series such as the behaviour of currency exchange rates [112], and to explore the underlying structure of the stock market [8]. A successful example of ICA decomposition in financial data is presented in [84]. The common independent factors, such as sudden seasonal changes caused by holidays, which affect the cash flow at several stores belonging to the same retail chain are identified by ICA. Their impact differs slightly from store to store depending on the skills of the individual manager. Thus, ICA can provide invaluable help in assessing each store's management.

Chapter 4

Practical Independent Component Analysis

There are numerous practical ICA algorithms which exploit the properties of the measures of statistical independence provided in section 3.6 in order to solve the BSS problem. Most of them can be considered variations of the approaches presented in section 4.2. Their popularity is also based on the sophisticated optimisation techniques and approximations which they employ in order to solve an intractable multidimensional problem in an accurate and efficient way. Infomax is a neural network method which is equivalent to maximum likelihood estimation. JADE employs cross-cumulants of signals as a measure of independence. Finally, special emphasis is put on FastICA due to its algorithmic simplicity, speed of convergence, and appealing properties in solving real world problems. In fact, our proposed algorithm, presented in chapter 5, employs the approximations of negentropy introduced in FastICA to incorporate prior knowledge about one or more original sources. Finally, the BSS problem in noisy environments is addressed in section 4.3. First, some necessary preprocessing of the data is described in section 4.1, which also shows the poverty of Principal Component Analysis (PCA) over ICA. Application of ICA in simulated data is performed in section 4.4.

4.1 Data Preprocessing

Most practical ICA algorithms, including JADE and FastICA, are significantly simplified when the input data are preprocessed. The typical preprocessing procedure consists of the

following two stages.

4.1.1 Centering

The first step is to make the observed data $\mathbf{x}$ a zero-mean vector by subtracting its mean vector $\bar{\mathbf{x}} = E\{\mathbf{x}\}$. This implies that $\mathbf{s} = A^{-1}\mathbf{x}$ is also of zero-mean. After estimating the mixing matrix A, we add the mean vector $\bar{\mathbf{s}} = A^{-1}\bar{\mathbf{x}}$ of $\mathbf{s}$ to the centered estimates of $\mathbf{s}$. From now on, $\mathbf{x}$ will denote the centered vector of the observed data.

4.1.2 Whitening

The second preprocessing step is important. Its purpose it to make the observed data $\mathbf{x}$ uncorrelated and of unit variance. This transformation is called *whitening* or *sphering*, and is always possible. If by $\mathbf{z}$ we denote the *whitened* (*sphered*) data, then

$$E\{\mathbf{z}\mathbf{z}^T\} = I \tag{4.1}$$

One method for whitening employs the singular value decomposition (SVD) [46] of the $(M \times M)$ covariance matrix $C = E\{\mathbf{x}\mathbf{x}^T\}$ of $\mathbf{x}$:

$$E\{\mathbf{x}\mathbf{x}^T\} = U\Lambda V^T \tag{4.2}$$

where U is the orthogonal $(M \times M)$ matrix whose columns are the eigenvectors of CC^T, Λ is the diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_M$ of CC^T, i.e. $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$, and V is the $(M \times M)$ orthogonal matrix whose columns represent the eigenvectors of C^TC. The elements $\lambda_i^{1/2}$ are also called *singular values* of C. However, since C is a symmetric matrix, $U = V$ and equation 4.2 is rewritten as

$$E\{\mathbf{x}\mathbf{x}^T\} = U\Lambda U^T \tag{4.3}$$

The whitened data $\mathbf{z}$ are given by

$$\mathbf{z} = \Lambda^{-1/2}U^T\mathbf{x} \tag{4.4}$$

where $\Lambda^{-1/2}$ is trivially given by $\Lambda^{-1/2} = \mathrm{diag}(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_M^{-1/2})$.

The $(M \times M)$ matrix $Z = \Lambda^{-1/2}U^T$ is often called *whitening matrix*. Using equations 4.3 and 4.4, it is easy to show that $E\{\mathbf{z}\mathbf{z}^T\} = I$. In addition,

$$\mathbf{z} = \Lambda^{-1/2}U^TA\mathbf{s} = \tilde{A}\mathbf{s} \tag{4.5}$$

where $\tilde{A} = \Lambda^{-1/2}U^T A$ is the new mixing matrix. Since the original sources are considered statistically independent and of unit variance, the covariance matrix $E\{\mathbf{s}\mathbf{s}^T\}$ of $\mathbf{s}$ equals the identity matrix, i.e. $E\{\mathbf{s}\mathbf{s}^T\} = I$. Hence,

$$E\{\mathbf{z}\mathbf{z}^T\} = E\{\tilde{A}\mathbf{s}\mathbf{s}^T\tilde{A}^T\} = \tilde{A}E\{\mathbf{s}\mathbf{s}^T\}\tilde{A}^T = \tilde{A}\tilde{A}^T = I \tag{4.6}$$

If we further assume for the sake of simplicity that the number of mixed signals M equals to the number of sources N, then from equation 4.6 we see that the square $(N \times N)$ matrix $\tilde{A}$ is orthogonal. Therefore, $\tilde{A}$ is described by $\frac{N(N-1)}{2}$ degrees of freedom, whereas the mixing matrix A consists of N^2 parameters. In consequence, whitening reduces nearly by half the number of the parameters which have to be estimated in the BSS problem.

In fact, whitening using SVD is nothing more than a mere application of the well known *Principal Component Analysis* (PCA) to the data [37, 75]. PCA by itself is not sufficient to solve the BSS problem. It simplifies the problem by making the data uncorrelated, yet they are still dependent. Recall from p.28 that uncorrelatedness is weaker than statistical independence. Consequently, ICA can be considered as an extension of PCA.

After data whitening is performed, the modified task of ICA is to estimate the inverse of the new mixing matrix $\tilde{A}$, say $\tilde{W}$, which is used in the linear data mixing model $\mathbf{z} = \tilde{A}\mathbf{s}$. A graphical representation of the BSS procedure including the crucial step of whitening is provided in figure 4.1.

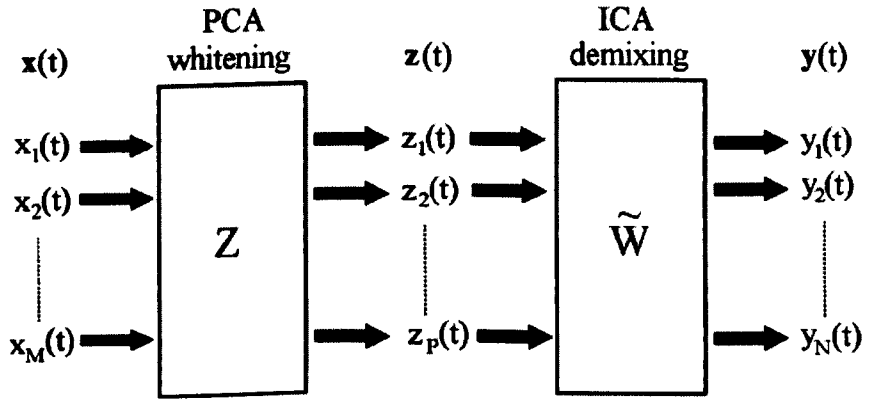
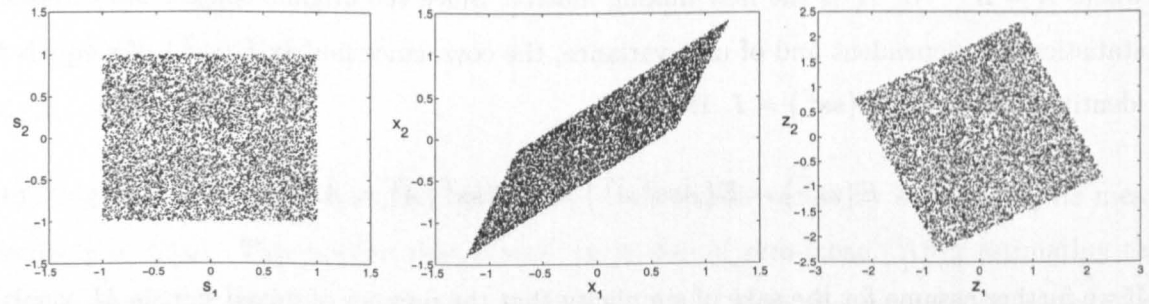


Figure 4.1: Schematic representation of PCA (whitening) and ICA processes. $\mathbf{x}(t)$: observed mixtures, $\mathbf{z}(t)$: whitened signals, and $\mathbf{y}(t)$: estimated source signals. Z: whitening matrix, and $\tilde{W}$: unmixing matrix (estimation of $\tilde{A}^{-1}$).



(a) original sources (b) mixed signals (c) whitened signals

Figure 4.2: Scatter-plots of the original sources (s_1,s_2), the mixed signals (x_1,x_2), and the whitened signals (z_1,z_2) (see also p.28).

For example, figure 4.2 provides a comparison of the scatter-plots of the two uniformly distributed sources given in p.28. The mixed signals, x_1 and x_2, are sphered. From the scatter-plot of the whitened data it is clear that the whitened signals, z_1 and z_2, are still dependent. However, we can see that the whitened data are now a rotated version of the original sources. Therefore, in a geometrical sense, the task of ICA is to estimate the angles producing that rotation.

PCA can be viewed as a rotation of the data axes to new orthogonal directions. The first principal component is the projection of the data on the direction in which the variance of the projection is maximised. The second principal component is the linear combination of the data that explains the maximum amount of variation not explained by the first component. Thus, the i^{th} principal component is the projection on the direction that contains the greatest amount of variation not described by the first $(i-1)$ principal components.

There can be as many principal components as the number of variables. However, PCA is frequently used to reduce the dimensionality of the problem while retaining as much information as possible. In practice, the eigenvalues of CC^T are sorted in a descending order, say $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$, and truncated by discarding the $(M-P)$ eigenvalues which are zero or too small. Therefore, the whitened data $\mathbf{z}$ is now a P-dimensional column vector given by the following linear transformation:

$$\mathbf{z} = \Lambda_P^{-1/2} U_P^T \mathbf{x} \tag{4.7}$$

where $\Lambda_P^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_P^{-1/2})$, and U_P is the $(M \times P)$ matrix whose columns

are the eigenvectors associated with the P most significant eigenvalues of CC^T. The new $(P \times M)$ whitening matrix is $Z = \Lambda_P^{-1/2} U_P^T$.

Dimensionality reduction also results in noise suppression. The subspace spanned by the first P principal components essentially contains most of the energy of the original sources $\mathbf{x}$, whereas the data contained in the remaining principal components can be considered mostly noise. Moreover, it prevents *overlearning* (also known as *overfitting*) which is observed in ICA when it is performed in high dimensions with an insufficient sample size [71].

To sum up, whitening of the recorded data using PCA is an important preprocessing step for any practical ICA algorithm due to the following reasons:

- makes the data uncorrelated, and thus takes a right step towards independence;
- compresses the dataset with a minimum loss of information;
- suppresses noise;
- prevents overlearning.

An alternative method for whitening uses the eigenvalue decomposition (EVD) [46] of the covariance matrix $C = E\{\mathbf{x}\mathbf{x}^T\}$:

$$E\{\mathbf{x}\mathbf{x}^T\} = EDE^T \tag{4.8}$$

where E is the orthogonal matrix whose columns are the eigenvectors of C, and D is the diagonal matrix of its eigenvalues $d_1, d_2, \ldots, d_M$.

The whitened data $\mathbf{z}$ are now given by

$$\mathbf{z} = ED^{-1/2}E^T\mathbf{x} = ED^{-1/2}E^TA\mathbf{s} = \tilde{A}\mathbf{s} \tag{4.9}$$

where $D^{-1/2} = \text{diag}(d_1^{-1/2}, d_2^{-1/2}, \ldots, d_M^{-1/2})$, $\tilde{A} = ED^{-1/2}E^TA$ is the new mixing matrix, and $Z = ED^{-1/2}E^T$ is the whitening matrix. Using equations 4.8 and 4.9, it is easy to show that $E\{\mathbf{z}\mathbf{z}^T\} = I$ and $\tilde{A}$ is orthogonal.

EVD is closely related to SVD. Once again, we can reduce the dimensionality of the problem by disregarding the eigenvalues d_i which are zero or too small. However, SVD is preferred in badly conditioned systems because it tends to be more stable [51]. Note that a problem is said to be ill-conditioned when the *condition number*, defined as the ratio of the largest to smallest eigenvalue, is too large.

4.2 Practical ICA Algorithms

4.2.1 Infomax

The Infomax (Information Maximisation) algorithm was presented in [15]. The main idea is to maximise the output joint entropy $H(y_1, y_2, \ldots, y_N)$ of a neural network with non-linear outputs $y_i = g_i(\mathbf{w}_i^T \mathbf{x})$, where $\mathbf{x}$ is the input vector, g_i is a non-linear scalar function, and $\mathbf{w}_i$ are the weight vectors of the neurons. It has been proven that the Infomax algorithm is equivalent to maximum likelihood estimation when the non-linearities g_i are chosen as the cumulative distribution functions corresponding to the densities P_i of the original sources s_i, i.e. $g_i'(.) = f_i(.)$ [20].

The learning rules for a single layer neural network are given in [15]:

$$\Delta W \propto [W^T]^{-1} + (\mathbf{1} - 2\mathbf{y})\mathbf{x}^T \tag{4.10}$$

$$\Delta \mathbf{w}_0 \propto \mathbf{1} - 2\mathbf{y} \tag{4.11}$$

where W is the weight matrix, $\mathbf{w}_0$ is a bias vector, $\mathbf{1}$ is a vector of ones, and $\mathbf{y} = g(W\mathbf{x} + \mathbf{w}_0)$ is the output vector, with g being the logistic function:

$$g(u) = (1 + e^{-u})^{-1} \tag{4.12}$$

The speed of convergence of the algorithm in equation 4.10 is significantly improved if the computationally intensive matrix inversion can be avoided. This is feasible by using the natural gradient approach [3] which results in a simplified learning rule [4]:

$$\Delta W \propto [I - \mathbf{f}(\mathbf{y})\mathbf{y}^T]W \tag{4.13}$$

where $\mathbf{f}(\mathbf{y}) = [f(y_1)\ f(y_2)\ \ldots\ f(y_N)]^T$, with the activation function $f(y_i)$ defined by

$$f(y_i) = \frac{3}{4}y_i^{11} + \frac{25}{4}y_i^9 - \frac{14}{3}y_i^7 - \frac{47}{4}y_i^5 + \frac{29}{4}y_i^3 \tag{4.14}$$

The main practical drawback of Infomax is the number of parameters which have to be tuned. Successful separation of the original sources depends highly on the parameter values of the algorithm. In consequence, Infomax is not a good prospective candidate for real world BSS problems.

4.2.2 JADE

The JADE (Joint Approximate Diagonalisation of Eigenmatrices) algorithm was originally introduced in [23]. It is the most popular representative of a greater class of ICA algorithms, which are often called *Jacobi algorithms* because they maximise measures of non-Gaussianity using a technique known as the *Jacobi diagonalisation* [24]. Thus, JADE does not suffer from problems of convergence which are frequently met in gradient optimisation methods. This particular class of algorithms is described in detail in [22].

JADE is a statistic-based method. Instead of accessing the data in each iteration of the algorithm, JADE estimates the demixing matrix by operating on a set of statistics generated once and for all from the dataset at the beginning. These statistics include second and fourth order cross-cumulants [21]. JADE assumes that the input data are whitened.

The *cumulant matrix* $Q^{\mathbf{x}}(C) = [q_{ij}]$ of an $(N \times N)$ matrix $C = [c_{kl}]$ is defined element-wise in [23] as

$$q_{ij} \equiv \sum_{k,l=1}^{N} \kappa_4(x_i, x_j, x_k, x_l) c_{kl}, \quad 1 \leq i, j \leq N \tag{4.15}$$

where $\mathbf{x}$ is a N-dimensional random vector, and $\kappa_4(x_i, x_j, x_k, x_l)$ is the fourth order cross-cumulant.

In the ICA data model, $\mathbf{x} = A\mathbf{s}$, a cumulant matrix $Q^{\mathbf{x}}(C)$ can be easily expressed using equation 3.23 in p.32 as in [22]

$$Q^{\mathbf{x}}(C) = A\Delta(C)A^T \tag{4.16}$$

where $\Delta(C) = \mathrm{diag}\Big(\kappa_4(s_1)\mathbf{a_1}^T C \mathbf{a_1}, \ldots, \kappa_4(s_N)\mathbf{a_N}^T C \mathbf{a_N}\Big)$ is a diagonal matrix, and $\mathbf{a_i}$ is the i^{th} column of A, i.e. $\mathbf{a_i} = [a_{1i} \;\; a_{2i} \;\; \cdots \;\; a_{Ni}]^T$. Note that the mixing matrix A is assumed to be $(N \times N)$ for the sake of simplicity. Recall also that the fourth order cumulant κ_4 of a random variable s_i of unit variance equals to its kurtosis (see equation 3.20 in p.31). Hence, kurtosis appears only in the diagonal matrix $\Delta(C)$.

For N variables there are N^4 cross-cumulants. However, a cumulant matrix has only N^2 elements and therefore uses only a small part of the fourth order information. The problem is resolved by processing jointly a maximal set of cumulant matrices C [22, 23], i.e. an orthonormal basis of N^2 matrices for the linear space of $(N \times N)$ matrices. The aim of JADE is to make the cumulant matrices as diagonal as possible. This can be interpreted as making the data as independent as possible.

The strong point of JADE is that there are no parameters to tune. On the other hand, JADE is a memory demanding algorithm. The number of fourth order cross-cumulants grows as $O(N^4)$, where N is the number of original sources.

4.2.3 FastICA

FastICA was developed at the Helsinki University of Technology [58]. The original FastICA algorithm was presented in [70]. In that paper the authors introduced a simple, yet highly efficient, fixed-point iterative algorithm for finding the local extrema of kurtosis. FastICA was further progressed to include kurtosis within the general framework of negentropy [66].

As seen in section 3.6.2, negentropy is a degree of non-Gaussianity. FastICA exploits this statement by maximising the negentropy of a linear combination of the observed signals. However, instead of using the noise sensitive approximations of negentropy based on higher order cumulants as in equation 3.29 in p.34, FastICA aims at maximising a new set of negentropy approximations introduced by the same authors in [65] (see equation 3.30 in p.34). In the simplest case, these approximations have the following form:

$$J(y) \propto \Big[E\{G(y)\} - E\{G(y_{G\nu})\}\Big]^2 \tag{4.17}$$

where G is a non-quadratic function of a random variable y (assumed to be of zero mean and unit variance, i.e. $\mu(y) = 0$ and $\sigma^2(y) = 1$), and $y_{G\nu}$ is a standardised Gaussian variable (i.e. $\mu(y_{G\nu}) = 0$ and $\sigma^2(y_{G\nu}) = 1$). Note that even when these approximations of negentropy are not very accurate, equation 4.17 can still be used as a measure of non-Gaussianity since it takes non-negative values, and becomes zero if and only if y is Gaussian.

In consequence, in order to find a single independent component, denoted by $y_i = \mathbf{w}_i^T\mathbf{x}$ as in p.30, we should estimate the M-dimensional weight vector $\mathbf{w}_i$ which maximises function J_G given by

$$J_G(\mathbf{w}_i) = \Big[E\{G(\mathbf{w}_i^T\mathbf{x})\} - E\{G(y_{G\nu})\}\Big]^2 \tag{4.18}$$
$$\text{under the constraint of } E\{(\mathbf{w}_i^T\mathbf{x})^2\} = 1$$

so that the ambiguity of the scale is removed (see p.27). In practice, the expectation $E\{G(\mathbf{w}_i^T\mathbf{x})\}$ is computed by using a large sample of $\mathbf{x}_k$ vectors, where $\mathbf{x}_k$ is the observed vector $\mathbf{x}$ for the k^{th} time slice ($k = 1, 2, \ldots, K$).

The contrast function $J_G(\mathbf{w}_i)$ is a one-unit function. This means that the independent

components $\mathbf{w}_i^T\mathbf{x}$ can be estimated sequentially, one-by-one, by maximising $J_G(\mathbf{w}_i)$ for as many weight vectors $\mathbf{w}_i$ as the number of original sources we wish to separate. This is an attractive, unique feature of FastICA for application in real world problems without prior knowledge of the number of original sources, or when we wish to extract and study a particular independent component out of the whole set.

In fact, this particular property reveals the strong connection of ICA with *exploratory projection pursuit* (EPP). EPP is a statistical technique for identifying interesting structure in high dimensional datasets [43, 77]. According to [29, 62], the most interesting directions show the least Gaussian distribution. Therefore, ICA can be considered as a variation of EPP, with the assumption of statistical independence between the components.

The strong point of FastICA is the ability to extract a limited number of independent components. However, it is easy to extend the one-unit contrast function of equation 4.18 in order to compute the whole unmixing matrix W. Recall from equation 3.40 in p.36 which was derived under the assumption that the independent components are of unit variance, the mutual information is minimised when the sum of the negentropies of the components is maximised. Thus, in order to find the whole set of independent components, we should estimate the $(N \times M)$ unmixing matrix W which maximises function $J_G(W)$ given by

$$J_G(W) = \sum_{i=1}^{N} J_G(\mathbf{w}_i) = \sum_{i=1}^{N} \Big[E\{G(\mathbf{w}_i^T\mathbf{x})\} - E\{G(y_{G\nu})\}\Big]^2 \tag{4.19}$$

$$\text{under the constraint of } E\{(\mathbf{w}_l^T\mathbf{x})(\mathbf{w}_m^T\mathbf{x})\} = \delta_{lm}$$

where $\mathbf{w}_i^T$ is the i^{th} row of the matrix W and, at the maximum, $\mathbf{w}_i^T\mathbf{x}$ gives an estimation of the i^{th} independent component.

From equation 4.18 it is easy to realise the simplicity of the contrast function. However, the high efficiency of FastICA is also based on the way this contrast function is maximised. The FastICA algorithm is derived as an approximate Newton's method [66]. The derivation of the optimisation algorithm is given in detail in appendix A. The steps of the algorithm are summarised in table 4.1. The index i of the weight vector $\mathbf{w}_i$ has been dropped for the sake of clarity.

Before applying the actual algorithm, the observed data $\mathbf{x} = [x_1\, x_2 \ldots x_M]^T$ are preprocessed as described in section 4.1. The whitened data $\mathbf{z} = [z_1\, z_2 \ldots z_P]^T$ ($P \leq M$, with $P = M$ when the dimensionality is kept) are now used as input in the FastICA algorithm. Therefore,

the constraint of equation 4.18 can be written as

$$E\{(\mathbf{w}_i^T\mathbf{z})^2\} = E\left\{\sum_{j=1}^{P}(w_{ij}z_j)^2\right\} = w_{i1}^2 + w_{i2}^2 + \ldots + w_{iP}^2 = 1 \Rightarrow \|\mathbf{w}_i\| = 1 \qquad (4.20)$$

The goal of ICA is now slightly modified. The task is to estimate the inverse, say $\tilde{W}$, of the orthogonal new mixing matrix $\tilde{A}$ of equation 4.5. If $\mathbf{w}_i^T$ is one of the rows of $\tilde{W}$, the linear combination $\mathbf{w}_i^T\mathbf{z}$ will yield one of the original sources.

1. Take a random initial vector $\mathbf{w}_0$ of unit norm. Let $k = 1$.
2. Let $\mathbf{w}_k = E\{\mathbf{z}g(\mathbf{w}_{k-1}^T\mathbf{z})\} - E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\}\mathbf{w}_{k-1}$.
3. Normalise $\mathbf{w}_k$ by dividing $\mathbf{w}_k$ by its Euclidean norm $\|\mathbf{w}_k\|$.
4. If $|\mathbf{w}_k^T\mathbf{w}_{k-1}|$ is not close enough to 1, let $k = k + 1$ and go back to step 2. Otherwise output vector $\mathbf{w}_k$.

Table 4.1: FastICA algorithm for estimating one independent component. The final vector $\mathbf{w}_k$ provides an estimation of one of the original sources as $\mathbf{w}_k^T\mathbf{z}$. g is the derivative of the function G defined in p.48, and g' is the derivative of g. The index k denotes the k^{th} iteration of the algorithm. The index i of the weight vector $\mathbf{w}_i$ has been dropped for the sake of clarity.

A large set of independent components can be estimated by running the algorithm as many times as required. However, it is necessary to remove the information contained in the solutions already found, in order to estimate a different independent component each time. Recall that the mixing matrix $\tilde{A}$ is orthogonal. Therefore, an orthogonalising projection can be added inside the loop. Then the algorithm is modified to include the projection operation as it can be seen in table 4.2.

The contrast function J_G depends on function G. According to [66], the choice of G is only important if the performance of the method should be optimised. There have been suggested three different functions for G which are summarised in table 4.3 [63]. Their selection was based on statistical criteria, such as robustness and asymptotic covariance.

In practice, the expectations in equation 4.17 are replaced by averages over a finite sample of K observations. If G grows fast when y grows, then the average of $G(y)$ depends mostly on a few extreme observations. Therefore, in order to reduce the effect of possible outliers, G should grow as slowly as possible.

1. Let $\tilde{W}$ be a $(N \times N)$ matrix of zeros. Let $i = 1$.
2. Take a random initial vector $\mathbf{w}_{i_0}$ of unit norm. Let $k = 1$.
3. Let $\mathbf{w}_{i_k} = E\{\mathbf{z}g(\mathbf{w}_{i_{k-1}}^T\mathbf{z})\} - E\{g'(\mathbf{w}_{i_{k-1}}^T\mathbf{z})\}\mathbf{w}_{i_{k-1}}$.
4. Let $\mathbf{w}_{i_k} = \mathbf{w}_{i_k} - \tilde{W}^T\tilde{W}\mathbf{w}_{i_k}$ (orthogonalising projection).
   Normalise $\mathbf{w}_{i_k}$ by dividing $\mathbf{w}_{i_k}$ by its Euclidean norm $\|\mathbf{w}_{i_k}\|$.
5. If $|\mathbf{w}_{i_k}^T\mathbf{w}_{i_{k-1}}|$ is not close enough to 1, let $k = k + 1$
   and go back to step 3. Otherwise output vector $\mathbf{w}_{i_k}$ and
   put $\mathbf{w}_{i_k}^T$ at the i^{th} row of matrix $\tilde{W}$.
6. If $i < N$, let $i = i + 1$ and go to step 2.

Table 4.2: FastICA algorithm for estimating N independent components.

The asymptotic variance of $\mathbf{w}$ is the limit of the covariance matrix of $\mathbf{w}\sqrt{K}$ as $K \to \infty$. This gives an approximation of the mean-square error of $\mathbf{w}$. According to [63], the trace of the asymptotic variance of $\mathbf{w}$ is minimised when G has the form:

$$G(y) = c_1 \log P_i(y) + c_2 y^2 + c_3 \tag{4.21}$$

where c_1, c_2, and c_3 are arbitrary constants, and P_i is the density function of the independent component s_i. For simplicity,

$$G(y) = \log P_i(y) \tag{4.22}$$

However, there is no need to know the exact distribution of the original sources. It can be approximated using the following exponential power family of density functions:

$$P_i(s_i) = k_1 \exp(k_2 |s_i|^\alpha) \tag{4.23}$$

where k_1, k_2 are normalisation constants such that P_i is of unit variance, and α is a positive constant ($0 < \alpha < 2$: super-Gaussian, $\alpha = 2$: Gaussian, $\alpha > 2$: sub-Gaussian)

From equation 4.22, $G(y) = |y|^\alpha$, where the constants have been dropped for simplicity. However, this choice of G is not differentiable at 0 for $a \leq 1$. Hence, differentiable functions with the same qualitative behaviour should be used. In [63] it is suggested using function G_1 when $a = 1$, and G_2 when $a < 1$ (see table 4.3). Consequently,

- G_1 is a good general-purpose function.

- G_2 is derived for highly super-Gaussian sources, and is useful when robustness against outliers is important because it grows very slowly when y grows.
- G_3 is derived for sub-Gaussian variables. Note that for whitened data, G_3 is roughly equivalent to kurtosis (see equation 3.20 in p.31). Nevertheless, its use is not indicated when outliers are present.

$G(y)$	$g(y)$	$g'(y)$
$G_1(y) = \frac{1}{\alpha_1} \log\cosh(\alpha_1 y)$	$\tanh(\alpha_1 y)$	$\alpha_1(1 - \tanh^2(\alpha_1 y))$
$G_2(y) = -\frac{1}{\alpha_2} \exp(-\alpha_2 y^2/2)$	$y \exp(-\alpha_2 y^2/2)$	$(1 - \alpha_2 y^2) \exp(-\alpha_2 y^2/2)$
$G_3(y) = \frac{1}{4} y^4$	y^3	$3y^2$

Table 4.3: Choices for function G. The first and second derivative of G, which are denoted by g and g' respectively and used in the algorithm in tables 4.1-4.2, are also provided. $1 \leq \alpha_1 \leq 2$ and $\alpha_2 \approx 1$ are constants.

In conclusion, the success of FastICA depends greatly on the efficient optimisation algorithm. The algorithm uses an approximate Newton's method to maximise approximations of negentropy defined by non-quadratic functions G. These approximations include kurtosis as a special case (G_3). The algorithmic performance can be further optimised by selecting the proper non-linearity G. The algorithm converges to the right extrema, independently of the distribution of the original sources, in a quadratic way (or even cubic when kurtosis is used) [66], whereas gradient methods show linear convergence. Moreover, there are no step parameters to tune. However, the most appealing property of FastICA, compared with other ICA approaches, is the ability of estimating the independent components one-by-one. Thus, the algorithm is ideal for use in environments where a single or a limited number of components should be extracted.

4.3 Noisy ICA

In real world problems, the recorded, mixed signals are usually corrupted with additive noise. Therefore, the simple, linear, noise-free ICA mixing model, as described by equation 3.4 in p.25 is incomplete and should be revised. In practice, the j^{th} observed signal x_j at the k^{th}

time slice should be expressed as

$$x_{jk} = \sum_{i=1}^{N} a_{ji} s_{ik} + n_{jk} \tag{4.24}$$

where n_{jk} is the noise affecting the j^{th} sensor at the k^{th} time slice.

Using vector-matrix notation, the data model can be written as

$$\mathbf{x} = A\mathbf{s} + \mathbf{n} \tag{4.25}$$

where the extra term $\mathbf{n} = [n_1\ n_2\ \ldots\ n_M]^{\mathrm{T}}$ is the M-dimensional column vector of additive Gaussian noise which corrupts $\mathbf{x}$, assumed to be of zero mean for the sake of simplicity. This generative model is known as *noisy* ICA model. Note that the noise term $\mathbf{n}$ should not be confused with artefact signals which are linearly mixed with the signals of interest and contaminate the observed data.

The approach used in noise-free ICA can be used for noisy ICA as well, if only we have measures of non-Gaussianity whose values for the unknown original sources can be estimated from noisy observations. However, the important difference between noisy and noise-free ICA is that the presence of noise requires non-linear reconstruction of the independent components. Indeed, if W is an estimation of A^{-1}, then the independent components cannot be extracted simply as $W\mathbf{x}$. In general, the estimation of $\mathbf{s}$ involves complicated algebraic manipulations which can lead to closed-form solutions under specific approximations [64].

The additive noise components n_i are considered to be independent of one another and of the original sources s_i. In consequence, $E\{\mathbf{s}\mathbf{n}^T\} = \mathbf{0}$, and the noise covariance matrix is $\Sigma = E\{\mathbf{n}\mathbf{n}^T\}$, where Σ is a diagonal matrix with positive elements.

Let us further assume that Σ is known. Then, in order to take into account the effect of noise, the whitening preprocessing step should be replaced by an EVD of the symmetric matrix $C - \Sigma$, where $C = E\{\mathbf{x}\mathbf{x}^T\}$ is the covariance matrix of the noisy observations. This operation is called *quasi-whitening*. Hence,

$$C - \Sigma = U_r \Lambda_r U_r^T \tag{4.26}$$

where U_r is the orthogonal $(M \times M)$ matrix whose columns are the eigenvectors of $C - \Sigma$, and $\Lambda_r = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$ is the diagonal matrix containing the eigenvalues of $C - \Sigma$.

The quasi-whitened data $\mathbf{z}$ are given by

$$\mathbf{z} = U_r \Lambda_r^{-1/2} U_r^T \mathbf{x} = U_r \Lambda_r^{-1/2} U_r^T (A\mathbf{s} + \mathbf{n}) = \tilde{A}\mathbf{s} + \tilde{\mathbf{n}} \tag{4.27}$$

where $\Lambda_r^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_M^{-1/2})$, and $\tilde{A} = U_r \Lambda_r^{-1/2} U_r^T A$ is the new mixing matrix. The matrix $Z = U_r \Lambda_r^{-1/2} U_r^T$ is called *quasi-whitening matrix*. Using equations 4.26 and 4.27 it is easy to show that

$$E\{\mathbf{z}\mathbf{z}^T\} = I + \tilde{\Sigma} \tag{4.28}$$

where $\tilde{\Sigma} = E\{\tilde{\mathbf{n}}\tilde{\mathbf{n}}^T\} = Z\Sigma Z$ is the covariance matrix of the transformed noise $\tilde{\mathbf{n}} = Z\mathbf{n}$.

The FastICA algorithm can be modified in order to estimate the ICA model in the presence of Gaussian noise [67]. Step 2 of the algorithm presented in table 4.1 should be replaced by the following:

> 2. Let $\mathbf{w}_k = E\{\mathbf{z}g(\mathbf{w}_{k-1}^T\mathbf{z})\} - (I + \tilde{\Sigma})\mathbf{w}_{k-1}E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\}$.

An identical correction should be made in step 3 of the algorithm in table 4.2 for the estimation of several components.

In fact, this modification for noisy data does not decipher completely the BSS problem. The addition of $\tilde{\Sigma}$ results in the removal of the asymptotic bias which is produced by noise in standard ICA, and allows an accurate estimation of $\tilde{W} = \tilde{A}^{-1}$. However, it does not solve the additional problem of non-linear reconstruction of independent components $\mathbf{s}$. Not much work has been done in this field. The BSS problem has been already very difficult to solve in the noise-free case.

In general, this obstacle can be surpassed by joint maximum likelihood estimation of A and $\mathbf{s}$ [64]. However, the optimisation problem now involves $M \times N + N \times K$ variables, where M, N and K is the number of mixed signals, original sources, and time slices (or sample points) respectively. The complexity can be reduced by applying a series of approximations about the noise and the distribution of the independent components which may lead to loss of generality [64]. Moreover, in [64] the noise covariance matrix Σ is also assumed to be known in advance. An alternative MLE algorithm for noisy ICA based on an expectation-maximisation (EM) method [36] is presented in [116]. Although it has the advantage of estimating Σ as part of the algorithm, it is extremely computationally intensive. Therefore, it can be employed only in datasets of small dimensionality. Nevertheless, several approximations can be derived in low noise environments, or for independent components following certain distribution [16].

A promising idea is the use of *Independent Factor Analysis* (IFA) [7]. Factor Analysis (FA) was actually developed for Gaussian sources [54, 90]. According to the ordinary FA statistical

data model, the j^{th} mixed signal x_j $(j = 1, 2, \ldots, M)$ can be expressed in a compact way as

$$x_j = \sum_{i=1}^{N} \lambda_{ji} f_i + e_j \tag{4.29}$$

where f_i is the i^{th} common factor $(i = 1, 2, \ldots, N \leq M)$, and e_j is the residual component affecting the recorded signal x_j. The coefficient λ_{ji} is known as the *loading* of f_i in x_j.

In practice, equation 4.29 states the hypothesis that the recorded signals x_j are generated from an unobserved systematic part $\sum_{i=1}^{N} \lambda_{ji} f_i$, which is a linear combination of a smaller number of unobserved factor variables, and an unobserved error part e_j.

Using vector-matrix notation, the factor model can be written as:

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{e} \tag{4.30}$$

where $\mathbf{x} = [x_1\ x_2\ \ldots\ x_M]^T$ is the M-dimensional vector of the recorded signals, $\Lambda = [\lambda_{ji}]$ is the constant $(M \times N)$ matrix of factor loadings, $\mathbf{f} = [f_1\ f_2\ \ldots\ f_N]^T$ is the N-dimensional vector of common factors, and $\mathbf{e} = [e_1\ e_2\ \ldots\ e_M]^T$ is the M-dimensional vector of residual components affecting $\mathbf{x}$. Note that equation 4.30 is similar to the noisy ICA data model in equation 4.25.

In the *unrestricted* factor model, the factors f_i are assumed to be uncorrelated with unit variances (i.e. whitened). In this case, $E\{\mathbf{f}\mathbf{f}^T\} = I$. The residual components e_j are considered to be independent of one another and of the factors f_i. In consequence, $E\{\mathbf{f}\mathbf{e}^T\} = \mathbf{0}$, and the covariance matrix of $\mathbf{e}$ is $E\{\mathbf{e}\mathbf{e}^T\} = \Psi$, where Ψ is a diagonal matrix. Then the covariance matrix C of $\mathbf{x}$ is

$$C = E\{\mathbf{x}\mathbf{x}^T\} = E\{(\Lambda\mathbf{f} + \mathbf{e})(\Lambda\mathbf{f} + \mathbf{e})^T\} = \Lambda\Lambda^T + \Psi \tag{4.31}$$

The aim of factor analysis is to estimate both unknown Λ and Ψ when only $\mathbf{x}$ is known. It can be proven that for a given number N of factors f_i (with $N < M$), there is a unique Ψ, with positive diagonal elements, and a unique $(M \times N)$ matrix Λ which satisfy equation 4.31 [90]. A popular method to estimate Λ and Ψ is the maximum likelihood estimation. FA has been also used as a method of reducing the data dimensionality like PCA.

In the standard FA model, both $\mathbf{f}$ and $\mathbf{e}$ are also assumed to follow multivariate normal distributions with zero mean. However, IFA is an extension of ordinary FA in which the factors f_i are considered to be non-Gaussian variables modelled by a mixture of Gaussians [7]. The parameters of the data model, such as the mixing matrix, the noise covariance, and the source densities, are learned using an EM algorithm.

4.4 Experimental Results

Chapter 3 provided the theoretical foundation of ICA. The nature of the BSS problem was explained in detail, and the theoretical quantities involved in ICA were presented thoroughly. So far the current chapter has dealt with issues of applied ICA, such as an efficient practical algorithm and necessary data preprocessing. Now is the right time to examine the application of ICA in practical data and validate its efficiency. The purpose of the following experiments is to study the behaviour of ICA in noise-free and noisy environments with simulated data for which the true mixing parameters are known. A secondary goal is to demonstrate the major inherent ambiguity of ICA in determining the order of independent components as stated in section 3.4. Essentially this can be considered as the motivation towards the introduction in chapter 5 of a novel algorithm which takes into account prior knowledge in order to favour the extraction of a particular independent component of interest. Finally, we will show how ICA can be employed in practice in order to remove a particular original source signal from the observed mixed data. The intrinsic simplicity of this procedure is the appealing feature of ICA in removing artefact signals from contaminated recordings, such as in MEG, compared with other artefact rejection techniques which were presented in section 1.1.2.

4.4.1 Noise-free ICA with simulated data

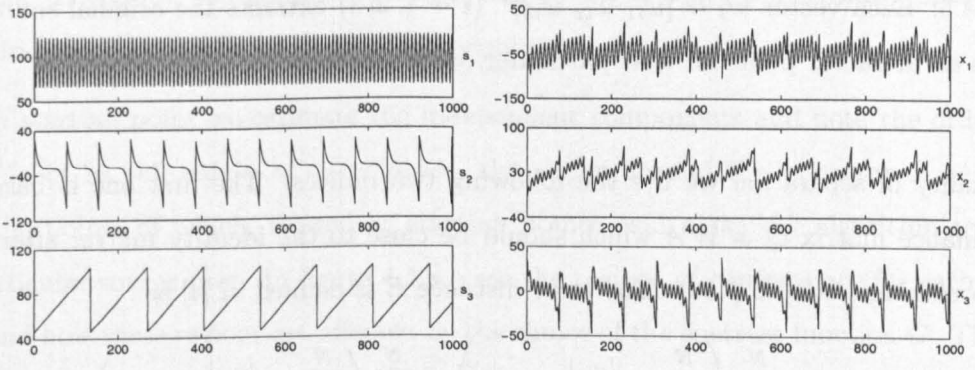
First, we will examine the efficiency of ICA under noise-free conditions. Three known simulated signals, s_1, s_2, and s_3, of 1000 sample points each are generated and linearly mixed with a known (3×3) mixing matrix A (see figure 4.3). The mixed signals, x_1 x_2, and x_3, are centered by subtracting their mean values. Finally, the data are whitened using an EVD as described in section 4.1.2. Some elementary statistical characteristics of the signals, calculated from the samples, are provided in Table 4.4. The sinusoidal signal s_1 and the sawtooth s_3 are sub-Gaussian, whereas the funny-shaped signal s_2 is super-Gaussian. Note that the ideal value of kurtosis of a sinusoidal signal such as s_1 is -1.5. The mixing matrix A and the whitening matrix Z are given below:

$$A = \begin{pmatrix} -0.999 & -0.538 & 0.434 \\ -0.267 & -0.409 & 0.588 \\ 0.282 & 0.591 & -0.137 \end{pmatrix} \quad \text{and} \quad Z = \begin{pmatrix} 0.012 & 0.105 & 0.135 \\ 0.076 & -0.086 & 0.061 \\ -0.028 & -0.015 & 0.014 \end{pmatrix}$$

The whitened data, z_1, z_2, and z_3, are used as input in the FastICA algorithm. Thus, the

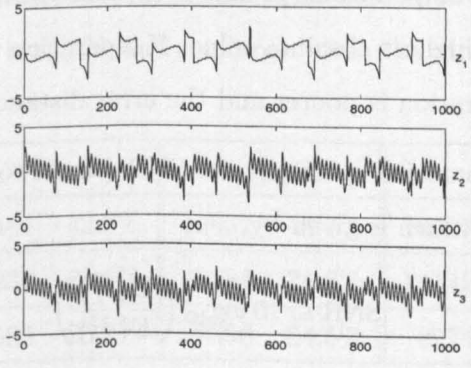
new mixing matrix is now $\tilde{A} = ZA$. Since $\tilde{A}$ is known, we can calculate the weight vector $\mathbf{w}_i^T$ $(i = 1, 2, 3)$, which yields each of the original sources s_i, as the i^{th} row of $\tilde{W} = \tilde{A}^{-1}$. Recall from equation 4.20 in p.50 the constraint $\|\mathbf{w}_i\| = 1$. Therefore, each row of $\tilde{W}$ is divided elementwise by its norm. The normalised weight vectors $\mathbf{w}_i$ are summarised in table 4.5.

Each $\mathbf{w}_i$ defines an attractor point in 3D space. In fact, the task of estimating one of the vectors $\mathbf{w}_i$ is reduced to a 2D optimisation problem due to the constraint $\|\mathbf{w}_i\| = 1$. The weight coefficient w_{i3} can be computed as $w_{i3} = \sqrt{1 - w_{i1}^2 - w_{i2}^2}$, for given values of w_{i1} and w_{i2}. Note that due to the ambiguity of the sign, we decide to use always the positive sign when we take the square root without any loss of generality. This implies also that, for each original source s_i , there are actually two attractors which are defined by the vectors $\mathbf{w}_i$ and $-\mathbf{w}_i$.



(a) original sources s_1, s_2, s_3

(b) mixed signals x_1, x_2, x_3



(c) whitened signals z_1, z_2, z_3

Figure 4.3: Three known signals are linearly mixed and then centered and whitened.

In practice, any ICA algorithm gives an estimation of the separation matrix $\tilde{W}$. To quantify

	original sources			mixed signals			whitened signals		
	s_1	s_2	s_3	x_1	x_2	x_3	z_1	z_2	z_3
mean	94.32	-28.55	74.28	-46.60	30.14	-0.38	0	0	0
stdev σ	19.48	19.29	16.74	23.40	13.71	12.96	1	1	1
kurtosis	-1.50	3.08	-1.21	-0.67	-0.13	1.91	-0.37	0.01	-0.13

Table 4.4: Basic statistical properties of the artificially generated signals.

	w_{i1}	w_{i2}	w_{i3}	original source
$\mathbf{w}_1$	-0.024	-0.694	0.720	sinusoidal s_1
$\mathbf{w}_2$	0.578	0.563	0.592	funny-shaped s_2
$\mathbf{w}_3$	-0.815	0.433	0.385	sawtooth s_3

Table 4.5: Each vector $\mathbf{w}_i = [w_{i1}\ w_{i2}\ w_{i3}]^T$ $(i = 1, 2, 3)$ extracts the original source s_i, and defines an attractor point for that particular original source.

the quality of separation we use the following two indices. The first one is based on the performance matrix $Q = \tilde{W}\tilde{A}$ which should be close to the identity matrix after rescaling and reordering (see also p.30). The error distance E is defined in [4] as

$$E \equiv \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\frac{|q_{ij}|}{\max_k |q_{ik}|} - 1\right) + \sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{|q_{ij}|}{\max_k |q_{kj}|} - 1\right) \tag{4.32}$$

where q_{ij} are elements of the performance matrix Q, and $\max_k |q_{ik}|$ denotes the element of the i^{th} row of Q with the highest absolute value. If separation is perfect, then $E = 0$. As $\tilde{W}$ goes away from $\tilde{A}^{-1}$, separation is poorer and the error distance E increases.

The second measure of separation quality is the SNR (signal-to-noise ratio) of the estimated independent components, which is given by

$$\text{SNR} \equiv 10\log_{10}\left(\frac{E\{s_i^2\}}{E\{d_i^2\}}\right) \tag{4.33}$$

where s_i is the original source, and $d_i = y_i - s_i$ is the undesired error with y_i the estimated independent component which corresponds to s_i. The mean value $E\{.\}$ is calculated from the sample points.

Note that in order to apply equation 4.33, y_i should have the same energy as s_i (or equivalently the same variance). However, the estimated components y_i are of unit variance. Therefore,

the original sources s_i are normalised to unit variance as well. In addition, since ICA cannot determine the order of separation, the extracted independent components should be carefully matched to their respective original sources.

We apply ICA using the deterministic FastICA algorithm to the whitened data $\mathbf{z}$. We examine all three contrast functions G_1 ($\alpha_1 = 1.5$), G_2 ($\alpha_2 = 1$), and G_3 (see p.52). Different values of the parameters α_1 and α_2 do not really affect the following results. Hence, we keep them fixed for all simulations to follow. For the sake of simplicity, from now on we will also refer to function G_3 as kurtosis.

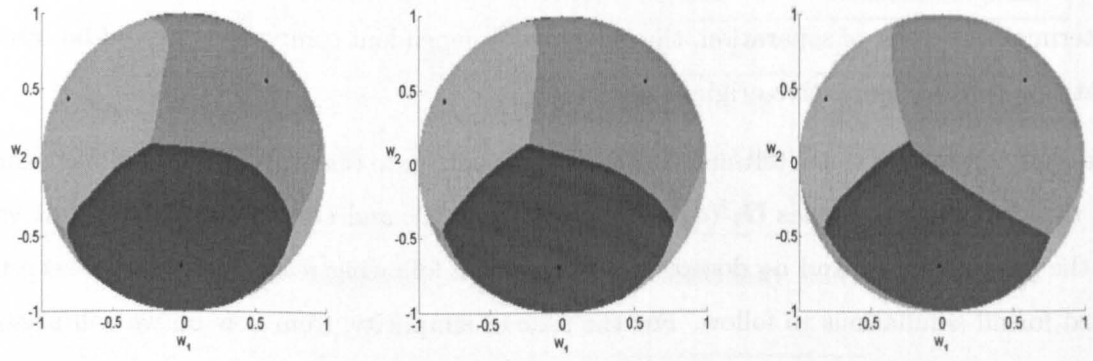
As mentioned above, the estimation of a particular independent component is a two-variable optimisation problem. We scan the 2D space of w_1 and w_2 with a step of 0.01 in both directions. If $w_1^2 + w_2^2 \leq 1$, then the point (w_1,w_2) can be used as a starting point in the algorithm. Therefore, the population of starting points is contained within a circle of unit radius. In total, 31413 starting points are examined.

For each starting point we estimate the independent components and note the order of extraction. Let us call *region of convergence*, the region around the attractor of an original source the points of which, if used as starting points, will make the algorithm to extract that particular source first. In figure 4.4 we see the regions of convergence for each original source and how these regions are affected by the choice of the contrast function G. The graph confirms the ambiguity of the extraction order in standard, unconstrained ICA. The order of separation depends on the starting point. Table 4.6 shows quantitatively the size of each region of convergence for different G. For example, note that if kurtosis (G_3) is used, the region of convergence of s_2 is significantly enlarged over the others, since s_2 has the highest absolute value of kurtosis.

	G_1		G_2		G_3	
s_1	13444	42.8%	12931	41.2%	10997	35.0%
s_2	9937	31.6%	10786	34.3%	14142	45.0%
s_3	8032	25.6%	7696	24.5%	6274	20.0%

Table 4.6: Size of region of convergence. Each column provides the number of starting points which extract a particular component s_i for different choices of G. See also figure 4.4.

In order to assess the performance of the algorithm, we estimate the two quality indicators described above. The results are summarised in tables 4.7 and 4.8. The mean and standard



(a) G_1 (b) G_2 (c) G_3

Figure 4.4: Regions of convergence for FastICA. Depending on the starting point a particular component is extracted first. A starting point in the lower dark area yields the sinusoidal signal s_1, in the upper right mid-dark area the funny-shaped signal s_2, and in the upper left light area the sawtooth signal s_3. The attractor points for the original sources are noted with black dots. The region of convergence depends also on the choice of the contrast function G. Note the small lower right light area which yields the sawtooth signal s_3 with inversed sign.

deviation of error distance E are calculated from the total population of starting points independently of which component is extracted first. However, the SNR performance index is computed only for the first extracted independent component. This is due to the fact that an unsuccessful separation does not remove completely that particular component from the mixed signals, and thus distorts the separation of the remaining components. Of course, this affects the error distance E as well. However, we intend to use E as a quality index of the overall separation procedure.

From table 4.8 we confirm that kurtosis (G_3) yields the worst results when separating the super-Gaussian s_2. On the other hand, kurtosis performs better than G_1 and G_2 for the extraction of the sub-Gaussian s_1. Functions G_1 and G_2 have similar performance. Overall, the source separation is proved to be successful for all starting points. In practice, even for the worst separation case between all choices of G ($G_3, E = 0.240$), the estimated signals coincide perfectly with the original sources (normalised to unit variance due to the ambiguity of the scale, and with the inverse sign if necessary due to the ambiguity of the sign) (see figure 4.5). The quality of separation is comparable for all choices of G (although it is slightly worse when G_3 is used). In fact, as noted in [66], the choice of G is important only if the performance of

the method should be optimised, or in special cases such as in presence of outliers.

	Error Distance E			
	mean	std	min	max
G_1	0.143	0.025	0.087	0.169
G_2	0.144	0.018	0.094	0.166
G_3	0.205	0.033	0.103	0.240

Table 4.7: Error distance for different choices of G. The statistics are calculated from the total population of starting points, independently of which component is extracted first, and thus provide an overall separation quality index.

	SNR of s_1 (dB)	SNR of s_2 (dB)	SNR of s_3 (dB)
G_1	47.3	33.6	42.2
G_2	52.9	33.6	39.7
G_3	62.6	29.5	37.4

Table 4.8: SNR of the first extracted independent component for different choices of G. Note that G_1 and G_2 have similar performance. G_3 (kurtosis) works best for the most sub-Gaussian s_1, and worst for the super-Gaussian s_2 when compared with G_1 and G_2.

In practice, there may be no need for the separation of all sources. Let us assume that we extract only one independent component. Then it is easy to remove this component from the observed mixed signals. In consequence, this BSS algorithm is particularly useful for the identification and removal of specific artefact signals which contaminate the recordings.

If $\mathbf{w}$ is the estimated weight vector of a single component $y = \mathbf{w}^T\mathbf{z}$, the "clean" whitened data $\mathbf{z}_c$ (i.e. the whitened data where the contribution of that particular component has been eliminated) are given by $\mathbf{z}_c = \mathbf{z} - \mathbf{w}\mathbf{w}^T\mathbf{z}$. Then in order to return to the zero-mean recordings, we simply inverse the whitening transformation: $\mathbf{x}_c = Z^{-1}\mathbf{z}_c$. Finally, to complete the reconstruction we should add the mean vector $\bar{\mathbf{x}}$ of $\mathbf{x}$ which we removed when we centered the data: $\mathbf{x}_C = \mathbf{x}_c + \bar{\mathbf{x}}$. To sum up, the observations $\mathbf{x}_C$, cleaned from a particular component $\mathbf{w}^T\mathbf{z}$, are given by

$$\mathbf{x}_C = Z^{-1}(\mathbf{z} - \mathbf{w}\mathbf{w}^T\mathbf{z}) + \bar{\mathbf{x}} \tag{4.34}$$

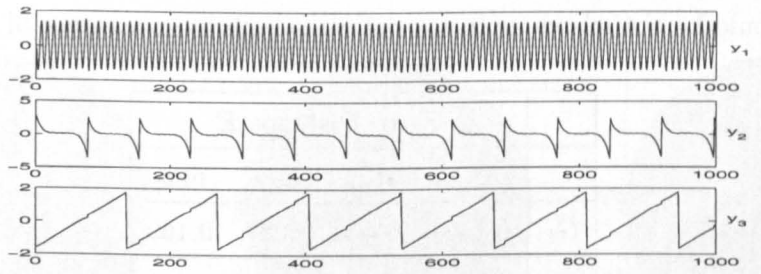


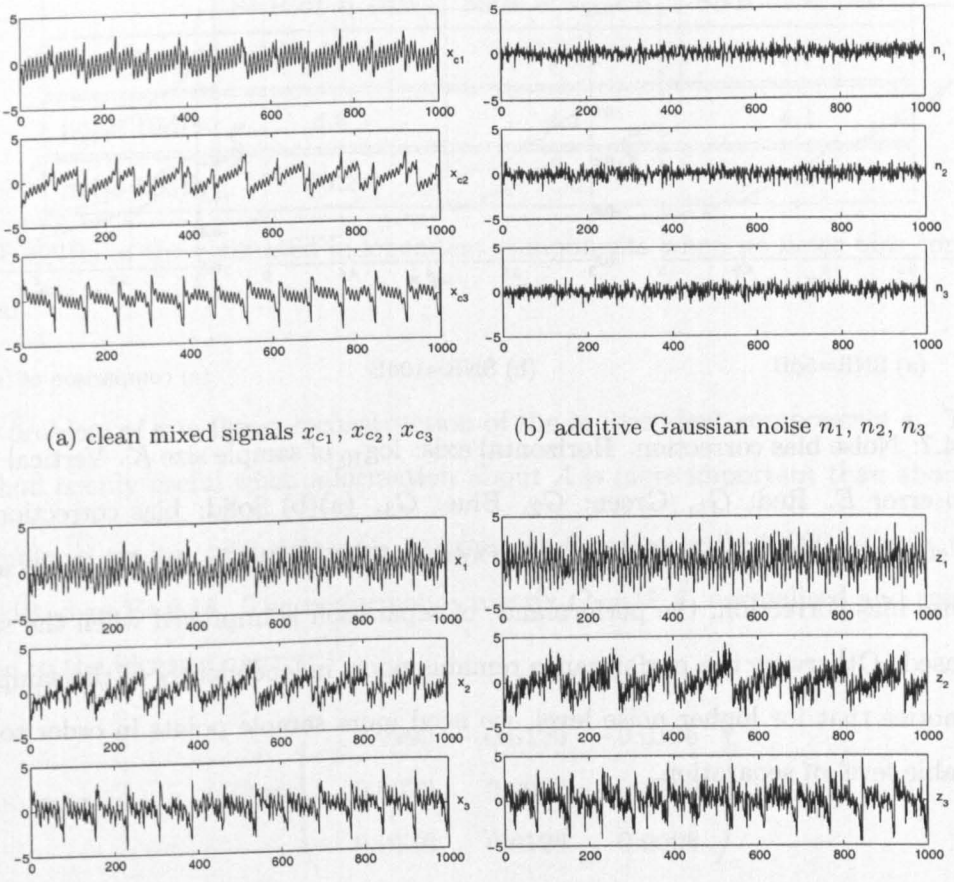
Figure 4.5: Estimated independent components for the worst separation case (G_3, $E = 0.240$). The separation is almost perfect. The original sources are shown in figure 4.3(a).

4.4.2 Noisy ICA with simulated data

The original sources s_i of the previous section, normalised to unit variance, are also used in this section. Now we will examine the behaviour of ICA in noisy conditions. For the sake of simplicity and without any loss of generality, the mixing matrix A is normalised so that the variance of each "clean" mixed signal $x_{c_i} = As_i$ equals to unity. We add Gaussian noise with known covariance matrix $\Sigma = \sigma^2 I$ to the clean mixtures (see figure 4.6). The SNR is calculated from equation 4.33 in p.58, where the numerator is $E\{x_{c_i}^2\} = 1$, and the denominator is $E\{n_i^2\} = \sigma^2$ $(i = 1, 2, 3)$. The noisy observations $\mathbf{x}$ are given by $\mathbf{x} = \mathbf{x_c} + \mathbf{n}$. The noisy data are then quasi-whitened as described in section 4.3. The new mixing matrix is now $\tilde{A} = ZA$, where Z is the quasi-whitening matrix.

We consider a fixed starting point $[w_1, w_2] = [-0.20, -0.60]$, and then we apply ICA with noise bias correction for various sample sizes from $K = 1000$ to $K = 100000$ sample points. To obtain data of different size, we simply replicate the original data of 1000 sample points as many times as needed. Of course, in noise-free ICA such a procedure does not yield different results for different sample sizes. The extra noiseless sample points do not bear any additional statistical information. However, it is helpful in noisy ICA in order to show how the noise bias is reduced when the statistics of the contrast function are estimated from a large sample of noisy observations.

For each sample size we perform 100 trials with different additive noise of the same level. The quality of separation in each trial is expressed by the error distance E of equation 4.32 in p.58. For a given K, the error E is computed as the mean of the error distances of all 100 trials. The procedure is repeated for all three possible choices of function G. Finally, we repeat the simulations without bias correction. We examine two different noise levels: (a)



(a) clean mixed signals x_{c_1}, x_{c_2}, x_{c_3}

(b) additive Gaussian noise n_1, n_2, n_3

(c) noisy signals x_1, x_2, x_3

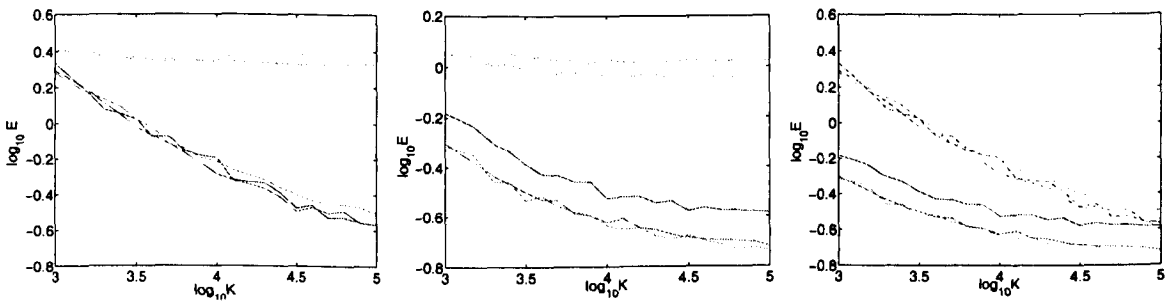
(d) quasi-whitened signals z_1, z_2, z_3

Figure 4.6: Additive Gaussian noise $\mathbf{n}$ corrupts the clean mixed signals $\mathbf{x_c}$ (SNR=5dB). The noisy data $\mathbf{x}$ are quasi-whitened before performing ICA.

SNR=5dB, and (b) SNR=10dB.

The results are depicted in figure 4.7. We notice that when no bias correction is employed, the quality of separation is relatively poor, and gets poorer as the noise level increases. In addition, the sample size does not affect the performance. On the other hand, with noise bias correction, the error E tends asymptotically to zero as more noisy samples are used for the estimation of the statistics of any function G. Note also that the higher the noise level, the more samples we need in order to reach a comparable level of separation.

Note that this method of noise bias correction assumes that we are aware of the presence of noise. Moreover, we have prior knowledge about the noise which is expressed through the known noise covariance matrix Σ. Hence, quasi-whitening is possible. However, in practice



(a) SNR=5dB (b) SNR=10dB (c) comparison of (a) & (b)

Figure 4.7: Noise bias correction. Horizontal axis: $\log_{10}$ of sample size K. Vertical axis: $\log_{10}$ of mean error E. Red: G_1. Green: G_2. Blue: G_3. (a)(b) Solid: bias correction. Dotted: no bias correction. (c) Solid: SNR=10dB. Dotted: SNR=5dB. From (a) and (b) we see that with noise bias correction, the performance of separation is improved when the sample size is increased. Otherwise the performance remains poor, independently of the sample size. In (c) we notice that for higher noise level, we need more sample points in order to achieve a comparable level of separation.

we may not know that we have noise. In this case, the actual noisy data are whitened as described in section 4.4.1. As we mentioned above, simple whitening of noisy data results in poor quality of separation. This behaviour is expected since the ICA model was established for noise-free data. Figure 4.8 shows the estimated independent components if there is no information available about the noise, and simple whitening is performed to the noisy data of figure 4.6 for SNR=5dB, 10dB, and 15dB. The estimated independent components deviate from the actual source signals as the noise level increases (compare also dotted lines in figures 4.7(a) and 4.7(b) depicting error distance E when noise bias correction is not applied). The SNR of the estimated independent components, which was defined by equation 4.33 in p.58, is summarised in table 4.9. Note also that our experiments showed that the regions of convergence demonstrate a similar behaviour as in figure 4.4. However, since the estimated independent components are now distorted, increased attention is required in identifying the order of extraction, especially for high levels of noise.

Nevertheless, the technique of noise bias correction solves only half of the BSS problem. Although it succeeds in estimating the inverse of the mixing matrix A (with the ambiguity of order, scale, and sign), provided a large sample of data is available, the method does not deal

	SNR of s_1 (dB)	SNR of s_2 (dB)	SNR of s_3 (dB)
noise 5dB	3.5	1.1	0.1
noise 10dB	6.6	4.1	4.1
noise 15dB	10.2	7.8	7.3

Table 4.9: SNRs of the estimated independent components when no noise bias correction is performed.

with the problem of non-linear reconstruction of the independent components s_i. Therefore, this method is only useful when information about A is more important than about s_i.

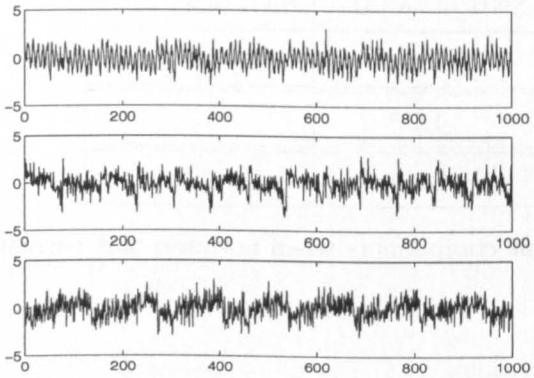
For example, a typical trial using G_1 as contrast function with 10000 sample points for SNR=10dB gives E=0.18. The performance matrix $Q = \tilde{W}\tilde{A}$, normalised and reordered, is very close to the identity matrix:

$$Q = \begin{pmatrix} 0.9999 & -0.0190 & -0.0135 \\ -0.0073 & 0.9997 & -0.0229 \\ 0.0146 & 0.0126 & 0.9996 \end{pmatrix}$$

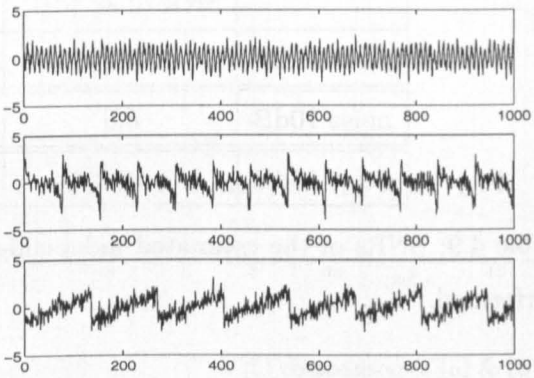
Hence, the separation is almost perfect. The linearly separated signals $\tilde{W}\mathbf{z}$ are shown in figure 4.9(a). The errors in the estimated components are due to noise and linear reconstruction. Since the "clean" mixed signals $\mathbf{x}_c = A\mathbf{s}$ are known in this example, we can reconstruct the independent components without the noise distortion as $\tilde{W}\mathbf{z}_c = \tilde{W}Z\mathbf{x}_c$ (see figure 4.9(b)). Note that now the estimated components are almost identical to the original sources (up to a scale factor). In table 4.10 we summarise the SNRs of the independent components, estimated either as $\tilde{W}\mathbf{z}$ or as $\tilde{W}\mathbf{z}_c$.

	SNR of $\tilde{W}\mathbf{z}$ (dB)	SNR of $\tilde{W}\mathbf{z}_c$ (dB)
s_1	4.1	28.2
s_2	5.1	32.7
s_3	4.3	36.0

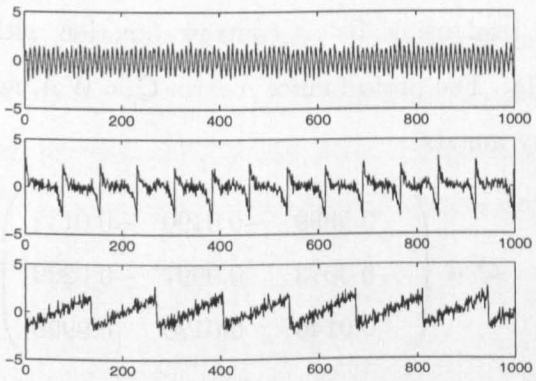
Table 4.10: SNRs of the independent components, estimated either from the noisy observations as $\tilde{W}\mathbf{z}$, or from the known clean mixtures as $\tilde{W}\mathbf{z}_c$. Note that, although noise bias correction estimates A^{-1} sufficiently, it does not actually help in unmixing the original sources by simple linear reconstruction.



(a) SNR=5dB



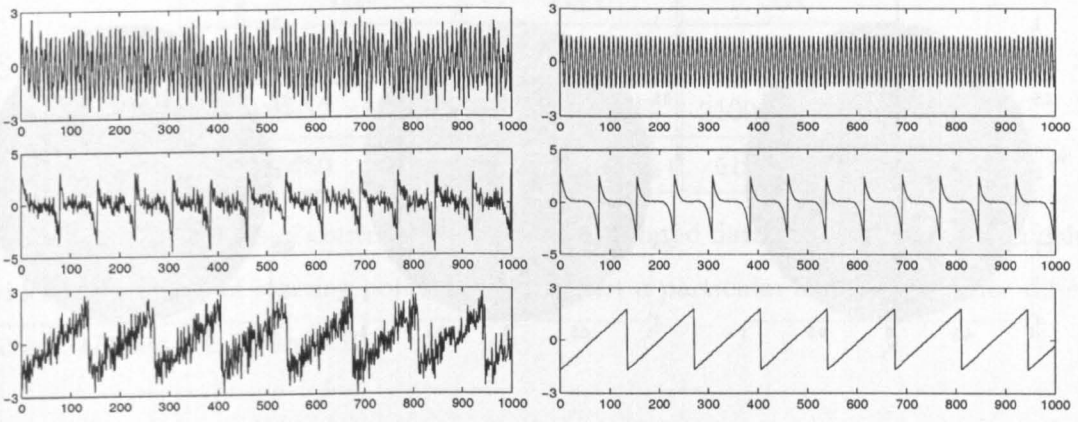
(b) SNR=10dB



(c) SNR=15dB

Figure 4.8: Estimated independent components of noisy data without noise bias correction. The performance of separation gets poorer when the noise level increases.

Finally, we demonstrate the ambiguity of the extraction order in noisy ICA as we did in the noise-free case. We consider additive noise to mixed data of 10000 samples with SNR=10dB. Since the updated mixing matrix $\tilde{A} = ZA$ is known, we can compute the weight vector $\mathbf{w}_i^T$ $(i = 1, 2, 3)$, which yields each of the original sources s_i, as the i^{th} row of $\tilde{A}^{-1}$. The normalised weight vectors $\mathbf{w}_i$ are summarised in table 4.11. Note that these vectors differ from those found in noise-free experiments in table 4.5. First of all, recall that the mixing matrix A used in this section has been normalised so that the variance of $x_c = As_i$ equals to unity. Therefore, it is expected to have different attractor points. This action was solely taken in order to simplify the procedure of adding noise of different level to the clean mixed signals x_c. However, even if we skip this conversion and keep A as in the noise-free case, we



(a) linearly separated components $\tilde{W}\mathbf{z}$ (b) separation of clean mixed signals $\tilde{W}\mathbf{z}_c$

Figure 4.9: Linear/non-linear reconstruction of independent components in noisy ICA.

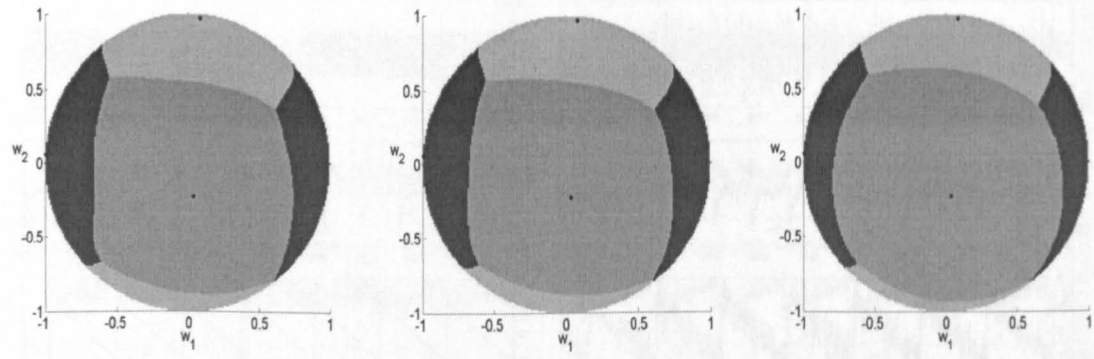
get different attractors because the quasi-whitening transformation always differs from the simple whitening preprocessing of the noise-free section 4.4.1. Hence, we cannot make direct comparisons between regions of convergence in noise-free and noisy situation. Nevertheless, it is interesting to note the intrinsic ambiguity of extraction order in noisy ICA.

	w_{i1}	w_{i2}	w_{i3}	original source
$\mathbf{w}_1$	-0.996	0.061	0.073	sinusoidal s_1
$\mathbf{w}_2$	0.035	-0.229	0.973	funny-shaped s_2
$\mathbf{w}_3$	0.081	0.971	0.227	sawtooth s_3

Table 4.11: Each vector $\mathbf{w}_i = [w_{i1}\ w_{i2}\ w_{i3}]^T$ $(i = 1, 2, 3)$ defines an attractor point.

Keeping the noise constant, we scan the 2D space of w_1 and w_2 with the same step of 0.01 in both directions. For each starting point we estimate the independent components and note the order of extraction. The regions of convergence for each original source, as defined in p.59, are illustrated in figure 4.10 for different choices of the contrast function G. The size of each region is estimated quantitatively in table 4.12. Note again the enlargement of the region of convergence of s_2, which is the original source with the highest absolute value of kurtosis, when G_3 (kurtosis) is used.

Note that in this section, we examined the general case where Gaussian noise is added separately in each observed signal. However, if we consider the special case where a Gaussian



(a) G_1 (b) G_2 (c) G_3

Figure 4.10: Regions of convergence for noisy simulated data (SNR=10dB). A starting point in the dark area yields the sinusoidal signal s_1, in the mid-dark area the funny-shaped signal s_2, and in the light area the sawtooth signal s_3. The attractor points for the original sources are noted with black dots. Note that the region of convergence depends also on the choice of the contrast function G.

source is mixed with the other three original sources, application of ICA will extract that Gaussian as the fourth independent component. Recall that ICA considers non-Gaussian signals, and therefore it can extract one Gaussian component at most as the residual signal when all non-Gaussians have been estimated. The quality of separation and the order of extraction are identical as those in section 4.4.1.

4.4.3 Conclusions

This section demonstrated the application of ICA in practice. The algorithm which was selected to be tested was FastICA due to its appealing property of sequential extraction of independent components as explained in section 4.2.3. We showed that ICA exhibits exceptional performance in decomposing observed linear mixtures of source signals in noise-free and noisy cases. The quality of separation was tested for artificially mixed signals, for which we know the ground truth.

Different contrast functions G were examined providing comparable quality of separation. In fact, the choice of G affected the behaviour of the algorithm under specific situations. For example, the use of G_3 enlarged the region of convergence of s_2 which has the highest

	G_1		G_2		G_3	
s_1	6906	22.0%	6770	21.6%	5123	16.3%
s_2	18057	57.5%	18225	58.0%	21091	67.1%
s_3	6450	20.5%	6418	20.4%	5199	16.6%

Table 4.12: Size of region of convergence in noisy simulated data (SNR=10dB). Each column provides the number of starting points which extract a particular component s_i for different choices of G. See also figure 4.10.

absolute value of kurtosis between the original sources. In addition, G_3 provided slightly better separation for the most sub-Gaussian s_1.

Noisy ICA is an extremely difficult task. The version of noisy ICA which was examined in our study considered the noise covariance matrix Σ to be known. The main difference with the noise-free case was the replacement of the whitening preprocessing step by a procedure called quasi-whitening which takes into account Σ. Moreover, we saw that this technique solves only partially the BSS problem. Although it succeeded in estimating the mixing matrix A (with the expected ambiguities of order, sign, and scale), it could not reconstruct the original sources, unless the "clean" mixed signals were given. The particular inherent weakness of noisy ICA limits further its application in real world problems. In fact, the only way to solve the problem completely is the joint maximum likelihood estimation of both A and s. However, the optimisation problem become intractable and can be applied only in small datasets. Nevertheless, for our simulated data we showed that the particular noise correction which was used, reduced asymptotically the noise bias as we increased the number of noisy observations taken into account for the statistics of ICA. After all, being a pure statistical method, the success or failure of ICA depends on the availability of a sufficient number of samples. A final note about the noise is that the higher the noise level, the more samples we needed in order to achieve a comparable level of separation.

However, in real world applications the noise covariance matrix Σ is often unknown and cannot be estimated. Hence, since there is no prior knowledge about the noise, the noise bias correction technique cannot be applied. Our experiments showed that in this case the standard noise-free ICA can be still applied with relative success as long as the noise level is low.

We also displayed how an estimated independent component can be eliminated from the mixed

data. This procedure is particularly useful for artefact removal in contaminated recordings.

Finally, the intrinsic indeterminacy of ICA in estimating the order of extraction was demonstrated in both noise-free and noisy cases. This is due to the inability of ordinary ICA to incorporate prior knowledge about one or more source signals. This issue will be confronted in the following chapter.

Chapter 5

Constrained Independent Component Analysis

This chapter deals with the practical problem of affecting the extraction order of the independent components. This issue has a major impact in real world applications where near-real time signal processing is required. In section 5.2 we introduce a novel algorithm which incorporates prior information about the source signals in order to favour their extraction. Section 5.3 examines the most widely used optimisation techniques in conjunction with our proposed quality function for constrained ICA with simulated data in order to select the one that performs best. A stochastic method, namely the simulated annealing, is proven to be the most appropriate for our case. Our algorithm is validated with real MEG data in section 5.4. This section also shows the effectiveness of ICA in removing artefacts from heavily contaminated recordings.

5.1 The Motivation

An intrinsic weakness of ordinary ICA is its inability to incorporate prior knowledge. This drawback results in the separation of the original sources in a random order. In the previous chapter we observed how the separation order depends on the choice of the contrast function J_G, and on the starting point which is used in the ICA algorithm. In many applications this may not be a significant issue. Nevertheless, there are cases, such as in clinical MEG environment, where the order of extraction is important in order to either identify the presence

of a particular source in the observed mixed signals, to study a single independent component, or to clean the recordings from an unwanted artefact signal. Moreover, the extraction of the desired source as the first component can result in increased speed in further processing, especially in vast datasets.

Let us assume that some statistical property of one of the original sources is known in advance. Then the standard ICA contrast function can be modified so as to include a penalty term which takes into account this statistical information about that particular signal. The penalty term forces the extraction of that signal first, before the remaining sources are separated as per normal.

In the past, independent components were sequentially extracted in decreasing order of their absolute normalised kurtosis [26], or rearranged depending on their frequency [73]. More recent studies managed to extract periodic signals using the autocorrelation function of the output independent component at a particular time delay $\tau = 1/F$, where F is the fundamental frequency of the desired signal to be extracted [10, 11]. However, the main problem in practical implementation is that of estimating the optimal time delay τ. Recently, [98] introduced constraints into standard ICA in order to eliminate the indeterminacy by solving the constrained optimisation problem using the method of Lagrange multipliers. The proposed algorithm requires the knowledge of a reference signal, and thus the success in extracting the desired independent component depends strongly on the choice of the reference signal and the closeness parameters. Previous work in this domain shows that it is also possible to incorporate prior knowledge into the BSS problem using Bayesian formalism [85, 111]. For example, we can use knowledge about the prior probabilities of the source positions and source amplitudes to determine the prior probability density for a given element of the mixing matrix A [85].

On the other hand, our approach does not really require a reference signal, but rather the prior knowledge of some statistical property of the component we wish to extract first. Therefore, our constraints allow greater flexibility. In our study, we assume that we know the autocorrelation function of the wanted signal. However, any other statistical function may be used with the appropriate modification of the following analysis.

5.2 Incorporating Prior Knowledge in ICA

Let us consider the autocorrelation function of the unknown desired source to be readily available. If $\mathbf{w} = [w_1\ w_2\ \cdots\ w_M]^T$ is the M-dimensional weight vector which gives the estimate of that independent component $y = \mathbf{w}^T\mathbf{x}$, then the value of y at the k^{th} time slice can be written as

$$y_k = w_1x_{1k} + w_2x_{2k} + \ldots + w_Mx_{Mk}, \text{ where } k = 1, 2, \ldots, K \tag{5.1}$$

Therefore, the autocorrelation function r_{yy} of y is given by

$$r_{yy}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} y_{k+\tau}y_k = \frac{1}{K}\sum_{k=0}^{K-1}\left[(w_1x_{1(k+\tau)} + \ldots + w_Mx_{M(k+\tau)})(w_1x_{1k} + \ldots + w_Mx_{Mk})\right]$$

$$\Rightarrow\ r_{yy}(\tau) = w_1\sum_{j=1}^{M} w_jr_{1j}(\tau)\ +\ w_2\sum_{j=1}^{M} w_jr_{2j}(\tau)\ +\ \ldots\ +\ w_M\sum_{j=1}^{M} w_jr_{Mj}(\tau)\ \Rightarrow$$

$$r_{yy}(\tau) = \sum_{i=1}^{M}\left[w_i\sum_{j=1}^{M} w_jr_{ij}(\tau)\right] \tag{5.2}$$

where τ is the time shift (lag), and $r_{ij}(\tau) = \sum_{k=0}^{K-1} x_{i(k+\tau)}x_{jk}$ is the cross-correlation of signals x_i and x_j for time shift τ. When $i = j$, then r_{ii} is the autocorrelation of signal x_i. Having K time slices in total, the lag τ can take K values ($\tau = 0, 1, \ldots, K-1$).

Using vector-matrix notation, equation 5.2 can be written as

$$r_{yy}(\tau) = \mathbf{w}^TA_{\mathbf{x}}(\tau)\mathbf{w} \tag{5.3}$$

where $A_{\mathbf{x}}(\tau)$ is the $(M \times M)$ matrix :

$$A_x(\tau) = \begin{pmatrix} r_{11}(\tau) & r_{12}(\tau) & \cdots & r_{1M}(\tau) \\ r_{21}(\tau) & r_{22}(\tau) & \cdots & r_{2M}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ r_{M1}(\tau) & r_{M2}(\tau) & \cdots & r_{MM}(\tau) \end{pmatrix}$$

Since the signals x_j $(j = 1, 2, \ldots, M)$ are known, it is easy to compute the elements $r_{ij}(\tau)$ of matrix $A_{\mathbf{x}}(\tau)$ for all possible values of τ.

In *constrained* ICA (cICA) we introduce here, the estimated independent component y should match a model signal, say s_{model}. The model signal is unknown, but we assume that we know some statistic of it, say its autocorrelation function $r_{model}(\tau)$ for all possible values of τ.

Consequently, in cICA we aim to choose the weight vector $\mathbf{w}$ so that we minimise the penalty term:

$$J_C(\mathbf{w}) = \sum_{\tau=0}^{K-1} [r_{yy}(\tau) - r_{s_{model}}(\tau)]^2 \tag{5.4}$$

Thus, from equations 5.3 and 5.4:

$$J_C(\mathbf{w}) = \sum_{\tau=0}^{K-1} \left[\mathbf{w}^{\mathrm{T}} A_{\mathbf{x}}(\tau)\mathbf{w} - r_{s_{model}}(\tau)\right]^2 \tag{5.5}$$

The last equation can be rewritten in a more analytical way as

$$J_C(w_1, w_2, \ldots, w_M) = \sum_{\tau=0}^{K-1} \left\{ \sum_{i=1}^{M} \left[w_i \sum_{j=1}^{M} w_j r_{ij}(\tau) \right] - r_{model}(\tau) \right\}^2 \tag{5.6}$$

The two terms, J_G of standard ICA (see equation 4.18 in p.48) and J_C of cICA, can be combined into a single quality function J which should be maximised and can he either

$$J_1(\mathbf{w}) = J_G(\mathbf{w}) - \lambda J_C(\mathbf{w}) \tag{5.7}$$

or alternatively

$$J_2(\mathbf{w}) = J_G(\mathbf{w}) + \frac{\lambda}{J_C(\mathbf{w}) + \alpha} \tag{5.8}$$

where $\lambda > 0$ is a weighting factor, and α is a small constant in order to avoid singularities.

Recall from p.34 that the approximations of negentropy J_G require y to be of unit variance. Hence, function J (either as J_1 or J_2) has to be maximised under the constraint:

$$E\{(\mathbf{w}^T\mathbf{x})^2\} = 1 \tag{5.9}$$

If we further whiten the data as in section 4.1.2, the constraint is written as

$$\sum_{j=1}^{M} w_j^2 = 1 \tag{5.10}$$

Note that as we have seen in the previous chapter, any approximation of negentropy J_G yields comparable levels of separation. Therefore, without loss of generality and in order to simplify our calculations, we choose to approximate J_G using contrast function G_3. In other words, we will make use of kurtosis for the standard ICA term J_G for the experimental analysis that follows. Similar results can be obtained by employing either G_1 or G_2 modifying the algorithms accordingly. In consequence, if $\mathbf{z}$ denotes the whitened data,

$$J_G(\mathbf{w}) = \left(E\{(\mathbf{w}^T\mathbf{z})^4\} - 3\right)^2 = \left[\frac{1}{K}\sum_{k=1}^{K}\left\{\left(\sum_{j=1}^{M} w_i z_{ik}\right)^4\right\} - 3\right]^2 \tag{5.11}$$

5.3 Optimisation Process

There are several methods to maximise function J. In the following, we will examine the most popular optimisation techniques which are used in signal processing, and choose the most suitable for our quality function. For a comprehensive presentation of the optimisation techniques used in this section, see [6, 48, 130].

The optimisation methods will be validated with the simulated data used in the previous chapter. First, the mixed signals $\mathbf{x}$ are centered and whitened as in section 4.4.1. See also figure 4.3 in p.57. The attractor points of the source signals $\mathbf{s}$ for unconstrained ICA were given in table 4.5.

Our study focuses on the possibility of forcing the extraction of a particular signal, which is already known to exist in the set of source signals and for which we know only the autocorrelation function, as the first component independently of the starting point used in the optimisation algorithm. For example, assume that we are aware of the autocorrelation function of a model signal similar to that of the sawtooth signal s_3. The normalised autocorrelation function for the model signal, essential for the term J_C of our quality function J, is shown in figure 5.1.

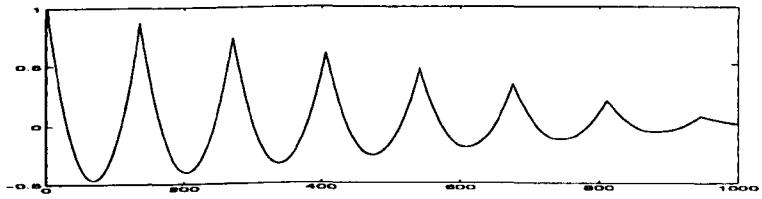


Figure 5.1: Normalised autocorrelation function of a model sawtooth signal.

5.3.1 Steepest Ascent

5.3.1.1 The Method

Steepest ascent is a straightforward maximisation technique. It belongs to the general family of *gradient methods* which are named after the use of first-order derivatives of the quality (contrast) function. In our case, the derivatives can be obtained in an explicit form using analytical methods. However, the drawback is that the computation of the gradient vector ∇J requires significantly more processing time than evaluating the quality function J at a given

point in the M-dimensional space. The calculation of ∇J, along with some simplifications which are essential in order to ease the computational burden, are provided in appendix B. The quality function which we opt to maximise with steepest ascent is

$$J(\mathbf{w}) = J_G(\mathbf{w}) - \lambda J_C(\mathbf{w}) \tag{5.12}$$

Note that steepest ascent gives the optimum results for quadratic functions.

Function J is a non-linear function of M unknowns. It is expected to have several local maxima. In order to take into consideration the constraint that the weight vector has to have unit length, we opt to consider unknown w_M as dependent variable through the equation:

$$w_M = \left(1 - \sum_{i=1}^{M-1} w_i^2\right)^{1/2} \tag{5.13}$$

Hence, the problem is reduced to an (M-1)-dimensional optimisation task. Note also that we decide to use always the positive sign when we take the square root in equation 5.13. This does not imply any loss of generality, since ICA suffers from the ambiguity of the sign.

The algorithm is summarised step by step in table 5.1. Recall that before applying the actual algorithm the data should be whitened. Then we must choose a starting point $\mathbf{w}_0 = [w_{1_0}, w_{2_0}, \ldots, w_{M_0}]$ within the region of feasible solutions which is defined by the constraint of equation 5.10. For that point, we calculate the gradient vector $\nabla J(\mathbf{w}_0)$. This provides the direction along which the next point $\mathbf{w}_1$ is to be chosen for some small step μ as $\mathbf{w}_1 = \mathbf{w}_0 + \mu \nabla J(\mathbf{w}_0)$. The gradient is re-evaluated at the new point, and another point is determined. The procedure is repeated until a point is found where the gradient becomes sufficiently small (or ideally zero), or when a user-specified maximum number of iterations Q is exceeded. Recall that the unknown w_M is a dependent variable calculated from equation 5.13, hence the gradient vector $\nabla J(\mathbf{w})$ consists of $M - 1$ elements.

The step parameter μ can be either constant for all iterations, or proportional to the magnitude of $\nabla J(\mathbf{w})$ at each iteration. The latter allows big steps on steep surfaces and small steps on rather flat ones. However, in this case it is possible to overshoot the maximum, and thus the algorithm will oscillate about the maximum point. Therefore, we opt for constant step providing a safer but slower convergence than the variable step method. Moreover, the gradient vector is normalised in each iteration so as to have a constant step size, equal to μ.

Note that the algorithm fails if by chance a stationary point (i.e. a point yielding a zero or near-zero gradient) turns out to be a saddle point instead of a maximum point. In this case,

we can pick a nearby point in the direction of increasing J and continue with the steepest ascent.

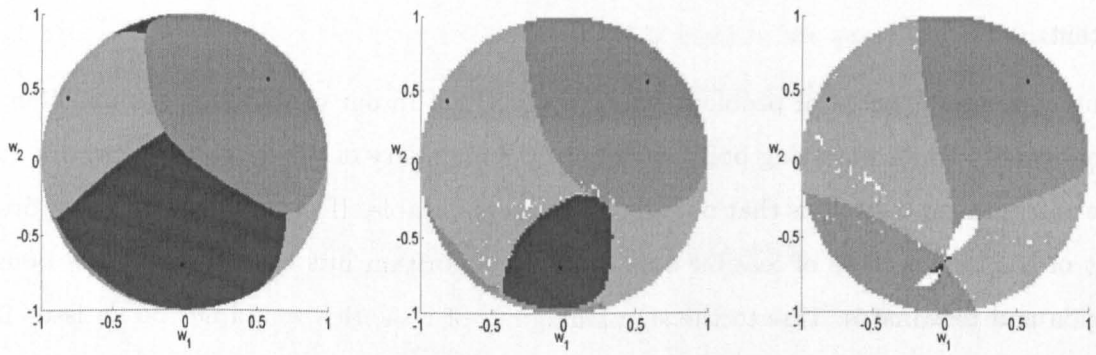
Finally, a minor technical problem which is often met in our constrained optimisation task (particularly when a starting point is close to the boundary of the feasible region) due to the use of a gradient method is that not all directions are usable. If the gradient vector is directed out of the hypersphere of feasible solutions, the algorithm hits the barrier of the bounded region and terminates. Due to the sign ambiguity of ICA, this would not be an issue if the dependent variable w_M was allowed to take the negative sign in the square root of equation 5.13. However, the problem can be solved if we inverse the signs of the independent variables w_i $(i = 1, 2, \ldots, M-1)$ just before they hit the boundary, and continue applying steepest ascent. Geometrically this operation is equivalent to transferring the point which is under examination to its symmetric point inside the hypersphere.

1. Take a random initial vector $\mathbf{w}_0$ of unit norm. Calculate $\nabla J(\mathbf{w}_0)$.
   Let $k = 1$.
2. Normalise $\nabla J(\mathbf{w}_{k-1})$. Let $\mathbf{w}_k = \mathbf{w}_{k-1} + \mu \nabla J(\mathbf{w}_{k-1})$.
3. Calculate $\nabla J(\mathbf{w}_k)$. If $|\nabla J(\mathbf{w}_k)| < \tau$, output vector $\mathbf{w}_k$.
   Otherwise let $k = k + 1$. If $k < Q$ go back to step 2. If not, terminate.

Table 5.1: Algorithm for gradient method of steepest ascent in cICA. The final vector $\mathbf{w}_k$ provides an estimation of the desired original source as $\mathbf{w}_k^T \mathbf{z}$, where $\mathbf{z}$ are the whitened data. The index k denotes the k^{th} iteration of the algorithm. μ is the step size which is considered to be constant in each iteration, and τ is the tolerance, i.e. an arbitrarily chosen small value which is used to terminate the algorithm, Q is the maximum number of iterations allowed.

5.3.1.2 Experimental Results with Simulated Data

Now is the time to test our quality function J employing the steepest ascent method with our simulated data. First, we maximise J considering no constraint (i.e. $\lambda = 0$). We scan the 2D space of feasible starting points with a step of 0.01 in both directions. The regions of convergence depending on the starting point are depicted in figure 5.2(a). Notice that the steepest ascent performs smoothly in estimating the local maxima of J_G. Yet the convergence requires significantly more iterations than the FastICA algorithm.



(a) $\lambda = 0$ (b) $\lambda = 0.005$ (c) $\lambda = 0.025$

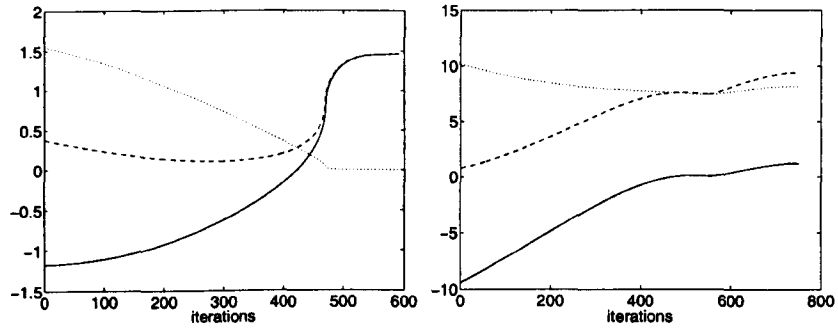
Figure 5.2: Regions of convergence for simulated data in cICA using steepest ascent. A starting point in the dark area yields the sinusoidal signal s_1, in the mid-dark area the funny-shaped signal s_2, and in the light area the sawtooth signal s_3. The attractor points for the original sources are noted with black dots. The white spots indicate starting points for which the algorithm fails to converge. Notice that an increase in the weighting factor λ enlarges the region of convergence of the desired sawtooth s_3.

Next, we gradually increase the value of the weighting factor λ in equation 5.12 in order to take into account the prior knowledge about the model autocorrelation function of figure 5.1, and repeat the procedure. In practice, the use of steepest ascent with our quality function was proven to be problematic when the constraint term J_C was activated (i.e. for $\lambda > 0$) (see figure 5.2).

Note that since the funny-shaped signal s_2 has the highest value of absolute kurtosis, it is expected that the starting points which normally converge to its attractor, will require an increased value for λ in order to redirect them to the attractor point of the desired sawtooth signal s_3. However, despite experimenting with diverse running conditions of the algorithm, such as weighting factor λ, step size μ, or tolerance τ, the constraint λJ_C fails effectively to influence these starting points which still extract signal s_2 (see also figure 5.3).

Moreover, there are quite a few starting points for which the algorithm reaches a plateau and fails to converge to one of the attractors. Nevertheless, an increase in λ results in the enlargement of the region of convergence of the desired sawtooth signal s_3 over the sinusoidal signal s_1. However, it is clear that the method of steepest ascent is very sensitive in maximising our quality function J, and thus alternative optimisation techniques should be

tested.



(a) $w_{1_0} = 0.40$, $w_{2_0} = -0.40$ (b) $w_{1_0} = 0.40$, $w_{2_0} = 0.00$

Figure 5.3: Simulated data. $\lambda = 0.100$, $\mu = 0.001$, $\tau = 0.001$. Standard ICA term J_G (dashed line), constraint term λJ_C (dotted line), and quality function $J = J_G - \lambda J_C$ (solid line) are shown as functions of iterations for two different starting points $[w_{1_0}, w_{2_0}]$. Graph 5.3(a) describes the typical behaviour of a starting point which extracts the desired signal. Graph 5.3(b) corresponds to a starting point for which the algorithm fails to influence the extraction despite experimentation with a wide range of parameters λ,μ,τ.

5.3.2 Simplex

5.3.2.1 The Method

Another popular optimisation approach is the simplex method [119]. Simplex does not require derivatives; only function evaluations. However, it is not one of the most efficient optimisation techniques in terms of number of iterations. The simplex method can be explained geometrically. In an M-dimensional space, a simplex is a geometrical figure that consists of $M+1$ fully interconnected vertices. For example, in optimisation problems of three variables, the simplex is a tetrahedron with four vertices. The method does not just start with a single point, but with $M+1$ points defining the initial, non-degenerate simplex, i.e. a simplex which encloses a finite inner M-dimensional volume. The aim is to enclose the maximum inside the volume of the final simplex.

Then the algorithm takes a series of steps, each moving a point in the simplex away from where the function is lowest. The simplex can be expanded, contracted, and reflected, depending on the minimal/maximal values of the quality function found at the corner points of the simplex.

The algorithm finds first the points where the quality function is highest (high point) and lowest (low point). Then it reflects the simplex away from the low point. If the solution is better (i.e. higher value of the quality function), it tries an expansion in that direction; otherwise it picks an intermediate point. Each operation defines new simplex corner points by linear combinations of the existing corner points. If no improvement is reached after a number of steps, the simplex is contracted, and started again.

The size of simplex is continuously changed and mostly diminished, so that finally it is small enough to contain the maximum with the desired accuracy. The algorithm terminates when the increase in the value of the quality function in the terminating step is smaller than some tolerance. Note that this termination criterion makes the method sensitive to a single anomalous step, and therefore it may fail in finding the maximum.

Preliminary experiments showed that the quality function J which was defined by equation 5.12 and optimised with steepest ascent in the previous section, yields mediocre results with simplex. Note that steepest ascent is ideal for quadratic or near-quadratic functions. Hence, the quality function which is to be maximised with simplex is:

$$J(\mathbf{w}) = J_G(\mathbf{w}) + \frac{\lambda}{J_C(\mathbf{w}) + \alpha_1} \tag{5.14}$$

where λ is a weighting factor, and α_1 is a small constant in order to avoid singularities.

The constraint of equation 5.10 can be incorporated into J as an additional penalty term J_W which has to be minimised:

$$J_W(\mathbf{w}) = \left(\sum_{j=1}^{M} w_j^2 - 1 \right)^2 \tag{5.15}$$

In consequence, the quality function J can be rewritten as

$$J(\mathbf{w}) = J_G(\mathbf{w}) + \frac{\lambda}{J_C(\mathbf{w}) + \alpha_1} + \frac{\mu}{J_W(\mathbf{w}) + \alpha_2} \tag{5.16}$$

where μ is the additional weighting factor for J_W, and α_2 is a constant in order to eliminate singularities as before. The constraint term J_C, as defined in equation 5.6, is highly computationally intensive when calculated in a straightforward way in an iterative algorithm, especially for real MEG data which typically consist of thousands of time samples. For this reason, some rearrangements in the order of summations are required. The simplifications are given in appendix C.

5.3.2.2 Experimental Results with Simulated Data

First, we perform ordinary ICA to our simulated data without any constraint (i.e. $\lambda = 0$) using the simplex method. The constant α_2 is set to 0.001. Its value does not really affect the algorithm, since the penalty term J_W is specifically adjusted by parameter μ.

If we scan the 2D space of feasible points (w_1, w_2) with a step of 0.015 in both dimensions, we get 13952 discrete points which are used as starting points in the simplex algorithm. Note that w_1, w_2, and w_3 are actually considered as three independent variables, and the simplex corner points are allowed to lie outside the field of feasible solutions in any particular step. However, the penalty term J_W restores the final point within the feasible field.

Our first task is to determine the optimum value for the weighting factor μ. Therefore, we commence our search by giving to μ very low values, close to zero. We notice that when μ increases, the number of points which converge to one of the three possible attractors also increases. In fact, when μ is so high that the penalty term $\frac{\mu}{J_W+\alpha_2}$ is comparable with the standard ICA term J_G, then all starting points converge to one of the three attractors. This happens when $\mu = 0.0108$. If we further increase μ, most of the starting points still converge to the attractors. However, for some of them the simplex method fails to converge; the higher the value of μ, the higher the number of starting points for which the algorithm fails (see table 5.2). Even so, for $\mu = 1$ the algorithm succeeds for more than 99% of the starting points.

Then we apply constrained ICA maximising J of equation 5.16. As before, the component we aim to extract first is the sawtooth signal s_3. We study the convergence of our set of starting points for several different values of μ and λ. The results are summarised in tables 5.3-5.8. From these tables we can clearly see that if we increase the value of λ, we can enlarge the size of the region of convergence for our desired sawtooth component (see also figure 5.4). Moreover, if λ is high enough (e.g. $\lambda = 2000$) the algorithm can extract the desired component first with almost absolute success. Comparing tables 5.3-5.8 we also notice that λ and μ are independent variables. The percentage of starting points converging to each of the attractors depends only on the value of λ, and is nearly the same for different values of μ. Therefore, in order to have the optimum result, we should choose a rather small value of μ (e.g. $\mu = 0.05$) so that we reduce the number of starting points for which the algorithm fails, and a high value of λ (e.g. $\lambda = 2000$) so that the constraint is strong enough to force the extraction of our desired independent component first.

μ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0.0108	5379	5367	3206	13952	0
	38.5%	38.5%	23.0%	100%	0%
0.05	5394	5374	3184	13952	0
	38.7%	38.5%	22.8%	100%	0%
1	5358	5363	3136	13857	95
	38.7%	38.7%	22.6%	99.3%	0.7%
5	5224	5330	3041	13595	357
	38.4%	39.2%	22.4%	97.4%	2.6%
10	5152	5298	2967	13417	535
	38.4%	39.5%	22.1%	96.2%	3.8%
20	4983	5226	2889	13098	854
	38.0%	39.9%	22.1%	93.9%	6.1%
50	4508	5110	2539	12157	1795
	37.1%	42.0%	20.9%	87.1%	12.9%

Table 5.2: ICA without constraint ($\lambda = 0$) using simplex. The weighting factor μ adjusts penalty term J_W. Columns $2-4$ give the number of starting points converging to the attractors of the source signals. The minimum value of μ in order to achieve convergence for all starting points is $\mu = 0.0108$. However, if we increase μ, we also increase the number of points for which the algorithm fails to converge.

In conclusion, we can see that the simplex method is rather successful in maximising our proposed quality function J of equation 5.16. Applying the right parameters of the weighting factors λ and μ, we can always extract the desired independent component first. However, simplex is indicated for optimising a function of a small number of variables [17]. In consequence, the use of simplex in multidimensional real world applications, such as in MEG data, can be shown to be erratic. Thus, we should try to identify an alternative optimisation method which can be applied in multivariate problems in an equally efficient way.

λ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0	5379	5367	3206	13952	0
	38.5%	38.5%	23.0%	100%	0%
500	1685	3213	8800	13698	254
	12.3%	23.5%	64.2%	98.2%	1.8%
1500	15	1503	11912	13430	522
	0.1%	11.2%	88.7%	96.3%	3.7%
2000	2	1	10842	10845	3107
	0.02%	0.01%	99.97%	77.7%	22.3%

Table 5.3: $\mu = 0.0108$. ICA using simplex with constraint ($\lambda > 0$). Columns $2-4$ give the number of starting points converging to the attractors of the source signals. For small values of λ ($\lambda = 500$) the algorithm is partially effective. It extracts the desired sawtooth signal as the first component for more starting points than when $\lambda = 0$. However, there are still many points which converge to undesired attractors. Increasing the value of λ finally leads to the extraction of our desired component with almost absolute success ($\lambda = 2000$).

λ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0	5394	5374	3184	13952	0
	38.7%	38.5%	22.8%	100%	0%
500	1696	3440	8816	13592	0
	12.1%	24.7%	63.2%	100%	0%
1500	16	1527	12400	13943	9
	0.1%	11.0%	88.9%	99.9%	0.1%
2000	3	4	13940	13947	5
	0.02%	0.03%	99.95%	99.96%	0.04%

Table 5.4: $\mu = 0.05$. ICA using simplex with constraint ($\lambda > 0$). Columns $2-4$ give the number of starting points converging to the attractors of the source signals.

λ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0	5358	5363	3136	13857	95
	38.7%	38.7%	22.6%	99.3%	0.7%
500	1713	3536	8646	13895	57
	12.3%	25.5%	62.2%	99.6%	0.4%
1500	6	1627	12285	13918	34
	0.04%	11.69%	88.27%	99.8%	0.2%
2000	2	8	13917	13927	25
	0.01%	0.06%	99.93%	99.8%	0.2%

Table 5.5: $\mu = 1$. ICA using simplex with constraint ($\lambda > 0$). Columns $2-4$ give the number of starting points converging to the attractors of the source signals.

λ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0	5152	5298	2967	13417	535
	38.4%	39.5%	22.1%	96.2%	3.8%
500	1669	3502	8500	13671	281
	12.2%	25.6%	62.2%	98.0%	2.0%
1500	10	1611	12098	13719	233
	0.1%	11.7%	88.2%	98.3%	1.7%
2000	1	28	13736	13765	187
	0.01%	0.20%	99.79%	98.7%	1.3%

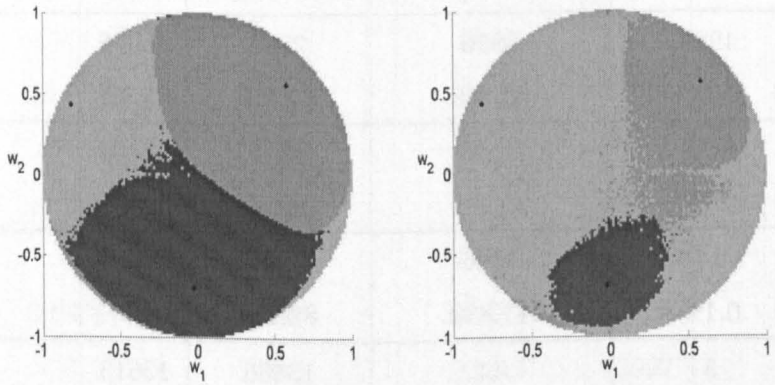
Table 5.6: $\mu = 10$. ICA using simplex with constraint ($\lambda > 0$). Columns $2-4$ give the number of starting points converging to the attractors of the source signals.

λ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0	4983	5226	2889	13098	854
	38.0%	39.9%	22.1%	93.9%	6.1%
500	1653	3466	8371	13490	462
	12.3%	25.7%	62.0%	96.7%	3.3%
1500	14	1606	11985	13605	347
	0.1%	11.8%	88.1%	97.5%	2.5%
2000	3	44	13566	13613	339
	0.02%	0.32%	99.65%	97.6%	2.4%

Table 5.7: $\mu = 20$. ICA using simplex with constraint ($\lambda > 0$). Columns $2-4$ give the number of starting points converging to the attractors of the source signals.

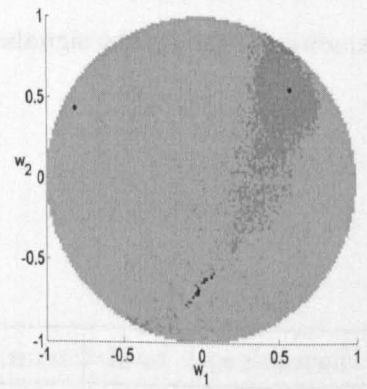
λ	sinusoidal s_1	funny-shaped s_2	sawtooth s_3	total	no convergence
0	4508	5110	2539	12157	1795
	37.1%	42.0%	20.9%	87.1%	12.9%
500	1558	3393	8184	13135	817
	11.9%	25.8%	62.3%	94.1%	5.9%
1500	26	1549	11734	13309	643
	0.2%	11.6%	88.2%	95.4%	4.6%
2000	8	72	13041	13121	831
	0.1%	0.5%	99.4%	94.0%	6.0%

Table 5.8: $\mu = 50$. ICA using simplex with constraint ($\lambda > 0$). Columns $2-4$ give the number of starting points converging to the attractors of the source signals.

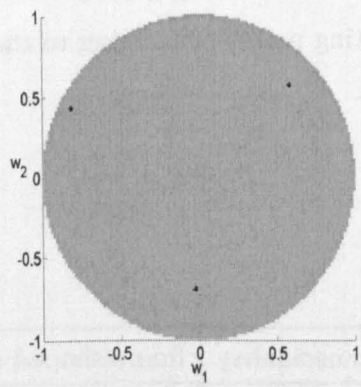


(a) $\mu = 0.05$, $\lambda = 0$

(b) $\mu = 0.05$, $\lambda = 500$



(c) $\mu = 0.05$, $\lambda = 1500$



(d) $\mu = 0.05$, $\lambda = 2000$

Figure 5.4: Regions of convergence for simulated data in cICA using simplex. λ and μ are the weighting factors of terms J_C and J_W respectively. Depending on the starting point a particular component is extracted first. A starting point in the dark area yields the sinusoidal signal s_1, in the mid-dark area the funny-shaped signal s_2, and in the light area the sawtooth signal s_3. The attractor points for the original sources are noted with black dots. We notice that the increase in λ results in enlargement of the region of convergence for our desired sawtooth signal. For a high value of λ ($\lambda = 2000$) our algorithm manages to extract the desired sawtooth signal as the first independent component regardless of the starting point.

5.3.3 Simulated Annealing

5.3.3.1 The Method

Simulated annealing is an intelligent, stochastic technique which is often used for optimisation problems of large scale [83, 130]. It is suitable for those cases where a desired global extremum is hidden among many local extrema. Similarly to simplex, simulated annealing uses only function evaluations instead of derivatives.

Simulated annealing imitates the way that metals cool and anneal. At high temperatures the molecules of a metal in liquid state are able to move freely with respect to each another. If the metal is cooled slowly, then its atoms line themselves up and form ordered crystals. The energy of this system is the lowest possible. However, if the metal is cooled quickly, it forms an amorphous mass having somewhat higher energy.

When simulated annealing is used in function optimisation, the function that has to be optimised plays the role of the energy. There is also a control parameter T which is the simulation analogue to temperature. In general, when simulated annealing is used for function maximization, it takes an uphill step most of the times. However, sometimes it allows taking a downhill step in order to find the global maximum. The efficiency of the method depends on the rate of cooling α and on how low the control parameter T can go. If temperature T is being reduced too fast, we will reach a local maximum, but not necessarily the global maximum.

In our case we wish to maximise the quality function J:

$$J(\mathbf{w}) = J_G(\mathbf{w}) + \frac{\lambda}{J_C(\mathbf{w}) + \alpha_1} \tag{5.17}$$

Simulated annealing is a randomised technique which always evaluates J within the field of feasible points. There are no forbidden directions as in steepest ascent, or corner points which may lie outside the allowed field as in simplex. In consequence, there is no need to incorporate the constraint of equation 5.10 as an extra penalty term J_W as we did before. Hence, we opt to consider unknown w_M as a dependent variable through the equation 5.13. The computation of the constraint term J_C defined in equation 5.6, is simplified in appendix C.

The algorithm is described step by step in table 5.9. To start the algorithm we set an initial value T_0 for the control parameter. We pick a random configuration of the weighting vector

$\mathbf{w}$, and calculate the quality function for this configuration. Then we choose at random a new value for w_1, and compute the quality function again. If the new value of the quality function $J(\mathbf{w}_{new})$ is higher, we accept the new value of w_1, and proceed in the same way to w_2. However, if the quality function is lower, then we pick a random number q between 0 and 1. If the quantity $exp(\frac{J(\mathbf{w}_{new})-J(\mathbf{w})}{T})$ is higher than q, we accept the value of w_1, otherwise we keep the old configuration of $\mathbf{w}$. This step is important because by allowing a downhill step we avoid a potential local maximum. When we have visited all w_i, we consider that we have completed a full iteration. Then we reduce the temperature T and repeat the above procedure. Note that as the temperature T is being lowered, the possibility of a downhill step is reduced. The algorithm continues until the system has cooled down satisfactorily.

1. Set an initial value T_0 for the control parameter T, and $k = 1$.
2. Pick a random configuration of the weighting vector $\mathbf{w}$.
3. Calculate $J(\mathbf{w})$ for this configuration, say $J(\mathbf{w}_{old})$.
4. Set $T_k = \alpha T_{k-1}$, where α is the rate of cooling.
5. Choose at random a new value for w_i, and compute $J(\mathbf{w})$ again, say $J(\mathbf{w}_{new})$.
6. If $J(\mathbf{w}_{new}) > J(\mathbf{w}_{old})$, accept the new value of w_i.
   Set $J(\mathbf{w}_{old}) = J(\mathbf{w}_{new})$ and go to step 8.
7. If $J(\mathbf{w}_{new}) < J(\mathbf{w}_{old})$, pick a random q between 0 and 1.
   If $exp(\frac{J(\mathbf{w}_{new})-J(\mathbf{w}_{old})}{T}) > q$, accept the new value of w_i.
   Set $J(\mathbf{w}_{old}) = J(\mathbf{w}_{new})$.
8. If we have not visited all w_i, go to step 5 for a new w_i.
   Otherwise $k = k + 1$. If $k < Q$, go back to step 4.
   If not, output vector $\mathbf{w}_{old}$ and terminate.

Table 5.9: Algorithm for simulated annealing in cICA. The final vector $\mathbf{w}_{old}$ provides an estimation of the desired original source as $\mathbf{w}_{old}^T\mathbf{z}$, where $\mathbf{z}$ are the whitened data. The index k denotes the k^{th} iteration. Q is the maximum number of iterations allowed.

5.3.3.2 Experimental Results with Simulated Data

Now we shall validate our constrained ICA algorithm with the simulated data maximising the quality function J with simulated annealing. As before, the component we aim to extract first is the sawtooth signal s_3.

For example, let us use as a starting point w_s of the weighting vector $\mathbf{w}$ the point $w_s = [0.00 \; -0.60 \; 0.80]$. As we can see in figures 5.2-5.4, for that starting point ordinary ICA extracts the sinusoidal signal s_1 as the first independent component. We apply our method using simulated annealing for different values of the weight factor λ, and the simulated annealing parameters (initial temperature T_0 and maximum number of iterations Q). The rate of cooling is set to $\alpha = 0.99$. For that starting point w_s and for each value of λ, T_0 and Q, we perform constrained ICA 1000 times using a different seed for the random number generator at each trial. The constant α_1 is set to 0.001.

If the value of weight factor λ is small ($\lambda = 0.001$), the constraint term J_C effectively does not participate in the optimisation process. In consequence the quality function J always converges to the point of maximum kurtosis, thus extracting the funny-shaped component s_2 (see table 5.10). Increasing the maximum number of iterations allowed in the simulated annealing algorithm reduces the variance of the point of convergence.

λ	T_0	Q	% tries	$\bar{w}_1$	σ_{w_1}	$\bar{w}_2$	σ_{w_2}	component
0.001	10	2000	100	0.556	0.013	0.556	0.014	funny-shaped s_2
0.001	10	5000	100	0.556	0.008	0.558	0.008	funny-shaped s_2
0.001	10	10000	100	0.556	0.006	0.557	0.006	funny-shaped s_2

Table 5.10: For small values of λ the constraint J_C is ineffective. The algorithm always extracts the funny-shaped s_2 as the first component, since it is the component with the highest kurtosis. More iterations Q reduce the variance of the point of convergence.

Let us increase the value of λ up to a point that the constraint term $\frac{\lambda}{J_C+a_1}$ is comparable with the ICA term J_G. For example, for $\lambda = 0.01$, we notice that our method extracts the desired component. However there are still many instances where the funny-shaped s_2 is extracted first, instead of the sawtooth model (see table 5.11).

A solution to this problem is to make finer adjustments to the values of the simulated annealing parameters (iterations Q and initial temperature T_0). Since simulated annealing is a

λ	T_0	Q	% tries	$\bar{w}_1$	σ_{w_1}	$\bar{w}_2$	σ_{w_2}	component
0.01	10	2000	51.6	-0.815	0.004	0.430	0.011	sawtooth s_3
0.01	10	2000	48.4	0.556	0.014	0.557	0.014	funny-shaped s_2
0.01	10	10000	97.8	-0.815	0.003	0.431	0.009	sawtooth s_3
0.01	10	10000	2.2	0.554	0.004	0.558	0.005	funny-shaped s_2

Table 5.11: Increasing the value of λ leads to the extraction of the desired component. However $\lambda = 0.01$ is still not high enough to allow the extraction of the sawtooth signal s_3 as the first component in all runs of the algorithm. The issue can be rectified to a great extent by increasing the number of iterations Q which are allowed.

stochastic method, if we increase the number of iterations, in practice we increase our chances to converge to the desired point. Compare in table 5.11 the hugely improved results when 10 000 iterations are allowed instead of 2 000. The desired component s_3 is extracted in 97.8% of the trials.

We can also improve the situation by adjusting the value of the initial temperature T_0. As we can see in table 5.12 for the same number of iterations and keeping λ constant, a higher T_0 increases the number of successful tries, but not significantly after a certain point. Note that if T_0 is very low, our algorithm never moves away from the starting point and fails to extract any significant component.

For an even higher value of λ ($\lambda = 0.1$) our algorithm is always successful in extracting the desired sawtooth component as long as the initial temperature T_0 is higher than a particular threshold (see table 5.13). Otherwise the algorithm stays most of the times at the starting point and fails as before to extract any component with physical significance.

In figure 5.5 the standard ICA term J_G (left graph), the constraint term $\frac{\lambda}{J_G+\alpha_1}$ (middle graph), and the quality function J (right graph) are depicted as functions of the number of iterations (for $\lambda = 0.1$, $T_0 = 10$, $\alpha = 0.99$). Notice that in the beginning of simulated annealing, the quality function is mainly influenced by J_G, but after some point the constraint takes over, and the algorithm successfully extracts the desired independent component. The quality function J is not clique dependent like the cases in image processing where simulated annealing is usually applied. Hence, the algorithm allows the system to randomly sample the solution space. That is why the quality function J appears to improve in sudden jumps when

the search reaches the right part of the solution space.

λ	T_0	Q	% tries	$\bar{w}_1$	σ_{w_1}	$\bar{w}_2$	σ_{w_2}	component
0.01	1	5000	34.4	-0.815	0.003	0.432	0.009	sawtooth s_3
0.01	1	5000	3.4	0.552	0.007	0.559	0.005	funny-shaped s_2
0.01	1	5000	62.2	0.000	x	-0.600	x	error
0.01	10	5000	81.4	-0.815	0.004	0.431	0.010	sawtooth s_3
0.01	10	5000	18.6	0.557	0.008	0.557	0.008	funny-shaped s_2
0.01	100	5000	84.6	-0.816	0.004	0.430	0.011	sawtooth s_3
0.01	100	5000	15.4	0.555	0.008	0.558	0.009	funny-shaped s_2

Table 5.12: Increasing the initial temperature T_0 increases our chances to extract the desired sawtooth component s_3 first for the same λ and number of iterations Q.

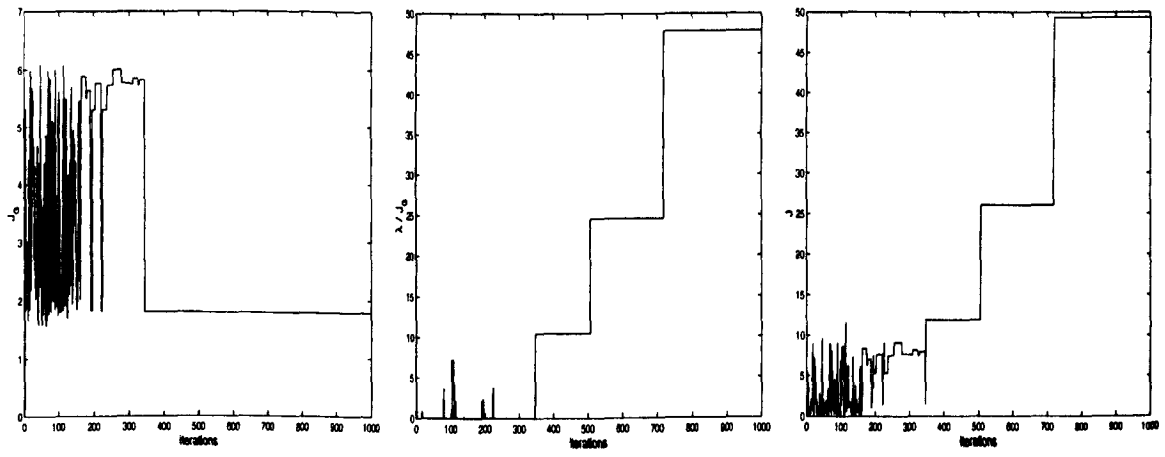
λ	T_0	Q	% tries	$\bar{w}_1$	σ_{w_1}	$\bar{w}_2$	σ_{w_2}	component
0.1	1	2000	5.9	-0.816	0.004	0.428	0.009	sawtooth s_3
0.1	1	2000	94.1	0.000	x	-0.600	x	error
0.1	10	2000	12.8	-0.815	0.005	0.430	0.013	sawtooth s_3
0.1	10	2000	87.2	0.000	x	-0.600	x	error
0.1	100	2000	100.0	-0.815	0.009	0.429	0.018	sawtooth s_3

Table 5.13: For $\lambda = 0.1$ and choosing a sufficiently high initial value for temperature T_0, we manage to extract the desired sawtooth component s_3 with absolute success.

Note that in our simulated data we considered the linear, noise-free ICA model. As explained in detail in sections 4.3 and 4.4.2, the addition of a noise term to the observations makes the problem very difficult to solve even in the unconstrained form. Certain hypotheses should be made which in general are not valid in real world applications. However, we can examine how our algorithm performs when the autocorrelation function is noisy. For that purpose we add Gaussian noise n to the normalised autocorrelation function r_{model} of the model. We notice that our algorithm can extract the desired sawtooth signal when the SNR is as low as 15dB, where the SNR is defined by

$$SNR = 10 \log_{10} \frac{\sum_\tau r^2_{model}(\tau)}{\sum_\tau n^2(\tau)}$$

To sum up, we have showed that simulated annealing can be employed successfully as an



(a) J_G (b) $\frac{\lambda}{J_G+a_1}$ (c) J

Figure 5.5: Simulated data. $\lambda = 0.1$, $T_0 = 10$, $\alpha = 0.99$. Standard ICA term J_G, constraint term $\frac{\lambda}{J_G+a_1}$, and quality function J are shown as functions of the number of iterations. In the beginning, J is mainly influenced by J_G, but after some point the constraint takes over, and the algorithm successfully extracts the desired component.

alternative to simplex in maximising our quality function J which was defined by equation 5.8. The principal advantage of simulated annealing is that it can be used effectively in multidimensional problems which are frequently met in biomedical studies, whereas simplex is recommended for cases of small scale.

In this section we validated with simulated data the modification of the ordinary ICA contrast function so as to incorporate prior knowledge about one or more source signals. Upon the extraction of the desired source signal, we can return to standard ICA to extract the remaining components in random order. If we need to extract in the second place another component for which we know only the autocorrelation function, we may repeat the above procedure. Next, we will apply our method for real MEG data in order to remove artefact signals which contaminate the recordings.

5.4 Application of cICA in Real MEG Data

5.4.1 Data

The MEG data used in this study were collected at the Pitié Salpêtrière Hospital in Paris, France [132]. A participant underwent a series of electrical stimulations applied at the tips of four of his left hand fingers (namely, thumb, index, middle, and little finger). The fingers were stimulated in random order preventing the brain from getting used to a particular pattern. However, the recording was continuous with no gap period between successive stimulations. In total, 1600 trials were recorded. The distribution of the trials based on the stimulated finger is given in table 5.14. The generated magnetic fields were recorded using an 151-channel MEG scanner (Omega 151 Adjustable, CTF Systems Inc. [32]). The topography of the channels over the scalp is provided in figure 5.6. During the recordings the participant kept his eyes open in order to block the activity of alpha rhythm (brainwaves of 8-12Hz) [88].

Each trial is a 300ms recording with a sampling frequency of 1250Hz (i.e. 375 time samples per channel per trial). The first 48ms correspond to the pre-stimulus period (i.e. first 60 samples). The electrical stimulation arrives at the 61^{st} sample of each trial. The total duration of the experiment was 8 minutes (i.e. 1600 trials $\times$ 0.3s per trial). Note that the biological signals of interest (i.e. the signals generated due to the electric stimulus) are expected to arise in the right parietal area of the brain where the somatosensory cortex is located [13] (area marked as *RP* in figure 5.6).

finger	no of trials
thumb	413
index	402
middle	399
little	386
total	1600

Table 5.14: Distribution of trials according to the stimulated finger.

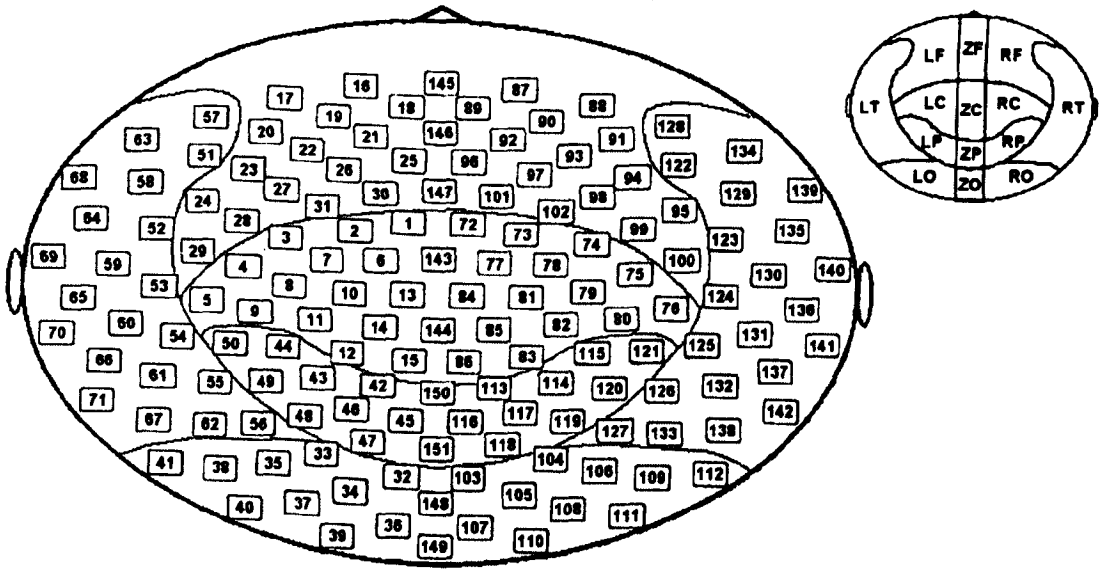
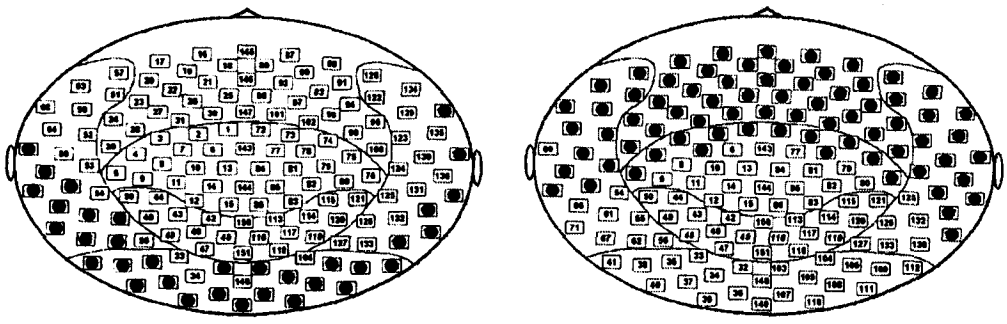


Figure 5.6: Channel topography over the scalp.

5.4.2 Artefacts

The data are contaminated by two major sources; heart interference and eye blinking. Both artefacts contaminate a vast range of channels (see figure 5.7). The cardiac contamination is stronger in the left occipital (LO) and left temporal (LT) areas due to the position and proximity of the heart. Channel 71 is one of the most-heavily heart contaminated channels. Unfortunately there was no parallel ECG recording during the data acquisition. Nevertheless, we can use channel 71 as a reference channel in order to detect spikes due to the heart in channels with low cardiac interference. In addition, channel 71 can be used later in cICA to estimate the autocorrelation function of the model signal. A typical extract from this channel is provided in figure 5.8. The QRS complex which represents the cardiac electromagnetic activity can be seen in detail in figure 5.9.

The ocular artefact due to eye blinking is the dominant one. The artefact is stronger in the frontal (LF/RF) and temporal (LT/RT) areas (see figure 5.7). A 16s extract from a channel located close to the eyes, namely channel 16, is presented in figure 5.10. Note that the ocular artefact is so strong that affects clearly a huge range of channels, even those which are located in central areas. As before, since there is no EOG recorded in parallel with the MEG data, we are going to use channel 16 for the computation of the model autocorrelation function. Overall, 158 ocular bumps are counted in the 8 mins recording resulting in a mean frequency



(a) cardiac contamination (b) ocular contamination

Figure 5.7: Channels where cardiac and ocular interference can be detected by simple visual signal inspection are indicated with black dots.

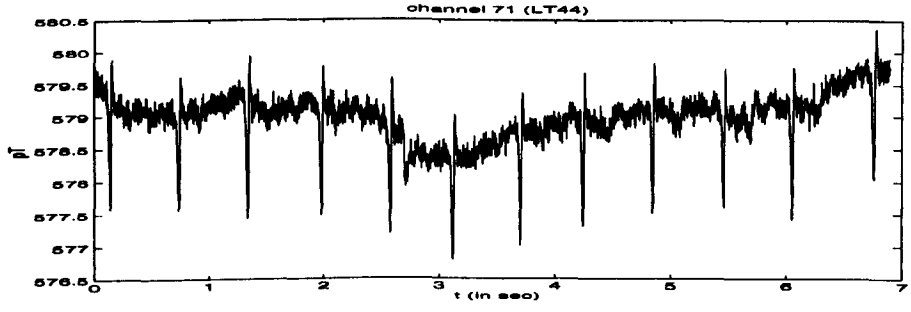


Figure 5.8: Channel 71 with cardiac interference - extract of 6.9s (trials 62-84).

of 20 eye blinks per minute. In figure 5.10 we see that the ocular artefact appears as very wide bumps rather than sharp spikes. This poses an even bigger problem since it affects a huge number of trials. The width of each bump is around 1500ms, therefore it affects at least 4 successive trials. In total it is estimated that at least 600 trials out of 1600 are contaminated. Moreover, due to its increased width it affects the whole trial and not just a small portion as it happens with the cardiac spike.

5.4.3 Artefact Rejection using cICA

The researchers who provided us with the data typically apply raw data averaging in order to cancel out noise or potential artefacts. However, the data are so heavily contaminated by the systematic artefacts described above, so that block trial averaging does not yield any significant improvement. For example, let us sort all the trials according to the finger which

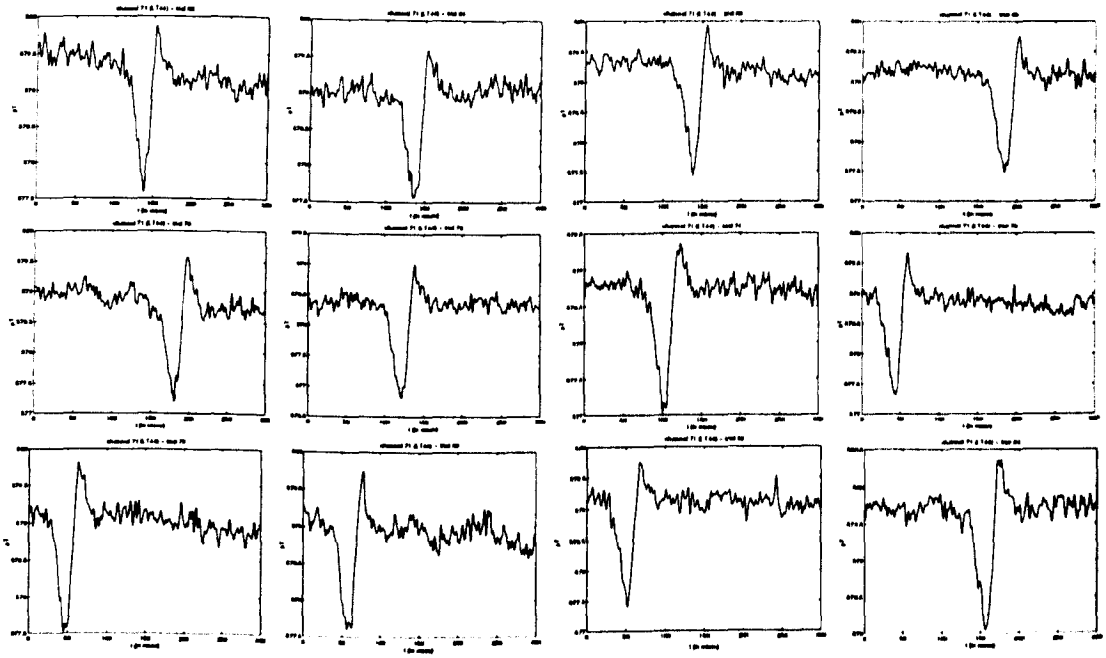


Figure 5.9: Channel 71 - trials {62 64 66 68} {70 72 74 76} {78 80 82 84}. The QRS complex due to cardiac activity is clearly present in those single trials which are affected by the heart interference.

is stimulated, and average them. In figure 5.11 we plot the averages from channel 117. This particular channel is located in the area of interest where the magnetic field produced by the stimulus is expected to be strong. According to [13], we expect a wide peak about 30-40ms after the stimulus (i.e. about 80-90ms from the beginning of each trial if we take into account the pre-stimulus recording). However, we notice that in raw averages it is rather difficult to observe the particular peak. Next, we will show how ICA can improve the data by removing the artefact signals.

The principal advantage of ICA is that it is performed on non-averaged data, and thus all the time samples are used in the computations without discarding arbitrarily any information. In the recent past, ICA has been applied to EEG and MEG data for the removal of ocular and cardiac artefacts [78, 79, 102, 152, 153]. Our MEG study can be considered as a typical BSS problem. Indeed, magnetic fields originated from different biological sources are mixed together instantaneously during their recording. The unknown mixing matrix A is roughly a function of the geometry of the sources. In order for the ICA data model to hold, A is assumed to be constant over time. Moreover, we have many channels available to guarantee that the number of sensors is higher than the number of sources. Their sta-

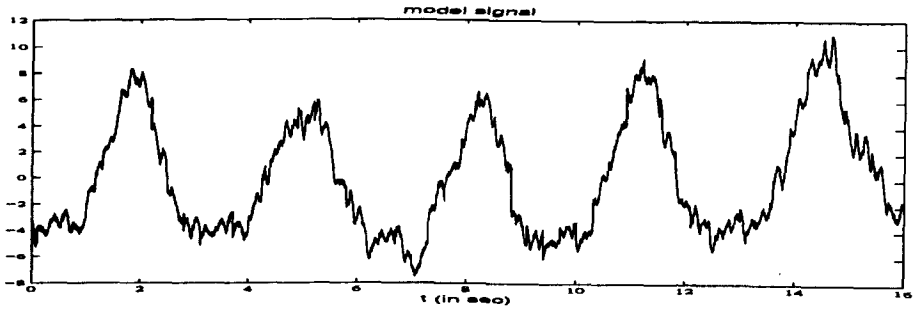
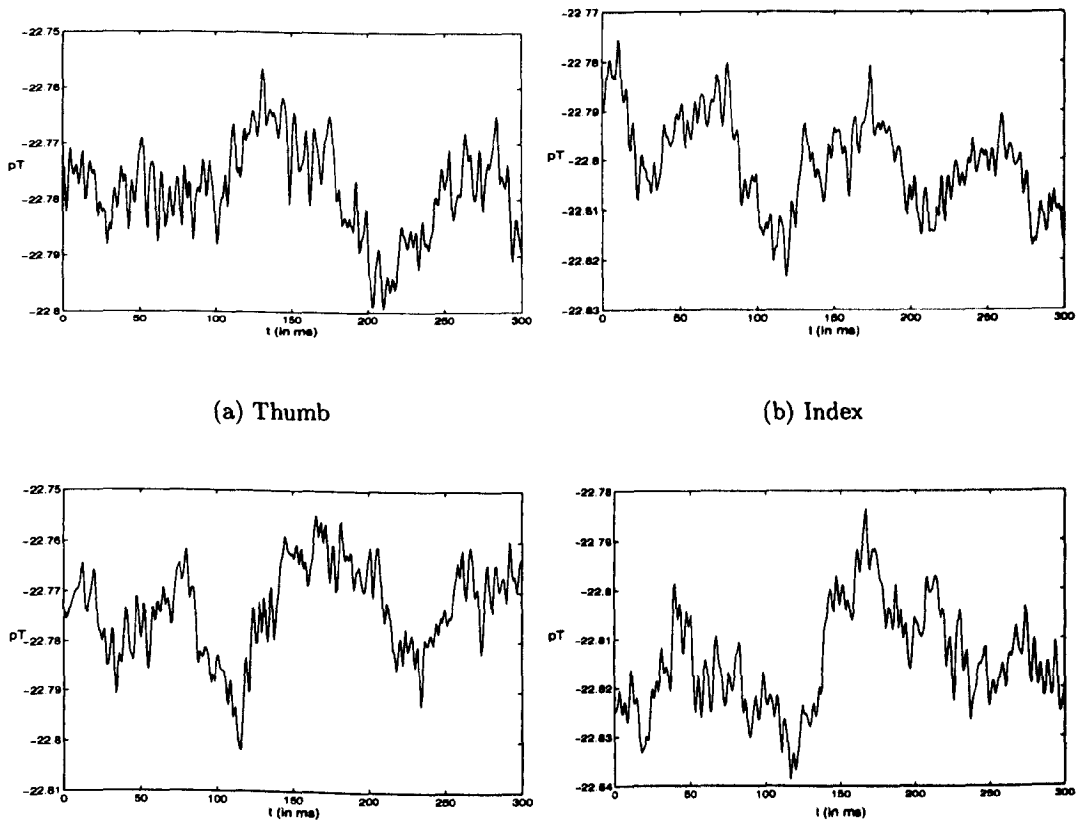


Figure 5.10: Channel 16 is a typical heavily contaminated MEG channel by the ocular artefact. It is positioned in the left frontal area and is also used to compute the model autocorrelation function of the contaminating signal.

tistical independence, which is the fundamental hypothesis of the ICA model, is verified by the different anatomical and physiological processes involved in the production of cerebral biomagnetic signals. However, the brain signals and the ocular artefact may be similarly time-locked to the stimulus, and therefore dependent for a very short time after the stimulus. This is true when an eye blink occurs as a response to the electric stimulation. Nevertheless, the statistical independence is calculated throughout the entire signal. In consequence, we can expect that their close relation during stimulation will not affect their global statistical independence. Finally, according to the researchers who perform quality control of the MEG scanner [133], the recordings can be considered virtually as noise-free since the SNR>30dB. Therefore, the data can be processed using the standard noise-free ICA model. Note that the artefact signals are considered as unknown sources as opposed to the additive electronic Gaussian noise.

Due to the vast size of the data (more than 700Mb), the recordings are split into 30 datasets of 16s each in an attempt to ease the data processing that follows. Therefore, each channel of a particular segment contains 20 000 time samples. The data of each segment are stored in a (151 $\times$ 20 000) matrix. For each segment we perform the following procedure. First, the data are whitened by performing an EVD as described in section 4.1.2. At the same time, we reduce the dimensionality of the dataset to 10 by keeping the ten highest eigenvalues. The percentage of eigenvalues retained ranges between 99.13% and 99.87% for all segments.

Now it is time to apply our cICA algorithm to MEG data in order to clean them from the artefacts. For example, to remove the strong ocular artefact we use the most contaminated channel of the original recordings (see figure 5.10) in order to compute the model autocor-



(a) Thumb

(b) Index

(c) Middle finger

(d) Little finger

Figure 5.11: Averages of raw data for channel 117. x-axis in ms, y-axis in pT. We expect a peak due to the electric stimulus at about 80-90ms. However, the peak of interest is not so easy to spot due to artefact contamination.

relation function. In general, this is not the proper way of estimating the autocorrelation of the artefact signal. However, since this artefact is exceptionally strong and the amplitude of the sources decreases as the inverse square of the distance from the source, we can safely assume that the contribution of other sources in that particular channel is insignificant. The normalised autocorrelation function is given in figure 5.12.

We test our cICA algorithm using simulated annealing for different values of the weight factor λ, and for different values of the simulated annealing parameters (number of iterations Q and initial temperature T_0). For a particular starting point and for each value of λ, Q, and T_0, we perform constrained ICA 100 times using different seed for the random number generator each time. We repeat the same procedure for 100 different starting points, and thus we perform 10 000 experiments in total.

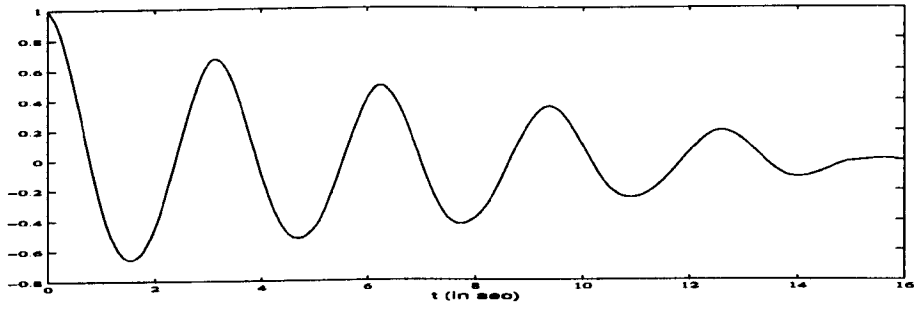


Figure 5.12: Normalised autocorrelation function of the ocular artefact.

If the value of λ is relatively small ($\lambda = 500$), the quality function J converges in the vast majority of trials for each different starting point to the point of maximum kurtosis failing to extract the desired ocular artefact.

If we increase the value of λ so that the constraint term is comparable with the ICA term, for example, for $\lambda = 1000$, the desired component in now extracted in more than half of the trials for each starting point (see figures 5.14). For that value of $\lambda = 1000$, if we increase the number of iterations, thus allowing the system to cool down more, the desired contaminating signal is extracted first in slightly more trials (see figure 5.15). A similar effect can be achieved if we increase the value of the initial temperature T_0 (see figure 5.16).

For an even higher value of λ ($\lambda = 2000$) the algorithm is almost always successful in extracting the desired component first (see figure 5.14). In practice, if λ is higher than 1500, then it is almost certain that the desired signal will be extracted first, as long as the initial temperature T_0 is high enough ($T_0 > 0.1$) to allow downhill steps in order to avoid local maxima. The extracted independent component which corresponds to the ocular artefact is depicted in figure 5.13.

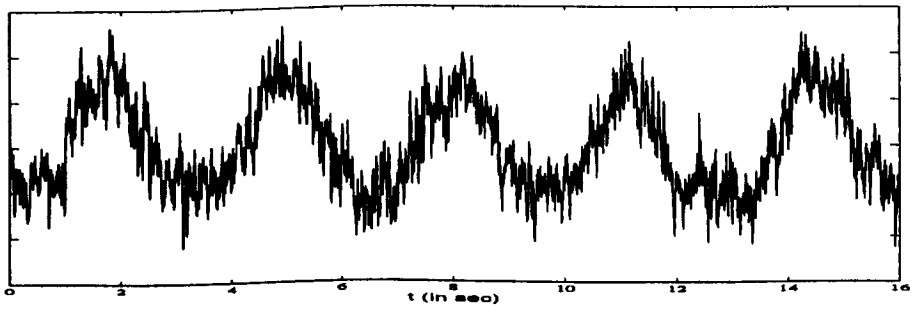


Figure 5.13: Constrained independent component corresponding to the ocular artefact.

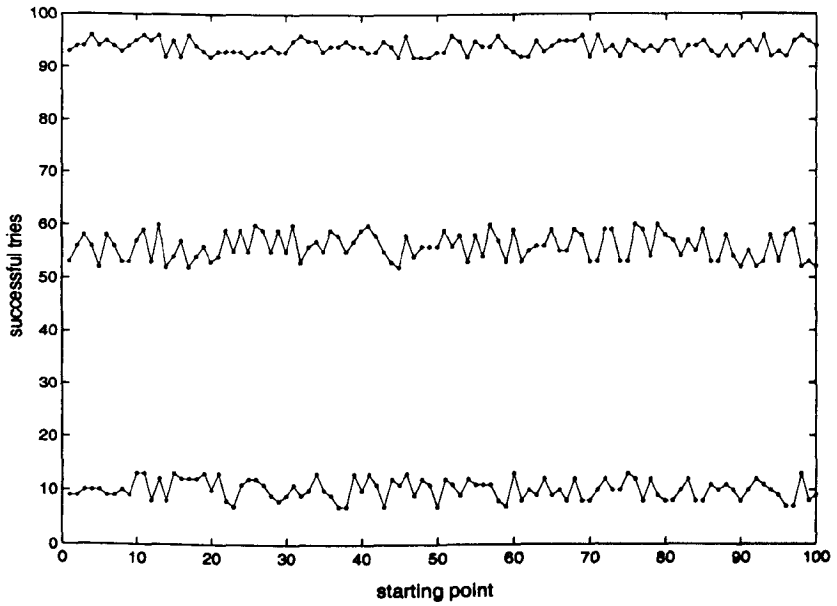


Figure 5.14: Successful tries in extracting the desired ocular component using cICA for each starting point for different values of λ ($T_0 = 10$, iterations $Q = 2000$). For small values of λ ($\lambda = 500$) the constraint is ineffective (lower line). The algorithm extracts most of the times the component with the highest kurtosis. Increasing the value of λ to $\lambda = 1000$ leads to better extraction of the desired contaminating component (middle line). The desired component in now extracted in more than half of the trials for each starting point. For higher values of λ, say $\lambda = 2000$ we manage to extract the contaminating ocular component with nearly absolute success (upper line).

In a similar way we can extract the cardiac interference as the second component calculating the autocorrelation function from the signal in figure 5.8. By applying ICA we actually estimate the weight vector $\mathbf{w}$ which describes the contribution of the artefact signals to each MEG channel. Therefore, it is now easy to remove the artefact signals from the original recordings and have them cleaned as we did in p.61 with the simulated data.

We repeat the procedure described above for all 30 datasets. The effectiveness of artefact cleansing can be confirmed if the cleaned data are split again in single trials, sorted according to the stimulated finger, and averaged as before (see figure 5.17). Comparing the latter figure with figure 5.11, we can now clearly see the peaks which are due to the electrical stimulation. In addition, in figure 5.18 we compare two channels, which are heavily contaminated by the ocular and the cardiac interference respectively, before and after artefact removal. We notice that ICA has achieved its goal in removing the artefact signals since the cardiac spikes and

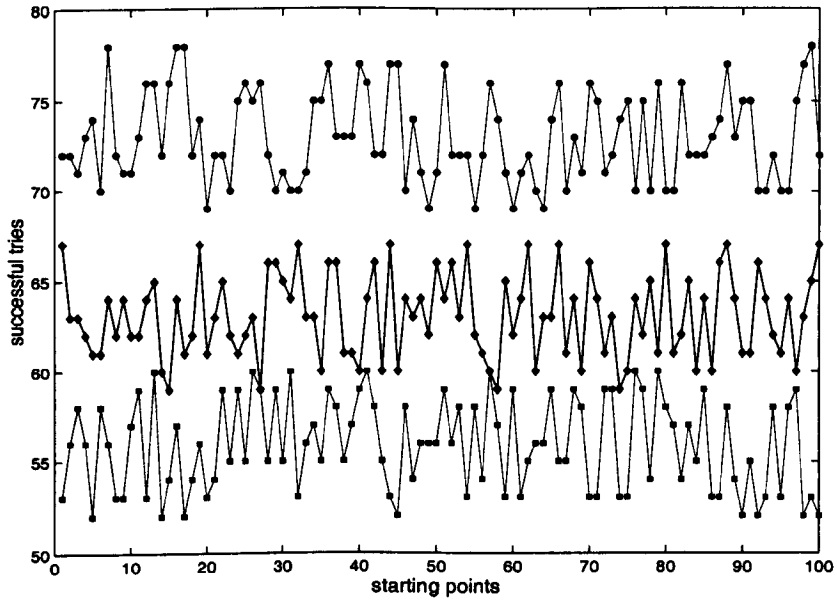


Figure 5.15: Successful tries in extracting the desired ocular component using cICA for each starting point for different values of iterations Q ($\lambda = 1000$, $T_0 = 10$). The extraction of the desired contaminating component can be improved if we increase the number of iterations allowed in the termination criterion of the algorithm. Upper line: 8000 iterations, middle line: 4000 iterations, lower line: 2000 iterations.

the ocular bumps have been eliminated to a great extent.

5.5 Conclusions

In this chapter we have introduced a modification of the standard ICA algorithm in order to cope with the intrinsic ambiguity of ICA in the extraction order of the independent components. We have shown that in case we have prior knowledge concerning one of the original signals, we can exploit that information by adding a penalty/constraint term to the standard ICA quality function in order to favour the extraction of that particular signal. The success of our method depends strongly on the selection of the parameter values of optimisation method that is employed. Our quality function was successfully optimised with both simplex and simulated annealing. The former performs well for small scale problems such as in the artificially generated data we used, whereas simulated annealing is indicated for multidimensional environments such as in our real MEG data. On the other hand, steepest ascent turns out to be very sensitive.

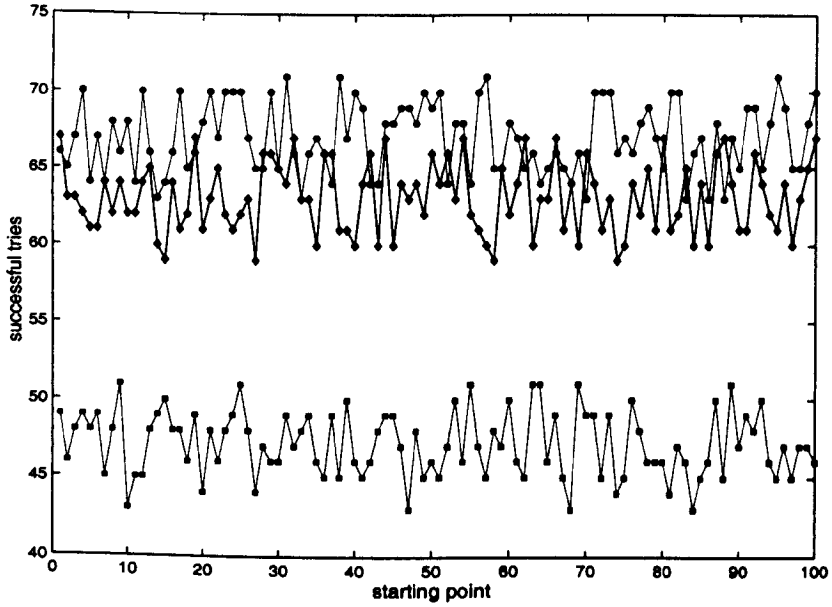
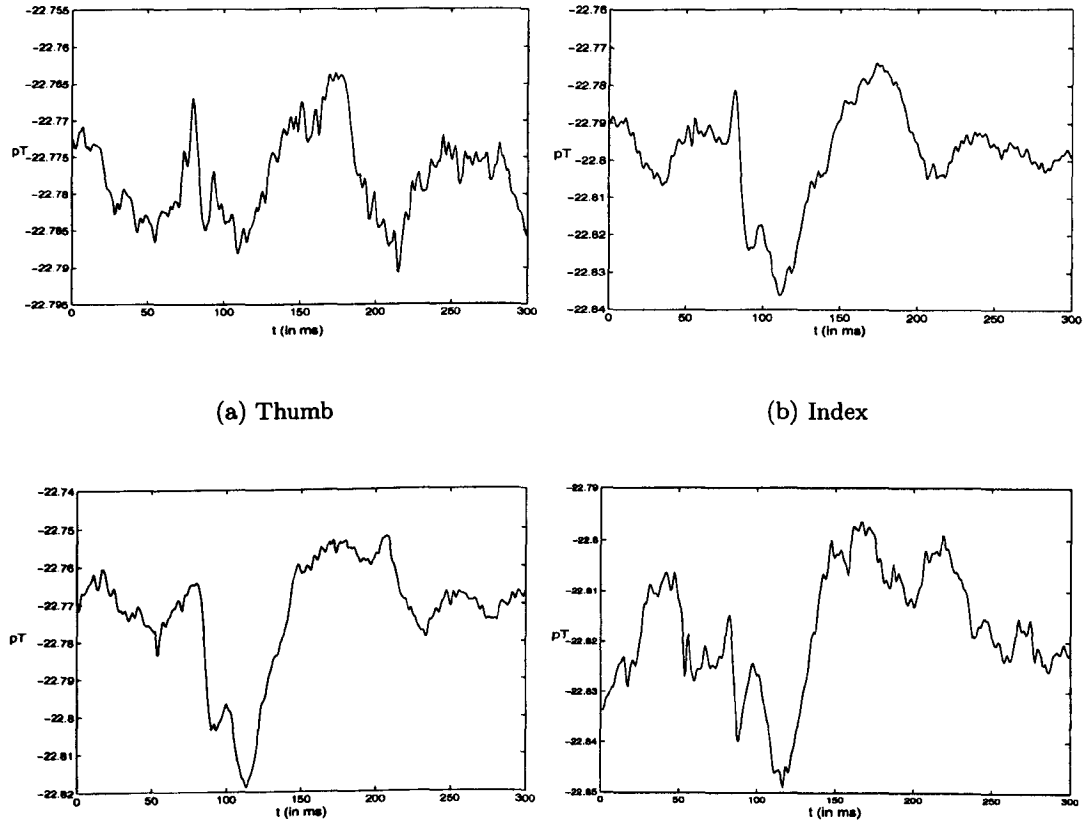


Figure 5.16: Successful tries in extracting the desired component for each starting point for different values of the initial temperature T_0 ($\lambda = 1000$, $Q = 4000$). The extraction of the desired contaminating component can be further improved if we increase the initial temperature T_0. Upper line: $T_0 = 100$, middle line: $T_0 = 10$, lower line: $T_0 = 1$.

By definition ICA uses all the available data points to extract the independent components, and therefore the task becomes computationally intensive in multichannel experiments. The importance of using *a priori* information in ICA becomes more significant in the real MEG data due to their vast size. In our MEG data we managed to employ constrained ICA successfully in identifying and removing the artefact fields. The signal of interest can then be extracted clean from any interference.

Compared with previous efforts in the so-called constrained ICA field, our method is not really based on an actual reference signal, but on the knowledge of some statistical properties of the desired independent component, e.g. the autocorrelation function. In addition there is no need to specify an optimum time lag. In fact, any statistical property about an original source can be used with the proper changes of the quality function. However, on the other hand our technique uses a significantly higher amount of data because all the cross-correlations between the different channels should be computed for all possible time lags.

The algorithm can be repeated for as many times as the number of independent components whose prior information is known. Then we can extract the remaining components by



(a) Thumb

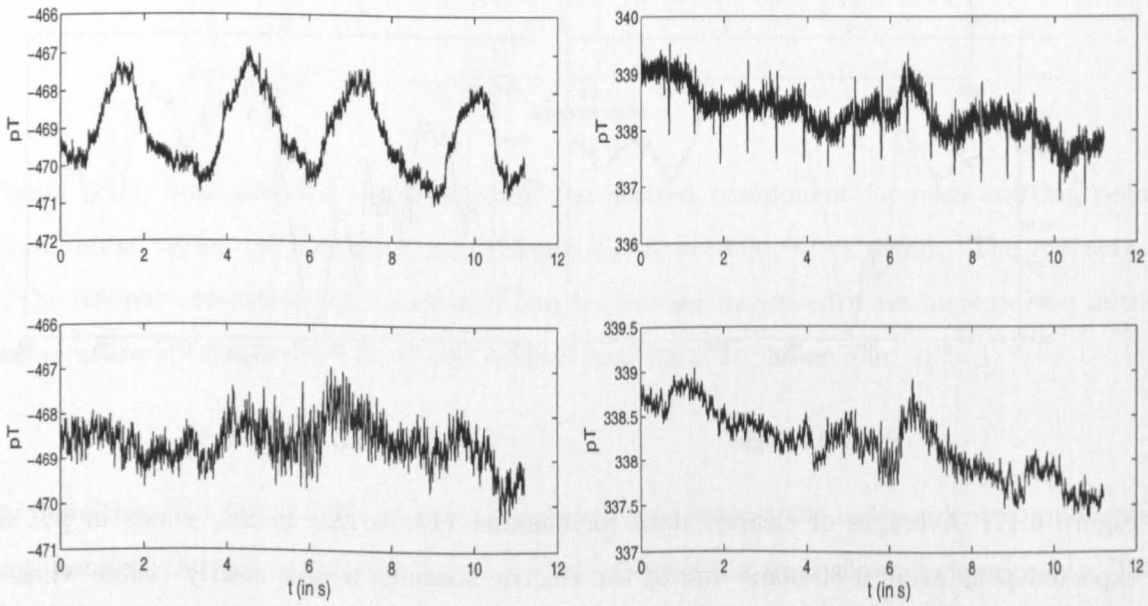
(b) Index

(c) Middle finger

(d) Little finger

Figure 5.17: Averages of cleaned data for channel 117. x-axis in ms, y-axis in pT. The expected peak around 80-90ms due to the electric stimulus is now clearly visible when the ocular and cardiac interference have been eliminated using cICA.

returning to the ordinary ICA algorithm.



(a) Channel 26

(b) Channel 40

Figure 5.18: Channels 26 and 40 - Before and after artefact removal using cICA. ICA succeeds in cleaning the recordings from the artefact signals.

Chapter 6

Independent Component Analysis in Source Localisation

This chapter completes the presentation of ICA in biomagnetic studies. In particular, it deals with the use of ICA in source localisation issues. First, section 6.1 introduces in brief the most commonly used head models. The selection of a proper model is crucial in order to achieve the desired accuracy in source localisation. Section 6.2 computes the forward problem for the simple spherical homogeneous head model. The importance of ICA in simplifying the ill-posed inverse problem is explained in section 6.3. Section 6.4 formulates the cost function for the inverse problem in our MEG study. Finally, section 6.5 demonstrates the practical use of ICA in source localisation for simulated and real MEG data.

6.1 Head Models

In order to solve the inverse problem, we have first to assume a model about the head. In general, there are three families of head models. The choice depends on the information which may be available from additional structural examinations, such as from detailed MRI scans. Moreover, the selection of a particular head model determines the computational complexity in tackling the inverse problem.

6.1.1 Spherical Homogeneous Model

The head is often modelled as a uniform conducting sphere [34]. The assumptions made in this model are the following:

1. Radial current dipoles have no measurable external magnetic fields.
2. Volume currents do not contribute to the magnetic field of interest.
3. Field recordings, dipole location and depth are linked through simple equations.

This particular model is extremely fast. The forward problem can be solved analytically without any information about the conductivity profile [131, 139]. Therefore, it can be employed when there is minimal knowledge about the anatomical and structural details of the participant. However, the main drawback is that this model does not represent well the temporal lobe [137].

6.1.2 Boundary Element Models

Localisation of biomagnetic sources is improved when a more realistic volume-conductor model is used for the head. This is approximated by compartments of isotropic and homogeneous conductivities, and is known as *boundary element model* (BEM) [33, 41]. This compartmental model does not actually need absolute conductivity values, but the ratio of them. BEM is limited only by the small number of compartments which are used. Typical layers used in BEM analysis are the scalp, and the outer and inner skull surface. In addition, this spherical model is extremely fast to compute.

Alternatively, there is a sensor-weighted overlapping-sphere head model proposed by [61]. It consists of multiple overlapping spheres on a sensor-by-sensor basis. It has almost the same computational cost as that of the single-sphere model, and similar accuracy to BEM for most regions of the brain at greatly reduced computational cost and complexity.

6.1.3 Finite Element Models

While BEM is a compromise between oversimplified spherically symmetric models and the real structure of the tissue, the *finite element model* (FEM) takes into account the inhomogeneities being present in practice within the human head, and the different tissue types [55]. Research

revealed that the magnetic fields are most sensitive to tissue resistivity changes very close to the source position (in particular, changes in the gray matter resistivity since the dipoles are located in the gray matter). That explains alterations in MEG patterns of patients with structural changes in the cortex. Unfortunately, the computational time for FEM is too high. Moreover, it is very difficult to determine the volume fine structure parameters.

6.2 Forward Problem

Unfortunately, our MEG data are not accompanied by an MRI scan. Therefore, in practice we have to limit ourselves in the use of a simple spherical homogeneous model. Let us consider a sensor C placed on the surface of a sphere of radius r (see figure 6.1). Denote by ω_1 the sensor azimuthal angle in the xy-plane from the x-axis with $0 \leq \omega_1 < 2\pi$, and by ω_2 the polar angle from the xy-plane with $-\frac{\pi}{2} \leq \omega_2 \leq \frac{\pi}{2}$. The Cartesian coordinates (c_x,c_y,c_z) of the sensor are given by

$$c_x = r\cos\omega_1\cos\omega_2 \tag{6.1}$$

$$c_y = r\sin\omega_1\cos\omega_2 \tag{6.2}$$

$$c_z = r\sin\omega_2 \tag{6.3}$$

Assume that a dipole S is lying within the sphere of radius r. The dipole produces a magnetic field which is recorded by the sensor. In 3D space, the dipole can be fully described by a set of six parameters: position (s_x,s_y,s_z), strength Q, and orientation (ϕ_1,ϕ_2), where ϕ_1 is the azimuthal angle of the dipole vector in the xy-plane from the x-axis with $0 \leq \phi_1 < 2\pi$, and ϕ_2 the polar angle of the dipole vector from the xy-plane with $-\frac{\pi}{2} \leq \phi_2 \leq \frac{\pi}{2}$. If r_q, θ_1 and θ_2 are the dipole position spherical coordinates, the dipole position in Cartesian coordinates is given by

$$s_x = r_q\cos\theta_1\cos\theta_2 \tag{6.4}$$

$$s_y = r_q\sin\theta_1\cos\theta_2 \tag{6.5}$$

$$s_y = r_q\sin\theta_2 \tag{6.6}$$

where $r_q \leq r$, $0 \leq \theta_1 < 2\pi$, and $-\frac{\pi}{2} \leq \theta_2 \leq \frac{\pi}{2}$.

Orientation (ϕ_1,ϕ_2) and dipole strength Q define the dipole vector $\mathbf{Q}$ (see figure 6.2). They can be calculated if we know alternatively the dipole moments in the x-, y- and z-direction,

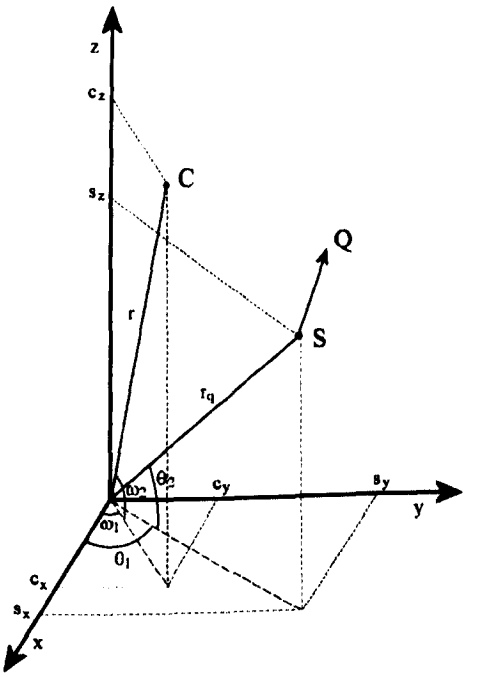


Figure 6.1: Geometry of sensor C and dipole S in 3D space.

say Q_x, Q_y and Q_z respectively:

$$Q_x = Q\cos\phi_1\cos\phi_2 \tag{6.7}$$

$$Q_y = Q\sin\phi_1\cos\phi_2 \tag{6.8}$$

$$Q_z = Q\sin\phi_2 \tag{6.9}$$

Thus

$$\tan\phi_1 = \frac{Q_y}{Q_x} \Rightarrow \phi_1 = \arctan\frac{Q_y}{Q_x} \tag{6.10}$$

and

$$\tan\phi_2 = \frac{Q_z}{Q_x}\sin\phi_1 \Rightarrow \phi_2 = \arctan\frac{Q_z\sin\phi_1}{Q_x} \tag{6.11}$$

If ϕ_1 and ϕ_2 are computed, then it is easy to calculate Q. In consequence, in 3D space we should know two triplets of parameters to describe the dipole: (s_x,s_y,s_z) for dipole position, and (Q_x,Q_y,Q_z) for dipole vector.

The magnetic field $\mathbf{B}(\mathbf{r})$ produced by dipole S with moment $\mathbf{Q}$ is calculated by the Biot-Savart law:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left[\mathbf{Q}\times\frac{(\mathbf{r}-\mathbf{r}_q)}{|\mathbf{r}-\mathbf{r}_q|^3} - \sum\sigma_j\int_{G_j}\nabla\phi\times\frac{(\mathbf{r}-\mathbf{r}_q)}{|\mathbf{r}-\mathbf{r}_q|^3}dv\right] \tag{6.12}$$

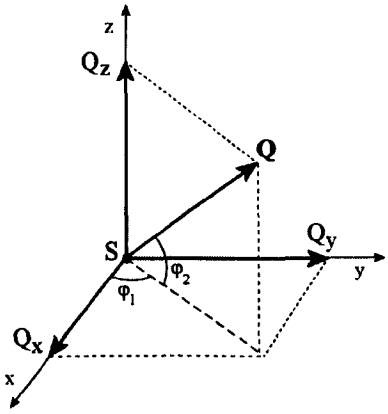


Figure 6.2: Dipole vector **Q** in 3D space.

where $\mu_0 = 4\pi 10^{-7}$ is the magnetic permeability of the vacuum, $\mathbf{r}$ is the point of measurement, and $\mathbf{r}_q$ is the dipole position vector. G_j indicates sub-volumes with different electrical conductivities σ_j, and ϕ is the electric potential.

For a spherical homogeneous medium, equation (6.12) can be expressed in an analytic closed form. According to [139], the magnetic field $\mathbf{B}(\mathbf{r})$, which is produced by dipole S enclosed in a homogeneous sphere at sensor point C outside the conductor, is calculated as follows (also known as the Sarvas formula):

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi F^2}\left[F\mathbf{Q}\times\mathbf{r}_q - (\mathbf{Q}\times\mathbf{r}_q\cdot\mathbf{r})\nabla F\right] \tag{6.13}$$

where the scalar function $F(\mathbf{r},\mathbf{r}_q)$ and the vector function $\nabla F(\mathbf{r},\mathbf{r}_q)$ are

$$F(\mathbf{r},\mathbf{r}_q) \equiv |\mathbf{d}|\left(|\mathbf{r}||\mathbf{d}| + |\mathbf{r}|^2 - \mathbf{r}_q\cdot\mathbf{r}\right) \tag{6.14}$$

and

$$\nabla F(\mathbf{r},\mathbf{r}_q) \equiv \left(\frac{|\mathbf{d}|^2}{|\mathbf{r}|} + \frac{\mathbf{d}\cdot\mathbf{r}}{|\mathbf{d}|} + 2|\mathbf{d}| + 2|\mathbf{r}|\right)\mathbf{r} - \left(\frac{\mathbf{d}\cdot\mathbf{r}}{|\mathbf{d}|} + |\mathbf{d}| + 2|\mathbf{r}|\right)\mathbf{r}_q \tag{6.15}$$

where

$$\mathbf{d} \equiv \mathbf{r} - \mathbf{r}_q \tag{6.16}$$

The Euclidean distance $|\mathbf{d}|$ between sensor C and dipole S is given by

$$|\mathbf{d}| = \left((c_x - s_x)^2 + (c_y - s_y)^2 + (c_z - s_z)^2\right)^{1/2} \tag{6.17}$$

Sensor C records only the component of the magnetic field normal to the sensor. In general, sensor C is not radially oriented. Denote by ζ_1 the azimuthal angle of the vector normal

to the sensor coil in the xy-plane from the x-axis with $0 \leq \zeta_1 < 2\pi$, and by ζ_2 the normal vector polar angle from the xy-plane with $-\frac{\pi}{2} \leq \zeta_2 \leq \frac{\pi}{2}$. Then the unit orientation vector $\mathbf{R}$ (normal vector) of the sensor coil is

$$\mathbf{R} = R_x\hat{\imath} + R_y\hat{\jmath} + R_z\hat{k} \tag{6.18}$$

where

$$R_x \equiv \cos\zeta_1 \cos\zeta_2 \tag{6.19}$$

$$R_y \equiv \sin\zeta_1 \cos\zeta_2 \tag{6.20}$$

$$R_z \equiv \sin\zeta_2 \tag{6.21}$$

Therefore, sensor C records the scalar projection X of $\mathbf{B}(\mathbf{r})$ onto the normal direction

$$X = \frac{\mu_0}{4\pi F^2}\left[F\mathbf{Q} \times \mathbf{r}_q - (\mathbf{Q} \times \mathbf{r}_q \cdot \mathbf{r})\nabla F\right] \cdot \mathbf{R} \tag{6.22}$$

In general, consider N dipoles S_i $(i = 1, 2, \ldots, N)$ and M sensors C_j $(j = 1, 2, \ldots, M)$. Then the signal X_{ji} recorded by sensor C_j due to dipole S_i is given by

$$X_{ji} = \frac{\mu_0}{4\pi F_{ji}^2}\left[F_{ji}\mathbf{Q}_i \times \mathbf{r}_{q_i} - (\mathbf{Q}_i \times \mathbf{r}_{q_i} \cdot \mathbf{r}_j)\nabla F_{ji}\right] \cdot \mathbf{R}_j \tag{6.23}$$

Sensor C_j records the contributions of all dipoles S_i

$$X_j = \sum_{i=1}^{N} X_{ji} = \sum_{i=1}^{N}\left\{\frac{\mu_0}{4\pi F_{ji}^2}\left[F_{ji}\mathbf{Q}_i \times \mathbf{r}_{q_i} - (\mathbf{Q}_i \times \mathbf{r}_{q_i} \cdot \mathbf{r}_j)\nabla F_{ji}\right] \cdot \mathbf{R}_j\right\} \tag{6.24}$$

The only assumption about the dipoles S_i is that their position $(s_{x_i}, s_{y_i}, s_{z_i})$, and their orientations (ϕ_{1_i}, ϕ_{2_i}) remain constant over time, whereas their strengths Q_i may vary with time. In fact, Q_i may even be zero when dipole S_i is not activated at that time slice.

For a particular time slice k we may write

$$X_j^k = \sum_{i=1}^{N} X_{ji}^k = \sum_{i=1}^{N}\left\{\frac{\mu_0}{4\pi F_{ji}^2}\left[F_{ji}\mathbf{Q}_i^k \times \mathbf{r}_{q_i} - (\mathbf{Q}_i^k \times \mathbf{r}_{q_i} \cdot \mathbf{r}_j)\nabla F_{ji}\right] \cdot \mathbf{R}_j\right\} \tag{6.25}$$

where Q_i^k is the dipole strength at the k^{th} time slice.

Equation (6.25) allows the calculation of sensor recordings produced by a configuration of multiple dipoles, and solves in an analytic closed form the *forward problem* for a spherical head model. This is the essential first step towards tackling the *inverse problem*. The forward problem is presented thoroughly for a wide range of head models in [114].

6.3 ICA and the assumption of a single dipole

The mainstream approach in source localisation is to consider simultaneously multiple dipoles and change their position, orientation, and strength [87]. For each different configuration, the forward solution is computed and compared with the actual sensor recordings. The configuration which yields the minimum error is considered to be the actual one. By its nature, this is an ill-posed problem which can be eased by using anatomical and physiological constraints to limit the number of possible solutions.

Recently ICA has been applied in simulated EEG studies to simplify the task of source localisation in a significant way [163]. According to the ICA data mixing model, the sensor recordings X_j $(j = 1, 2, \ldots, M)$ are considered to be linear mixtures of N unknown, statistically independent signals s_i $(i = 1, 2, \ldots, N)$. The problem is decomposed in N separate independent components. Each of them is considered to be produced by a unique dipole. Therefore, instead of working with N dipoles simultaneously, we only have to localise a single dipole for each independent component. The straightforward procedure of localising a single dipole is then repeated for all independent components.

For simplicity, from now and on we drop index i which was used to identify the dipole. The activation map produced by a single dipole S can be written analytically for each sensor C_j at the k^{th} time slice

$$X_j^k = \frac{\mu_0}{4\pi F_j^2}\left[F_j \mathbf{Q}^k \times \mathbf{r}_q - (\mathbf{Q}^k \times \mathbf{r}_q \cdot \mathbf{r}_j)\nabla F_j\right] \cdot \mathbf{R}_j \tag{6.26}$$

Using vector notation, the activation map produced by a single dipole S at the k^{th} time slice can be denoted by a vector $\mathbf{X}^k = [X_1^k \; X_2^K \; \ldots \; X_M^k]^T$.

6.4 Inverse Problem

Our motivation is to pinpoint the single dipole S which is associated with a particular independent component s_l and produces the respective activation map. To solve the inverse problem we should estimate the six dipole parameters that minimize the least-square function:

$$\Psi(s_x, s_y, s_z, Q_x, Q_y, Q_z) = \left(\mathbf{X}^{\text{fwd}} - \mathbf{X}^{\text{meas}}\right)^2 \tag{6.27}$$

where $\mathbf{X}^{\text{fwd}}$ is the activation map for a given dipole (Q_x,Q_y,Q_z) in a given location (s_x,s_y,s_z), ie $\mathbf{X}^{\text{fwd}} = \mathbf{X}^{\text{fwd}}(s_x, s_y, s_z, Q_x, Q_y, Q_z)$, and $\mathbf{X}^{\text{meas}}$ is the actual, recorded activation map due

to source S.

In our study, the activation map $\mathbf{X}^{\text{meas}}$ produced by a single dipole is provided by ICA. Since ICA estimates the $(M \times N)$ mixing matrix A, and thus the independent components s_i $(i = 1, 2, \ldots, N)$, the contribution of a particular independent component s_l to the sensor recordings can be computed. By zeroing all other components s_i (for $i = 1, 2, \ldots, l-1, l+1, \ldots, N$) except for the one s_l in which we are interested, we can have the estimated activation map $\mathbf{X}^{\text{meas}}$ due to that component s_l:

$$\mathbf{X}^{\text{meas}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1l} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2l} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Ml} & \cdots & a_{MN} \end{pmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ s_l \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{1l}s_l \\ a_{2l}s_l \\ \vdots \\ a_{Ml}s_l \end{bmatrix} \tag{6.28}$$

The process of solving the inverse problem does not require the whole time series. Since the position and the orientation of each dipole is considered to be constant over time, a single time slice, which is chosen at random, is enough. The only requirement is that the dipole, which we attempt to localise, is active during that time slice. Therefore, for simplicity we dropped index k which was used to identify the time slice. In consequence, equation (6.26) can be rewritten as

$$X_j = \frac{\mu_0}{4\pi F_j^2} [F_j \mathbf{Q} \times \mathbf{r}_q - (\mathbf{Q} \times \mathbf{r}_q \cdot \mathbf{r}_j) \nabla F_j] \cdot \mathbf{R}_j \tag{6.29}$$

Now we will show that the recorded signal X_j can be expressed as a linear combination of the dipole moments Q_x, Q_y and Q_z.

Indeed, we have

$$\mathbf{Q} \times \mathbf{r}_q = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ Q_x & Q_y & Q_z \\ s_x & s_y & s_z \end{vmatrix} = (s_z Q_y - s_y Q_z)\hat{\imath} + (s_x Q_z - s_z Q_x)\hat{\jmath} + (s_y Q_x - s_x Q_y)\hat{k} \tag{6.30}$$

and

$$\mathbf{Q} \times \mathbf{r}_q \cdot \mathbf{r}_j = (s_z Q_y - s_y Q_z) c_{xj} + (s_x Q_z - s_z Q_x) c_{yj} + (s_y Q_x - s_x Q_y) c_{zj} \tag{6.31}$$

Say $\nabla F_j = \Lambda_{1j}\hat{\imath} + \Lambda_{2j}\hat{\jmath} + \Lambda_{3j}\hat{k}$. Then

$$(\mathbf{Q} \times \mathbf{r}_q \cdot \mathbf{r}_j) \nabla F_j = A_{xj}\hat{\imath} + A_{yj}\hat{\jmath} + A_{zj}\hat{k} \tag{6.32}$$

where

$$A_{xj} \equiv \left[(s_z Q_y - s_y Q_z)c_{xj} + (s_x Q_z - s_z Q_x)c_{yj} + (s_y Q_x - s_x Q_y)c_{zj}\right] \Lambda_{1j} \tag{6.33}$$

$$A_{yj} \equiv \left[(s_z Q_y - s_y Q_z)c_{xj} + (s_x Q_z - s_z Q_x)c_{yj} + (s_y Q_x - s_x Q_y)c_{zj}\right] \Lambda_{2j} \tag{6.34}$$

$$A_{zj} \equiv \left[(s_z Q_y - s_y Q_z)c_{xj} + (s_x Q_z - s_z Q_x)c_{yj} + (s_y Q_x - s_x Q_y)c_{zj}\right] \Lambda_{3j} \tag{6.35}$$

Using (6.18),(6.30),(6.32), equation (6.29) can be written as

$$X_j = \beta_{xj} Q_x + \beta_{yj} Q_y + \beta_{zj} Q_z \tag{6.36}$$

where

$$\beta_{xj} \equiv \frac{\mu_0}{4\pi F_j^2}\left[F_j R_{zj} s_y - F_j R_{yj} s_z + (\Lambda_{1j} R_{xj} + \Lambda_{2j} R_{yj} + \Lambda_{3j} R_{zj})(s_z c_{yj} - s_y c_{zj})\right] \tag{6.37}$$

$$\beta_{yj} \equiv \frac{\mu_0}{4\pi F_j^2}\left[F_j R_{xj} s_z - F_j R_{zj} s_x + (\Lambda_{1j} R_{xj} + \Lambda_{2j} R_{yj} + \Lambda_{3j} R_{zj})(s_x c_{zj} - s_z c_{xj})\right] \tag{6.38}$$

$$\beta_{zj} \equiv \frac{\mu_0}{4\pi F_j^2}\left[F_j R_{yj} s_x - F_j R_{xj} s_y + (\Lambda_{1j} R_{xj} + \Lambda_2 R_{yj} + \Lambda_{3j} R_{zj})(s_y c_{xj} - s_x c_{yj})\right] \tag{6.39}$$

Parameters β_{xj}, β_{yj} and β_{zj} are functions of the dipole location (s_x,s_y,s_z). From equation (6.36) it is clear that X_j is a linear function of Q_x, Q_y and Q_z.

For any given dipole location (s_x,s_y,s_z), the optimal dipole moments $\bar{Q}_x$, $\bar{Q}_y$, and $\bar{Q}_z$ can be found in closed form.

Using vector-matrix notation, the activation map produced by a single dipole can be written as

$$\mathbf{X} = \beta_{\mathbf{x}} Q_x + \beta_{\mathbf{y}} Q_y + \beta_{\mathbf{z}} Q_z \tag{6.40}$$

where $\mathbf{X} = [X_1\ X_2\ \ldots\ X_M]^T$, $\beta_{\mathbf{x}} = [\beta_{x1}\ \beta_{x2}\ \ldots\ \beta_{xM}]^T$, $\beta_{\mathbf{y}} = [\beta_{y1}\ \beta_{y2}\ \ldots\ \beta_{yM}]^T$ and $\beta_{\mathbf{z}} = [\beta_{z1}\ \beta_{z2}\ \ldots\ \beta_{zM}]^T$

Equation (6.27) can be written as

$$\Psi(s_x, s_y, s_z, Q_x, Q_y, Q_z) = (\beta_{\mathbf{x}} Q_x + \beta_{\mathbf{y}} Q_y + \beta_{\mathbf{z}} Q_z - \mathbf{X}^{\text{meas}})^2 \tag{6.41}$$

The quadratic function (6.41) should be minimized, to obtain the least square error solution. We set its partial derivatives with respect to Q_x, Q_y and Q_z to zero:

$$\frac{\partial \Psi}{\partial Q_x} = 0 \Rightarrow 2\sum_{j=1}^{M}\left[\beta_{xj}\left(\beta_{xj} Q_x + \beta_{yj} Q_y + \beta_{zj} Q_z - X_j^{\text{meas}}\right)\right] = 0 \Rightarrow$$

$$\sum_{j=1}^{M}\left(\beta_{x_j}\beta_{x_j}Q_x+\beta_{x_j}\beta_{y_j}Q_y+\beta_{x_j}\beta_{z_j}Q_z-\beta_{x_j}X_j^{\text{meas}}\right)=0\Rightarrow$$

$$\left[\sum_{j=1}^{M}\left(\beta_{x_j}\beta_{x_j}\right)\right]Q_x+\left[\sum_{j=1}^{M}\left(\beta_{x_j}\beta_{y_j}\right)\right]Q_y+\left[\sum_{j=1}^{M}\left(\beta_{x_j}\beta_{z_j}\right)\right]Q_z=\sum_{j=1}^{M}\left(\beta_{x_j}X_j^{\text{meas}}\right)\Rightarrow$$

$$<\beta_{\mathbf{x}},\beta_{\mathbf{x}}>Q_x+<\beta_{\mathbf{x}},\beta_{\mathbf{y}}>Q_y+<\beta_{\mathbf{x}},\beta_{\mathbf{z}}>Q_z=<\beta_{\mathbf{x}},\mathbf{X}^{\text{meas}}> \tag{6.42}$$

where $<*,*>$ denotes inner product.

Similarly,

$$\frac{\partial\Psi}{\partial Q_y}=0\Rightarrow<\beta_{\mathbf{x}},\beta_{\mathbf{y}}>Q_x+<\beta_{\mathbf{y}},\beta_{\mathbf{y}}>Q_y+<\beta_{\mathbf{y}},\beta_{\mathbf{z}}>Q_z=<\beta_{\mathbf{y}},\mathbf{X}^{\text{meas}}> \tag{6.43}$$

and

$$\frac{\partial\Psi}{\partial Q_z}=0\Rightarrow<\beta_{\mathbf{x}},\beta_{\mathbf{z}}>Q_x+<\beta_{\mathbf{y}},\beta_{\mathbf{z}}>Q_y+<\beta_{\mathbf{z}},\beta_{\mathbf{z}}>Q_z=<\beta_{\mathbf{z}},\mathbf{X}^{\text{meas}}> \tag{6.44}$$

Equations (6.42), (6.43) and (6.44) constitute a (3 × 3) system of equations.

$$\begin{pmatrix}<\beta_x,\beta_x> & <\beta_x,\beta_y> & <\beta_x,\beta_z>\\ <\beta_x,\beta_y> & <\beta_y,\beta_y> & <\beta_y,\beta_z>\\ <\beta_x,\beta_z> & <\beta_y,\beta_z> & <\beta_z,\beta_z>\end{pmatrix}\begin{pmatrix}Q_x\\Q_y\\Q_z\end{pmatrix}=\begin{pmatrix}<\beta_x,X^{\text{meas}}>\\<\beta_y,X^{\text{meas}}>\\<\beta_z,X^{\text{meas}}>\end{pmatrix}$$

Using vector-matrix notation the system of equations can be rewritten as

$$B_1\mathbf{Q}=B_2 \tag{6.45}$$

where $B_1=\begin{pmatrix}<\beta_x,\beta_x> & <\beta_x,\beta_y> & <\beta_x,\beta_z>\\ <\beta_x,\beta_y> & <\beta_y,\beta_y> & <\beta_y,\beta_z>\\ <\beta_x,\beta_z> & <\beta_y,\beta_z> & <\beta_z,\beta_z>\end{pmatrix}$, and $B_2=\begin{pmatrix}<\beta_x,X^{\text{meas}}>\\<\beta_y,X^{\text{meas}}>\\<\beta_z,X^{\text{meas}}>\end{pmatrix}$

Its solution $(\bar{Q}_x,\bar{Q}_y,\bar{Q}_z)$ provides the optimal dipole moments for a given position (s_x,s_y,s_z). However, we can prove that matrix B_1 is not invertible ($\det(B_1)=0$) and $\text{rank}(B_1)=2$. Therefore, we have two main unknowns, say Q_x, Q_y, which can be expressed through a side unknown Q_z. In consequence, for a given position (s_x,s_y,s_z) there are infinite sets of optimal dipole moments $(\bar{Q}_x,\bar{Q}_y,\bar{Q}_z)$ for which the quadratic function (6.41) is minimized.

The estimation of the optimal dipole moments for any possible location in 3D space reduces function Ψ to a least-square error cost function Ω which is explicitly only a function of s_x,

s_y and s_z:

$$\Omega(s_x, s_y, s_z) \equiv \sum_{j=1}^{M} \left(\beta_{xj}\bar{Q}_x + \beta_{yj}\bar{Q}_y + \beta_{zj}\bar{Q}_z - X^{\text{meas}}\right)^2 \tag{6.46}$$

Equation 6.46 provides the cost function for the inverse problem of a single dipole based on the adoption of a spherical homogeneous head model. Next, we will demonstrate the ease in source localisation which is offered by ICA with simulated data.

6.5 Experimental Results

6.5.1 Simulated Data

Assume that we have three dipoles lying within the skull, acting as sources of the recorded magnetic fields. The dipoles are considered to have constant orientation and position. Their strength over time is presented in figure 4.3(a) in p.57. The dipole coordinates are given in table 6.1. The magnetic field produced by each dipole is calculated with equation 6.13. The fields are measured by 151 sensors placed outside the skull according to equation 6.23. The geometrical characteristics of the sensors are provided by the manufacturer of the MEG scanner [32]. Each sensor records the contribution of all three dipoles using equation 6.24. Therefore, the recordings are linear mixtures of the contribution of each source.

dipole	s_x (cm)	s_y (cm)	s_z (cm)
1st	4.45	4.68	-4.14
2nd	-2.82	-0.06	-4.71
3rd	-1.27	-0.96	-0.81

Table 6.1: Cartesian coordinates of sources

Then we apply ICA to the recordings in order to restore the original signals. Since we are not now interested in a particular order of extraction, we perform ordinary ICA. The original sources are accurately estimated as three independent components. Each independent component is assumed to be produced by a different dipole, and used as the measured activation map $\mathbf{X}^{\text{meas}}$ in cost function Ω of equation 6.46.

Our aim is to localise the sources in space. The procedure is straightforward. We minimise the cost function Ω using the simplex method which was explained in detail in section 5.3.2.

In order to avoid potential local minima, we can use several different points inside the skull as starting points. The coordinates of the estimated dipoles are given in table 6.2. The Euclidean distance between the estimated dipoles and the corresponding original sources is 0.06cm, 0.16cm, and 0.41cm respectively.

In general, the success or failure of this variation of source localisation depends on the appropriateness of the head model which is employed, and on the efficiency of ICA in estimating the mixing matrix A. However, working with simulated data we actually remove the dependence on the head model since we know in advance the exact mechanism which generates, propagates and mixes the magnetic fields of the original sources in the MEG sensors. In effect, ICA estimates the elements of the mixing matrix A which are linked with the dipole coordinates through equation 6.23. Therefore, the accuracy of source localisation using ICA with simulated data is actually an exclusive measure of the efficiency of ICA in separating the original sources.

Note also that since A is considered to be constant over time, there is no need to use the whole time series in order to solve the inverse problem. Any single time slice, chosen at random, yields identical results.

dipole	s_x (cm)	s_y (cm)	s_z (cm)
1st	4.50	4.71	-4.14
2nd	-2.84	-0.08	-4.86
3rd	-1.40	-0.88	-0.44

Table 6.2: Cartesian coordinates of estimated dipoles.

6.5.2 Real MEG Data

The extracranial magnetic fields which are produced due to the electrical stimulus are recorded with the 151-channel MEG system. The biological signals of interest are generated at the somatosensory cortex. Hence, they are expected to be localised in the right parietal (RP) area of the brain [13] (see figure 5.6). In the previous chapter we removed the two major sources of interference (namely cardiac contamination and eye blinking) from the data by applying cICA.

Now our task is to pinpoint the area in the brain which produces the signal of interest for

different fingers. The basic assumption is that different stimulated fingers have magnetic signals which are produced by dipoles in slightly different positions. Therefore, if we apply ICA in trials where the stimulated finger is different, we expect to find one or more independent components which are localised in close but distinct points of the somatosensory cortex. According to [13], the dipoles should be located within a few millimeters.

In order to have a sufficient number of samples to perform ICA, we concatenate the cleaned trials according to the stimulated finger. For each finger, we concatenate 50 trials (i.e. we use 18750 samples for ICA). Thus for the thumb and index, we have 8 new smaller subdatasets, and for the middle and little finger, 7 new subdatasets. Data concatenation is a perfectly valid concept if we take into account the nature of ICA as a statistical technique which uses all the available information in order to estimate the independent components without being actually interested in the order of time slices.

Then we perform ICA extracting 8 independent components for each subdataset of each finger. Note that the dimensionality of the data was reduced to 10 in the previous chapter by applying an EVD. Moreover, two independent components corresponding to the artefact signals were eliminated from the recordings. In consequence, we expect to have a maximum number of eight independent components left in our datasets.

Next we try to find which independent component of each subdataset can be localised in the area of interest. For example, let us examine one of the subdatasets of the index. We could use the simplex method to minimise the cost function 6.46 as we did before with the simulated data. However, since we know approximately where to search, and in order to avoid errors in the minimisation due to the simplex sensitivity, we decide to perform an exhaustive search in the RP area with a small step of 1mm in each direction of the 3D space. Recall that this a single time slice analysis, and the only requirement is that the dipole, which is associated with the biological signal of interest, is active during that time slice. Therefore, we perform our study for a time slice 30-40ms after the stimulus where the peak is expected, say 35ms. Unfortunately, none of the independent components can be localised in the area of interest. Consequently, we decide to expand our search field to the whole head using the simplex method in order to minimise the cost function Ω using several different starting points inside the skull. In table 6.3 we can see for each independent component of this subdataset which MEG channel is closer to the estimated source point. Similar results are obtained for all subdatasets for all fingers. Moreover, let us reverse the concatenation

process and split the estimated independent components for each subdataset to single trials. Then in order to enhance the quality of the biological signal of interest we perform averaging of the independent components for all trials of the same subdataset, and repeat the above procedure of source localisation. Unfortunately, we get more negative results in a similar way.

comp	channel	comp	channel
1	122	5	139
2	139	6	139
3	139	7	68
4	139	8	68

Table 6.3: Trials are concatenated according to the stimulated finger forming smaller datasets. ICA is applied to one of them, namely to a subdataset of the index, and all independent components are localised in 3D space. Each component is associated with an MEG channel which is the closest to the estimated source point of that component. We notice that none of the components can be localised in the area of interest.

The most probable cause of this catastrophic failure of ICA in real MEG data has to do with the choice of the head model. The solution of the inverse problem depends strongly on the forward model which is adopted. Therefore, although the spherical homogeneous model in conjunction with the Sarvas formula 6.13 has been widely used in phantom studies, it may not be realistic enough for our real MEG data. Unfortunately the lack of an MRI scan narrows the choices of a proper model. Realistic head models require detailed information about the changes of conductivity in the conducting medium.

In addition, ICA may have failed to extract the weak biological signal of interest in a single separate component. In this case, undetected traces of this signal are scattered throughout the entire population of independent components. In consequence, the method would be incapable of localising the component in the right area even if the head model was realistic and accurate.

To conclude, in this chapter we demonstrated in brief the concept behind source localisation in biomagnetic problems. ICA was presented as a mean of assisting the solution of the ill-posed inverse problem. However, recall that ICA does not effectively solve the inverse problem. It just offers a great simplification by allowing the sequential localisation of each potential

source. ICA reduces further the complexity by performing a single time slice analysis based on the assumption that the mixing matrix A is constant over time (a rather fundamental assumption in the ICA data model). However, the efficiency of the method still depends on the head model and the consequent approach which is used for the forward problem.

Chapter 7

Conclusions

This thesis examined the use of Independent Component Analysis (ICA) in magnetoencephalographic (MEG) data. This chapter provides a brief overview of the thesis in combination with the major contributions. The limitations of our work are also stated, and possible directions for future research are suggested

7.1 Overview

First, the basics of MEG were provided in chapter 2 with the intention to help understanding the challenging tasks of any biomagnetic study: (a) the identification and removal of artefact signals, and (b) the source localisation of the stimulated brain areas. This thesis dealt with the application of ICA in solving the first issue and simplifying the second one to a great extent.

The theoretical background of ICA was firmly established in chapter 3. ICA was initially developed in order to tackle the challenging problem of blind source separation. This problem is frequently met in biomagnetic studies due to the biological mechanisms which generate and propagate the magnetic fields. The key-element of ICA is the non-Gaussianity of the source signal. More assumptions about the ICA data model were introduced, and the intrinsic ambiguities of ICA were stated. In fact, one of them, namely the ambiguity of determining the order extraction, is of significant importance in an MEG clinical environment where fast processing is required. Finally, all the mathematical quantities which can be employed in expressing non-Gaussianity were presented in a comparative way.

Chapter 4 introduced the world of practical ICA. It presented the principal applied ICA algorithms with the intention to choose the algorithm, namely FastICA, which bears the most attractive property for real world applications. That is the capability of sequential extraction of independent components. This feature is exceptionally useful in datasets of large scale when only a few source signals are important for our study. The ambiguity which attracts our interest in this thesis was demonstrated with experimental studies in artificially generated data in noise-free and noisy environments. In addition, the problem of noise in ICA was explained in detail.

In chapter 5 we introduced a novel algorithm which succeeded in eliminating the intrinsic ambiguity of ICA in determining the order of extraction. This issue is particularly important in clinical environments where real-time processing is needed or when vast datasets are provided. In particular, we suggested a modification of the ordinary ICA quality function in order to take into account prior information about one or more source signals and favour their extraction as the first independent component under diverse running conditions of the algorithm. The proposed quality function was validated first with simulated data, and then with real MEG data. The most popular optimisation techniques were employed in order to select the one that gives the optimum result. The analysis showed that although simplex is sufficiently effective in small scale problems, simulated annealing should be applied in multidimensional problems such as in real MEG data. The suggested algorithm allows the sequential extraction of as many independent components as we have prior information about. If more components should be extracted, we can revert to the ordinary ICA algorithm.

The practical use of our method includes the verification of the presence of a suspected signal hidden in the observed mixed recordings in order to be used for further study or to remove it from the observations and reconstruct the original recordings free from that signal. Using the latter concept we managed to eliminate two major artefact signals in real MEG data, namely an ocular and a cardiac interference. Note the advantage of ICA in using non-averaged trials in rejecting the artefact signals. The researchers who provided us with the data typically apply a crude method of block trial averaging which in practice is proved to be inefficient under the presence of strong systematic artefacts such as in our case. Thus, ICA provides a novel way to remove artefacts which (a) does not require prolonged recordings, and (b) takes into account all information available without rejecting portions of data if they exceed an arbitrarily determined threshold of contamination.

Finally, in chapter 6 we discussed the use of ICA in tackling the second challenging problem in MEG (source localisation). Note that ICA does not attempt to solve the inverse problem by itself. In fact, no-one can solve it directly since it is an ill-posed problem. In general, further anatomical constraints should be imposed to ease the complexity. However, ICA removes to a great extent the computational burden associated with the current source localisation techniques. It provides invaluable information about the mixing matrix A which is roughly a function of the sensor-source relative geometry. This information can then be used in conjunction with existing techniques such as a minimum norm solution method. The main positive point of ICA is that it helps in localising each component sequentially by registering each source to a single component. Thus, there is no need to know in advance the number of dipoles to be fitted. In addition, since the mixing matrix is assumed to be constant over time in ICA, the analysis is performed on a single-time slice basis. We demonstrated the use of ICA with simulated data using the simple spherical homogeneous model. Next, we tried to validate the technique with real MEG data. Our valid hypothesis was that different stimulated fingers yield independent components which should be localised in distinct points in the somatosensory cortex. However, the lack of additional structural information diminished our options for selecting a realistic head model for the real MEG data, and the method failed to identify an independent component characterising a particular finger.

7.2 Limitations and Future Directions

The real MEG data, which are processed in this study, are limited in many ways. First of all, they are not accompanied by crucial supplemental information which could provide invaluable assistance in solving both fundamental biomagnetic problems of artefact removal and source localisation. In particular, the absence of an EOG and ECG, which should have been recorded in parallel with the MEG data, forces us to use the most heavily contaminated MEG channels in order to estimate the autocorrelation function used in our constrained ICA algorithm. Of course, this is not the most appropriate way and is partially effective. We have to assume that these particular artefacts are so strong that any contribution from other sources to these reference channels is insignificant.

Second, despite our intended pursuit for contaminated data in order to show the efficiency of ICA and in particular of our method in source signal decomposition, the artefact contamination seems to be overwhelming. This can be verified by the fact that during preprocessing

with an EVD, only a handful of principal components can explain the vast majority of variance in data. Thus, the dataset seems to be ill-conditioned. Especially the ocular artefact appears to affect all the data channels, even those located in the central areas and in the areas of interest.

Finally, an accurate source localisation requires realistic head models which go much further than our spherical homogeneous spherical model. In practice, we need a detailed profile of the changes in the conductivity in the conducting head medium which can be provided only with a structural MRI scan. Any localisation attempt is destined to fail no matter how efficient or helpful ICA is, when a non-realistic head model is employed for the real data.

The above limitations leave a number of issues which can motivate future work. Most application of ICA in the biomagnetic field, including this thesis, assume low levels of additive noise in order for the simple ICA data model to hold. As we have seen in section 4.3, the presence of noise makes the blind source separation problem almost impossible to solve. The optimisation task becomes virtually intractable for real world multidimensional applications. Thus some assumptions should be made about the distribution of the sources which may not be valid in practice [64]. Alternatively the problem can be partially solved when we have prior knowledge of the noise covariance matrix [67]. Therefore, it would be interesting if we could formulate a model which describes the external noise in MEG data. An alternative promising idea in noisy ICA is the use of Independent Factor Analysis (IFA), which is an extension of ICA in conjunction with Factor Analysis [7, 72].

Our constrained ICA algorithm uses a significant amount of data because all the cross-correlations between the different channels should be computed for all possible time lags. Therefore, we should find a way to keep as few time lags as possible in order to extract our desired signal with the minimal computational burden. Although our cICA algorithm was developed for use with the autocorrelation function, any other statistical property of a model signal can be used with the proper modification of the quality function. Further research should be conducted in selecting an optimum statistical discriminator. Finally, an alternative concept which shows great potential in incorporating prior knowledge in ICA and should be examined in detail is the application of the Bayesian theory in ICA [85].

Appendix A

FastICA Optimisation Algorithm

First, the observed data $\mathbf{x} = [x_1\ x_2\ \ldots\ x_M]^T$ are preprocessed as described in section 4.1. The whitened data $\mathbf{z} = [z_1\ z_2\ \ldots\ z_P]^T$ ($P \leq M$, with $P = M$ when the dimensionality is not reduced) are now used as input in the FastICA algorithm.

Recall from p.48 that the one-unit contrast function is :

$$J_G(\mathbf{w}_i) = \Big[E\{G(\mathbf{w}_i^T\mathbf{z})\} - E\{G(y_{G\nu})\}\Big]^2 \tag{A-1}$$

$$\text{under the constraint of } E\{(\mathbf{w}_i^T\mathbf{z})^2\} = 1$$

The maxima of $J_G(\mathbf{w}_i)$ are obtained at the extrema of $E\{G(\mathbf{w}_i^T\mathbf{z})\}$. Then $\mathbf{w}_i^T\mathbf{z}$ is the estimate of the i^{th} independent component.

Since the input data $\mathbf{z}$ are whitened, the corresponding constraint can be written as

$$E\{(\mathbf{w}_i^T\mathbf{z})^2\} = E\{\ (w_1z_1 + w_2z_2 + \ldots + w_Pz_P)^2\} = w_1^2 + w_2^2 + \ldots + w_P^2 = 1 \Rightarrow$$

$$\|\mathbf{w}\| = 1 \tag{A-2}$$

According to the Kuhn-Tucker conditions [99], the extrema of $E\{G(\mathbf{w}^T\mathbf{z})\}$ under the constraint $\|\mathbf{w}\|^2 - 1 = 0$ are obtained at points where:

$$\frac{dE\{G(\mathbf{w}^T\mathbf{z})\}}{d\mathbf{w}} - \lambda\frac{d(\|\mathbf{w}\|^2 - 1)}{d\mathbf{w}} = 0 \tag{A-3}$$

However,

$$\frac{dE\{G(\mathbf{w}^T\mathbf{z})\}}{d\mathbf{w}} = \begin{bmatrix} \frac{dE\{G(\sum w_iz_i)\}}{dw_1} \\ \vdots \\ \frac{dE\{G(\sum w_iz_i)\}}{dw_P} \end{bmatrix} = \begin{bmatrix} E\{z_1g\left(\sum w_iz_i\right)\} \\ \vdots \\ E\{z_Pg\left(\sum w_iz_i\right)\} \end{bmatrix} \Rightarrow$$

$$\frac{dE\{G(\mathbf{w}^T\mathbf{z})\}}{d\mathbf{w}} = E\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} \tag{A-4}$$

where g is the derivative of G, and

$$\frac{d(\|\mathbf{w}\|^2 - 1)}{d\mathbf{w}} = \frac{d\|\mathbf{w}\|^2}{d\mathbf{w}} = \frac{d\sum w_i^2}{d\mathbf{w}} = 2\mathbf{w} \tag{A-5}$$

Thus, (A-3) can be written as:

$$E\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} - \beta\mathbf{w} = 0 \tag{A-6}$$

where $\beta \equiv E\{\hat{\mathbf{w}}^T\mathbf{z}g(\hat{\mathbf{w}}^T\mathbf{z})\}$ is a constant, with $\hat{\mathbf{w}}$ the value of $\mathbf{w}$ at the optimum.

Let us solve (A-6) using Newton's method [6]. First, denote by F the vector function on the left-hand side:

$$F(\mathbf{w}) \equiv E\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} - \beta\mathbf{w} \tag{A-7}$$

in other words,

$$\begin{bmatrix} F_1 \\ \vdots \\ F_P \end{bmatrix} = \begin{bmatrix} E\{z_1 g(\sum w_i z_i)\} - \beta w_1 \\ \vdots \\ E\{z_P g(\sum w_i z_i)\} - \beta w_P \end{bmatrix}$$

and by J its Jacobian matrix:

$$J(\mathbf{w}) = \begin{bmatrix} \frac{\partial F_1}{w_1} & \cdots & \frac{\partial F_1}{w_P} \\ \vdots & & \vdots \\ \frac{\partial F_P}{w_1} & \cdots & \frac{\partial F_P}{w_P} \end{bmatrix} \tag{A-8}$$

Therefore,

$$J(\mathbf{w}) = \begin{bmatrix} E\{z_1^2 g'(\mathbf{w}^T\mathbf{z})\} - \beta & \cdots & E\{z_1 z_P g'(\mathbf{w}^T\mathbf{z})\} \\ \vdots & & \vdots \\ E\{z_1 z_P g'(\mathbf{w}^T\mathbf{z})\} & \cdots & E\{z_P^2 g'(\mathbf{w}^T\mathbf{z})\} - \beta \end{bmatrix} \Rightarrow$$

$$J(\mathbf{w}) = E\{\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\} - \beta I \tag{A-9}$$

where g' is the derivative of g.

However, if we consider the following approximation :

$$E\{\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\} \approx E\{\mathbf{z}\mathbf{z}^T\}E\{g'(\mathbf{w}^T\mathbf{z})\} = E\{g'(\mathbf{w}^T\mathbf{z})\}I \tag{A-10}$$

the Jacobian matrix becomes diagonal :

$$J(\mathbf{w}) = \left[E\{g'(\mathbf{w}^T\mathbf{z})\} - \beta\right] \mathrm{I} \tag{A-11}$$

and can be inverted easily :

$$J^{-1}(\mathbf{w}) = \left[E\{g'(\mathbf{w}^T\mathbf{z})\} - \beta\right]^{-1} \mathrm{I} \tag{A-12}$$

Assume now that there exists a vector $\hat{\mathbf{w}}$ such that $F(\hat{\mathbf{w}}) = 0$. If $J(\hat{\mathbf{w}}) \neq 0$, then it can be shown [105] that the sequence $\{\mathbf{w}_k\}_{k=0}^{\infty}$ defined by the iteration :

$$\mathbf{w}_k = \mathbf{w}_{k-1} - J^{-1}(\mathbf{w}_{k-1})F(\mathbf{w}_{k-1}) \text{ , for } k = 1, 2, \ldots \tag{A-13}$$

will converge to $\hat{\mathbf{w}}$ for any initial approximation $\mathbf{w}_0$.

Hence, we obtain the following formula for Newton iteration :

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \frac{E\{\mathbf{z}g(\mathbf{w}_{k-1}^T\mathbf{z})\} - \beta\mathbf{w}_{k-1}}{E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\} - \beta} \text{ , for } k = 1, 2, \ldots \tag{A-14}$$

A normalisation step is added after each iteration :

$$\mathbf{w}_k^* = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|} \tag{A-15}$$

where $\mathbf{w}_k^*$ denotes the normalised value of $\mathbf{w}$ after the k^{th} iteration.

The algorithm can be further simplified by multiplying both sides of (A-14) by the scalar $\beta - E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\}$:

$$\left[\beta - E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\}\right] \mathbf{w}_k = E\{\mathbf{z}g(\mathbf{w}_{k-1}^T\mathbf{z})\} - E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\}\mathbf{w}_{k-1} \tag{A-16}$$

Due to normalisation in (A-15), we can omit the scalar coefficient on the left-hand side of (A-16), ending up with the following iteration algorithm :

$$\mathbf{w}_k = E\{\mathbf{z}g(\mathbf{w}_{k-1}^T\mathbf{z})\} - E\{g'(\mathbf{w}_{k-1}^T\mathbf{z})\}\mathbf{w}_{k-1} \tag{A-17}$$

$$\mathbf{w}_k^* = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}$$

Appendix B

Gradient Optimisation in cICA

B.1 The Gradient of the ICA Term J_G

Let us denote by J_G the standard ICA term:

$$J_G = \left[\frac{1}{K} \sum_{k=1}^{K} \left\{ \left(\sum_{j=1}^{M} w_i z_{ik} \right)^4 \right\} - 3 \right]^2 \tag{B-1}$$

For $i \neq m$, we have:

$$\frac{\mathrm{d}J_G}{\mathrm{d}w_i} = \frac{\partial J_G}{\partial w_i} + \frac{\partial J_G}{\partial w_m} \frac{\partial w_m}{\partial w_i} \tag{B-2}$$

However,

$$\frac{\partial J_G}{\partial w_i} = 2 \left[\frac{1}{K} \sum_{k=1}^{K} \left\{ \left(\sum_{j=1}^{M} w_i z_{ik} \right)^4 \right\} - 3 \right] \frac{1}{K} \sum_{k=1}^{K} \left\{ z_{ik} \left(\sum_{p=1}^{M} w_p z_{pk} \right)^3 \right\} \tag{B-3}$$

Recall the constraint of the weight vector $\mathbf{w}$ for whitened data $\mathbf{z}$:

$$w_1^2 + w_2^2 + \ldots + w_M^2 = 1 \Rightarrow w_M = (1 - w_1^2 - w_2^2 - \ldots - w_{M-1}^2)^{1/2} \tag{B-4}$$

Note that w_M can be computed from equation B-4 if we know the values of all other weights. We decide to use the positive sign when we take the square root. This does not imply any loss of generality, since it is known that ICA recovers the independent components up to a constant factor, positive or negative.

From equation B-4 we have :

$$2w_i + 2w_M \frac{\partial w_M}{\partial w_i} = 0 \Rightarrow \frac{\partial w_M}{\partial w_i} = -\frac{w_i}{w_M} \tag{B-5}$$

B.2 The Gradient of the Constraint Term J_C

Let us denote by J_C the constraint term :

$$J_C = \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} \left[w_p \sum_{j=1}^{M} w_j r_{pj}(\tau) \right] - r_{model}(\tau) \right\}^2 \tag{B-6}$$

Thus

$$J_C = \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} \left[w_p^2 r_{pp}(\tau) + w_p \sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \right] - r_{model}(\tau) \right\}^2 \tag{B-7}$$

If

$$F(\tau) \equiv \sum_{p=1}^{M} \left[w_p^2 r_{pp}(\tau) + w_p \sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \right] - r_{model}(\tau) \tag{B-8}$$

then

$$J_C = \sum_{\tau=0}^{K-1} F^2(\tau) \Rightarrow \frac{\partial J_C}{\partial w_i} = 2 \sum_{\tau=0}^{K-1} F(\tau) \frac{\partial F(\tau)}{\partial w_i} \tag{B-9}$$

Hence

$$\frac{\partial J_C}{\partial w_i} = 2 \sum_{\tau=0}^{K-1} \left\{ F(\tau) \left(2 w_i r_{ii}(\tau) + \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right) \right\} \tag{B-10}$$

For $i \neq M$, we have :

$$\frac{\mathrm{d}J_C}{\mathrm{d}w_i} = \frac{\partial J_C}{\partial w_i} + \frac{\partial J_C}{\partial w_M} \frac{\partial w_M}{\partial w_i} \tag{B-11}$$

Let us now consider equation B-10 :

$$\frac{\partial J_C}{\partial w_i} = 4 w_i \sum_{\tau=0}^{K-1} \{ F(\tau) r_{ii}(\tau) \} + 2 \sum_{\tau=0}^{K-1} \left\{ F(\tau) \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right\}$$

Let us write :

$$\frac{\partial J_C}{\partial w_i} \equiv A + B \tag{B-12}$$

where

$$A \equiv 4 w_i \sum_{\tau=0}^{K-1} \{ F(\tau) r_{ii}(\tau) \} \tag{B-13}$$

and

$$B \equiv 2 \sum_{\tau=0}^{K-1} \left\{ F(\tau) \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right\} \tag{B-14}$$

Let us consider first the A term and substitute for $F(\tau)$:

$$A = 4w_i \sum_{\tau=0}^{K-1} \left\{ r_{ii}(\tau) \left(\sum_{p=1}^{M} \left[w_p^2 r_{pp}(\tau) + w_p \sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \right] - r_{model}(\tau) \right) \right\} =$$

$$= 4w_i \sum_{\tau=0}^{K-1} \left(r_{ii}(\tau) \sum_{p=1}^{M} w_p^2 r_{pp}(\tau) \right) + 4w_i \sum_{\tau=0}^{K-1} \left(r_{ii}(\tau) \sum_{p=1}^{M} w_p \left[\sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \right] \right) -$$

$$-4w_i \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{model}(\tau)$$

Thus

$$A \equiv A1 + A2 + A3 \tag{B-15}$$

where

$$A1 \equiv 4w_i \sum_{\tau=0}^{K-1} \left(r_{ii}(\tau) \sum_{p=1}^{M} w_p^2 r_{pp}(\tau) \right) \tag{B-16}$$

and

$$A2 \equiv 4w_i \sum_{\tau=0}^{K-1} \left(r_{ii}(\tau) \sum_{p=1}^{M} w_p \left[\sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \right] \right) \tag{B-17}$$

and

$$A3 \equiv -4w_i \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{model}(\tau) \tag{B-18}$$

Let us simplify the above terms, starting from $A1$:

$$A1 = 4w_i \sum_{\tau=0}^{K-1} \left(r_{ii}(\tau) \sum_{p=1}^{M} w_p^2 r_{pp}(\tau) \right) = 4w_i \sum_{\tau=0}^{K-1} \sum_{p=1}^{M} r_{ii}(\tau) w_p^2 r_{pp}(\tau) =$$

$$= 4w_i \sum_{p=1}^{m} w_p^2 \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{pp}(\tau)$$

Thus

$$A1 = 4w_i \sum_{p=1}^{m} w_p^2 A1_{\text{off}}(i,p) \tag{B-19}$$

where

$$A1_{\text{off}}(i,p) = \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{pp}(\tau) \tag{B-20}$$

and the subscript "off" means that this factor can be computed off-line. There are $\frac{M(M+1)}{2}$ such factors $A1_{\text{off}}$, since $A1_{\text{off}}$ is symmetric with respect to its arguments, and each of its arguments takes values from 1 to M.

Let us consider now term $A2$:

$$A2 = 4w_i \sum_{\tau=0}^{K-1} \left(r_{ii}(\tau) \sum_{p=1}^{M} w_p \left[\sum_{\substack{j=1 \\ j\neq p}}^{M} w_j r_{pj}(\tau) \right] \right) =$$

$$= 4w_i \sum_{\tau=0}^{K-1} \sum_{p=1}^{M} r_{ii}(\tau) w_p \sum_{\substack{j=1 \\ j\neq p}}^{M} w_j r_{pj}(\tau) = 4w_i \sum_{\tau=0}^{K-1} \sum_{p=1}^{M} w_p \sum_{\substack{j=1 \\ j\neq p}}^{M} w_j r_{ii}(\tau) r_{pj}(\tau) =$$

$$= 4w_i \sum_{p=1}^{M} w_p \sum_{\substack{j=1 \\ j\neq p}}^{M} w_j \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{pj}(\tau)$$

Thus

$$A2 = 4w_i \sum_{p=1}^{M} \left(w_p \sum_{\substack{j=1 \\ j\neq p}}^{M} w_j A2_{\text{off}}(i,p,j) \right) \tag{B-21}$$

where

$$A2_{\text{off}}(i,p,j) = \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{pj}(\tau) \tag{B-22}$$

There are M^3 such factors that can be computed off-line.

Term $A3$ is easy :

$$A3 = -4w_i A3_{\text{off}}(i) \tag{B-23}$$

where

$$A3_{\text{off}}(i) = \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{model}(\tau) \tag{B-24}$$

There are only M such factors.

Let us consider now the B term of equation B-14 :

$$B = 2 \sum_{\tau=0}^{K-1} \left\{ F(\tau) \sum_{\substack{j=1 \\ j\neq i}}^{M} w_j r_{ij}(\tau) \right\} =$$

$$= 2 \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} w_p^2 r_{pp}(\tau) \sum_{\substack{j=1 \\ j\neq i}}^{M} w_j r_{ij}(\tau) + \sum_{p=1}^{M} w_p \sum_{\substack{j=1 \\ j\neq p}}^{M} w_j r_{pj}(\tau) \sum_{\substack{n=1 \\ n\neq i}}^{M} w_n r_{in}(\tau) - r_{model}(\tau) \sum_{\substack{j=1 \\ j\neq i}}^{M} w_j r_{ij}(\tau) \right\}$$

where we have changed the dummy index in one of the sums from j to n so that to avoid confusion.

Thus we may write

$$B \equiv B1 + B2 + B3 \tag{B-25}$$

where

$$B1 \equiv 2 \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} w_p^2 r_{pp}(\tau) \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right\} \tag{B-26}$$

$$B2 \equiv 2 \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} w_p \sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \sum_{\substack{n=1 \\ n \neq i}}^{M} w_n r_{in}(\tau) \right\} \tag{B-27}$$

and

$$B3 \equiv -2 \sum_{\tau=0}^{K-1} \left\{ r_{model}(\tau) \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right\} \tag{B-28}$$

Let us simplify the above terms, starting from $B1$:

$$B1 = 2 \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} w_p^2 r_{pp}(\tau) \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right\} = 2 \sum_{p=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \left(w_p^2 w_j \sum_{\tau=0}^{K-1} r_{pp}(\tau) r_{ij}(\tau) \right)$$

Thus

$$B1 = 2 \sum_{p=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \left(w_p^2 w_j B1_{\text{off}}(p,i,j) \right) \tag{B-29}$$

where

$$B1_{\text{off}}(p,i,j) = \sum_{\tau=0}^{K-1} r_{pp}(\tau) r_{ij}(\tau) \tag{B-30}$$

Note that $B1_{\text{off}}(p,i,j) = A2_{\text{off}}(i,p,j)$, so these M^3 factors should only be computed once.

Term $B2$ is :

$$B2 = 2 \sum_{\tau=0}^{K-1} \left\{ \sum_{p=1}^{M} w_p \sum_{\substack{j=1 \\ j \neq p}}^{M} w_j r_{pj}(\tau) \sum_{\substack{n=1 \\ n \neq i}}^{M} w_n r_{in}(\tau) \right\} = 2 \sum_{p=1}^{M} \sum_{\substack{j=1 \\ j \neq p}}^{M} \sum_{\substack{n=1 \\ n \neq i}}^{M} \left\{ w_p w_j w_n \sum_{\tau=0}^{K-1} r_{pj}(\tau) r_{in}(\tau) \right\}$$

Thus

$$B2 = 2 \sum_{p=1}^{M} \sum_{\substack{j=1 \\ j \neq p}}^{M} \sum_{\substack{n=1 \\ n \neq i}}^{M} \left\{ w_p w_j w_n B2_{\text{off}}(p,j,i,n) \right\} \tag{B-31}$$

where

$$B2_{\text{off}}(p,j,i,n) = \sum_{\tau=0}^{K-1} r_{pj}(\tau) r_{in}(\tau) \tag{B-32}$$

There are $\frac{M^2(M^2+1)}{2}$ such factors because $B2_{\text{off}}$ is symmetric with respect to the two pairs of its arguments, i.e. $B2_{\text{off}}(p,j,i,n) = B2_{\text{off}}(i,n,p,j)$.

Finally, we turn to term $B3$:

$$B3 = -2 \sum_{\tau=0}^{K-1} \left\{ r_{model}(\tau) \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j r_{ij}(\tau) \right\} = -2 \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j \sum_{\tau=0}^{K-1} r_{ij}(\tau) r_{model}(\tau)$$

Thus

$$B3 = -2 \sum_{\substack{j=1 \\ j \neq i}}^{M} w_j B3_{\text{off}}(i,j) \tag{B-33}$$

where

$$B3_{\text{off}}(i,j) = \sum_{\tau=0}^{K-1} r_{ij}(\tau) r_{model}(\tau) \tag{B-34}$$

There are $M(M-1)$ such factors since $i \neq j$.

To sum up, we have :

$$\frac{\partial J_C}{\partial w_i} = A + B = A1 + A2 + A3 + B1 + B2 + B3 \tag{B-35}$$

where

$$A1 = 4w_i \sum_{p=1}^{M} w_p^2 A1_{\text{off}}(i,p) \tag{B-36}$$

$$A1_{\text{off}}(i,p) = \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{pp}(\tau) \tag{B-37}$$

$$A2 = 4w_i \sum_{p=1}^{M} \left(w_p \sum_{\substack{j=1 \\ j\neq p}}^{M} w_j A2_{\text{off}}(i,p,j) \right) \tag{B-38}$$

$$A2_{\text{off}}(i,p,j) = \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{pj}(\tau) \tag{B-39}$$

$$A3 = -4w_i A3_{\text{off}}(i) \tag{B-40}$$

$$A3_{\text{off}}(i) = \sum_{\tau=0}^{K-1} r_{ii}(\tau) r_{model}(\tau) \tag{B-41}$$

$$B1 = 2 \sum_{p=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \left(w_p^2 w_j A2_{\text{off}}(p,i,j) \right) \tag{B-42}$$

$$B2 = 2 \sum_{p=1}^{M} \sum_{\substack{j=1 \\ j\neq p}}^{M} \sum_{\substack{n=1 \\ n\neq i}}^{M} \left\{ w_p w_j w_n B2_{\text{off}}(p,j,i,n) \right\} \tag{B-43}$$

$$B2_{\text{off}}(p,j,i,n) = \sum_{\tau=0}^{K-1} r_{pj}(\tau) r_{in}(\tau) \tag{B-44}$$

$$B3 = -2 \sum_{\substack{j=1 \\ j\neq i}}^{M} w_j B3_{\text{off}}(i,j) \tag{B-45}$$

$$B3_{\text{off}}(i,j) = \sum_{\tau=0}^{K-1} r_{ij}(\tau) r_{model}(\tau) \tag{B-46}$$

Appendix C

Calculation of Constraint Term J_C

The constraint term $J_C(\mathbf{w})$ of equation 5.6 is highly computationally intensive when evaluated within an iterative algorithm such as simplex or simulated annealing where thousands of iterations may be required. The autocorrelation matrix $A_{\mathbf{x}}(\tau)$ consists of M^2 elements for each time lag τ. A huge amount of data should be loaded in computer memory in each iteration. Thus it is necessary to transfer as many calculations as possible off-line. We show here how the calculation of $J_C(\mathbf{w})$ can be simplified. Indeed,

$$\begin{aligned}
J_C(\mathbf{w}) &= \sum_{\tau=0}^{K-1} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w} - r_{model}(\tau)\right]^2 = \\
&= \sum_{\tau=0}^{K-1} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w} - r_{model}(\tau)\right]\left[\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w} - r_{model}(\tau)\right] = \\
&= \sum_{\tau=0}^{K-1} [\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w}\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w} - \mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w} r_{model}(\tau) - \\
&\quad -r_{model}(\tau)\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w} + r_{model}(\tau)r_{model}(\tau)] = \\
&= \sum_{\tau=0}^{K-1} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w}\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w}\right] - 2\mathbf{w}^T \left(\sum_{\tau=0}^{K-1} r_{model}(\tau)A_{\mathbf{x}}(\tau)\right)\mathbf{w} + \sum_{\tau=0}^{K-1} r_{model}^2(\tau)
\end{aligned}$$

Therefore we can write

$$J_C(\mathbf{w}) = A_1 - 2\mathbf{w}^T A_2 + A_3 \qquad \text{(B-1)}$$

where

$$A_1 \equiv \sum_{\tau=0}^{K-1} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w}\mathbf{w}^T A_{\mathbf{x}}(\tau)\mathbf{w}\right]$$

$$A_2 \equiv \sum_{\tau=0}^{K-1} r_{model}(\tau)A_{\mathbf{x}}(\tau)$$

$$A_3 \equiv \sum_{\tau=0}^{K-1} r^2_{model}(\tau)$$

Terms A_2 and A_3 can be calculated off-line as they do not depend on the unknown vector $\mathbf{w}$. The term A_1 can be simplified more. Since $\mathbf{w}$ is an M-dimensional vector, $\mathbf{w}\mathbf{w}^T$ is an $(M \times M)$ matrix, say B. Matrix B can be written as:

$$B = \sum_{i,j=1}^{M} B_{ij}$$

where B_{ij} is a $(M \times M)$ matrix the elements of which are all zero except the element at position (i,j) which is equal to $b_{ij} = w_i w_j$.

Thus, A_1 can be rewritten as

$$\begin{aligned} A_1 &= \sum_{\tau=0}^{K-1} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau) B A_{\mathbf{x}}(\tau)\mathbf{w}\right] = \sum_{\tau=0}^{K-1} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau) \left(\sum_{i,j=1}^{M} B_{ij}\right) A_{\mathbf{x}}(\tau)\mathbf{w}\right] = \\ &= \sum_{\tau=0}^{K-1} \sum_{i,j=1}^{M} \left[\mathbf{w}^T A_{\mathbf{x}}(\tau) B_{ij} A_{\mathbf{x}}(\tau)\mathbf{w}\right] = \sum_{\tau=0}^{K-1} \sum_{i,j=1}^{M} \left[\mathbf{w}^T b_{ij} A_{\mathbf{x}_i}(\tau) A_{\mathbf{x}_j}(\tau)\mathbf{w}\right] \\ &= \sum_{i,j=1}^{M} \left[\mathbf{w}^T b_{ij} \sum_{\tau=0}^{K-1} \left[A_{\mathbf{x}_i}(\tau) A_{\mathbf{x}_j}(\tau)\right] \mathbf{w}\right] = \sum_{i,j=1}^{M} \left[b_{ij} \mathbf{w}^T \sum_{\tau=0}^{K-1} \left[A_{\mathbf{x}_i}(\tau) A_{\mathbf{x}_j}(\tau)\right] \mathbf{w}\right] \end{aligned}$$

where $A_{\mathbf{x}_i}$ is the $(M \times 1)$ vector $(A_{\mathbf{x}_{1i}}\ A_{\mathbf{x}_{2i}}\ \ldots\ A_{\mathbf{x}_{mi}}\ )^T$ and $A_{\mathbf{x}_j}$ is the $(1 \times M)$ vector $(A_{\mathbf{x}_{j1}}\ A_{\mathbf{x}_{j2}}\ \ldots\ A_{\mathbf{x}_{jm}})$. In other words, $A_{\mathbf{x}_i} A_{\mathbf{x}_j}$ is the outer product of two vectors, i.e. an $(M \times M)$ matrix. The sum $\sum_{\tau=0}^{K-1} \left[A_{\mathbf{x}_i}(\tau) A_{\mathbf{x}_j}(\tau)\right]$ can be computed off-line.

For our real MEG data, $M = 10$ and τ takes 20 000 values in total. Therefore, if $J_C(\mathbf{w})$ is calculated in the straightforward way, 2 million numbers - elements of $A_{\mathbf{x}}(\tau)$ (i.e. 16Mb of data) should be loaded in the computer memory at each iteration. After the rearrangements, only 80kb of data have to be loaded, a reduction by a factor of 200.

Bibliography

[1] F. Acernese, A. Ciaramella, S. De Martino, R. De Rosa, M. Falanga, and R. Tagliaferri. Neural networks for blind-source separation of Stromboli explosion quakes. *IEEE Transactions on Neural Networks*, 14(1):167–175, 2003.

[2] F. Aires, A. Chédin, and J.P. Nadal. Independent component analysis of multivariate time series. Application to the tropical SST variability. *Journal of Geophysical Research*, 105:17437–17455, 2000.

[3] S. Amari. Neural learning in structured parameter spaces - natural riemannian gradient. In *Advances in Neural Information Procesing Systems*, volume 9, pages 127–133, 1997.

[4] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Procesing Systems*, volume 8, pages 752–763, 1996.

[5] Anon. Magnetoencephalography. *Lancet*, 335(8689):576–577, 1990.

[6] M. Aoki. *Introduction to optimization techniques.* The Macmillan Company, New York, 1971.

[7] H. Attias. Independent factor analysis. *Neural Computation*, 10(6):1373–1425, 1998.

[8] A.D. Back and A.S Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8(4):473–484, 1997.

[9] S. Baillet and L. Garnero. A bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. *IEEE Transactions on Biomedical Engineering*, 44(5):374–385, 1997.

[10] A.K. Barros and A. Cichocki. Extraction of specific signals with temporal structure. *Neural Computation*, 13(9):1995–2000, 2001.

[11] A.K. Barros, R. Vigario, V. Jousmaki, and N. Ohnishi. Extraction of event-related signals from multi-channel bioelectrical measurements. *IEEE Transactions on Biomedical Engineering*, 47(5):583–588, 2000.

[12] M. Bartlett and J.T. Sejnowski. Viewpoint invariant face recognition using independent component analysis and attractor networks. In *Advances in Neural Information Processing*, volume 9, pages 817–823, 1997.

[13] C. Baumgartner. *Analysis of the electrical activity of the brain*, chapter Clinical applications of source localisation techniques – the human somatosensory cortex, pages 271–308. John Wiley & Sons, 1997.

[14] M.F. Bear, B.W. Connors, and M.A. Paradiso. *Neuroscience: exploring the brain*, pages 22–129. Lippincott Williams and Wilkins, Philadelphia, 2nd edition, 2001.

[15] A.J. Bell and T.J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[16] O. Bermond and J.F. Cardoso. Approximate likelihood for noisy mixtures. In *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pages 325–330, Aussois, France, 1999.

[17] M.J. Box. A comparison of several current optimization methods and the use of transformations in constrained problems. *Computer Journal*, 9(1):67–77, 1966.

[18] M. Bundo, S. Inao, A. Nakamura, T. Kato, K. Ito, M. Tadokoro, R. Kabeya, T. Sugimoto, Y. Kajita, and J. Yoshida. Changes of neural activity correlate with the severity of cortical ischemia in patients with unilateral major cerebral artery occlusion. *Stroke*, 33(1):61–66, 2002.

[19] J.F. Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor. Blind indentification of more sources than sensors. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'91)*, pages 3109–3112, Toronto, Canada, 1991.

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

[21] J.F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, 1998.

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

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

[24] J.F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal of Matrix Analysis and Applications*, 17(1):161–164, 1996.

[25] C.I. Chang. Linear spectral random mixture analysis for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40(2):375–392, 2002.

[26] A. Cichocki, R. Thawonmas, and S. Amari. Sequential blind signal extraction in order specified by stochastic properties. *Electronic Letters*, 33:64–65, 1997.

[27] D. Cohen. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. *Science*, 161:784–786, 1968.

[28] D. Cohen. Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer. *Science*, 175:664–666, 1972.

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

[30] T.M. Cover and J.A. Thomas. *Elements of information theory.* Wiley-Interscience, New York, 1991.

[31] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen. CDMA delay estimation using a fast ICA algorithm. In *Proceedings of the IEEE International Symposium on Personal, Indoor, and Mobile Communications (PIMRC'00)*, pages 1117–1120, London, UK, 2000.

[32] CTF Systems Inc. 15 - 1750 McLean Ave., Port Coquitlam, B.C., V3C 1M9, Canada, <http://www.ctf.com/home.html>.

[33] B.N. Cuffin. Effects of head shapes on EEGs and MEGs. *IEEE Transactions on Biomedical Engineering*, 37:15–22, 1990.

[34] B.N. Cuffin and D. Cohen. Magnetic fields produced by models of biological current sources. *Journal of Applied Physics*, 48:3971–3980, 1977.

[35] A.M Dale and E. Halgren. Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology*, 11(2):202–208, 2001.

[36] A.P. Dempster, N.M Laird, and D.B. Rubin. Maximum likelihood from incompleted data via the EM algorithm. *Journal of the Royal Statistical Society ser. B*, 1–38:1977, 39.

[37] G.H. Dunteman. *Principal components analysis.* Sage Publications, 1989.

[38] J.S. Ebersole. Non-invasive pre-surgical evaluation with EEG/MEG source analysis. *Electroencephalography and Clinical Neurophysiology - Supplement*, 50:167–174, 1999.

[39] A.W.F. Edwards. *Likelihood.* The John Hopkins University Press, 1992.

[40] H. Eswaran, J. Wilson, H. Preissl, S. Robinson, J. Vrba, P. Murphy, D. Rose, and C. Lowery. Magnetoencephalographic recordings of visual evoked brain activity in the human fetus. *Lancet*, 360(9335):779–780, 2002.

[41] A.S. Ferguson, X. Zhang, and G. Stroink. A complete linear discretization for calculating the magnetic field using the boundary element method. *IEEE Transactions on Biomedical Engineering*, 41(5):455–459, 1994.

[42] I.K Fodor and C. Kamath. Using independent component analysis to separate signals in climate data. In *Independent Component Analysis, Wavelets, and Neural Networks, Proceedings of the SPIE*, volume 5102, pages 25–36, Orlando, USA, 2003.

[43] J. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C–23(9):881–889, 1974.

[44] M. Funaro, E. Oja, and H. Valpola. Independent component analysis for artefact separation in astrophysical images. *Neural Networks*, 16(3–4):469–478, 2003.

[45] M. Girolami. Hierarchic dichotomizing of polychotomous data – an ICA based data mining tool. *Computing and Information Systems*, 5(3):107–112, 1998.

[46] G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 1989.

[47] E. Gordon. Brain imaging technologies: how, what, when and why? *Australian and New Zealand Journal of Psychiatry*, 33(2):187–196, 1999.

[48] B.S. Gottfried and J. Weisman. *Introduction to optimization theory.* Prentice-Hall, New Jersey, 1973.

[49] M. Gray, J.R. Movellan, and J.T. Sejnowski. Dynamic features for visual speechreading: a systematic comparison. In *Advances in Neural Information Processing Systems*, volume 9, pages 751–757, 1997.

[50] M. Hajek, R. Huonker, C. Boehle, H.P. Volz, H. Nowak, and H. Sauer. Abnormalities of auditory evoked magnetic fields and structural changes in the left hemisphere of male schizophrenics - a magnetoencephalographic-magnetic resonance imaging study. *Biological Psychiatry*, 42(7):609–616, 1997.

[51] P. Hall, D. Marshall, and R. Martin. Adding and subtracting eigenspaces with EVD and SVD. *Image and Vision Computing*, 20(13–14):1009–1016, 2002.

[52] M. Hämäläinen, R. Hari, R.J. Ilmoniemi, J. Knuutila, and O.V. Lounasmaa. Magnetoencephalography – theory, instrumentation, and applications in noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1993.

[53] R. Hari and O.V. Lounasmaa. Recording and interpretation of cerebral magnetic fields. *Science*, 244(4903):432–436, 1989.

[54] H.H. Harman. *Modern factor analysis.* The University of Chicago Press, Chicago, 2nd edition, 1967.

[55] J. Haueisen, C. Ramon, M. Eiselt, H. Brauer, and H. Nowak. Influence of tissue resistivities on neuromagnetic fields and electric potentials studied with a finite element model of the head. *IEEE Transactions on Biomedical Engineering*, 44(8):727–735, 1997.

[56] E. Haykin, editor. *Unsupervised adaptive filtering – Volume I: Blind source separation.* Wiley-Interscience, New York, 2000.

[57] H. Helmholtz. Uber einige Gesetze der Vertheilung elektrischer Ströme in körperlichen Leitern, mit Anwendung auf die thierisch-elektrischen Versuche [Some laws concerning the distribution of electrical currents in conductors with applications to experiments on animal electricity]. *Annalen der Physik und Chemie*, 89(6):211–233,353–377, 1853.

[58] Helsinki University of Technology – Laboratory of Computer and Information Science – ICA Project. <http://www.cis.hut.fi/projects/ica/>.

[59] B. Horwitz and D. Poeppel. How can EEG/MEG and fMRI/PET data be combined? *Human Brain Mapping*, 17(1):1–3, 2002.

[60] P.O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network Computation in Neural Systems*, 11(3):191–210, 2000.

[61] M. Huang, J. Mosher, and R. Leahy. A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG. *Physics in Medicine and Biology*, 44(2):423–440, 1999.

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

[63] A. Hyvärinen. One-unit contrast functions for independent component analysis: a statistical analysis. In *Proceedings of IEEE Neural Networks for Signal Processing Workshop '97*, pages 388–397, Florida, USA, 1997.

[64] A. Hyvärinen. Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22:49–67, 1998.

[65] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, pages 273–279, 1998.

[66] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[67] A. Hyvärinen. Gaussian moments for noisy independent component analysis. *IEEE Signal Processing Letters*, 6(6):145–147, 1999.

[68] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[69] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* Wiley-Interscience, New York, 2001.

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

[71] A. Hyvärinen, J. Särelä, and R. Vigário. Spikes and bumps: artefacts generated by independent component analysis with insufficient sample size. In *Proccedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pages 425–429, Aussois, France, 1999.

[72] S. Ikeda. *Advances in Independent Component Analysis*, chapter ICA on noisy data: A factor analyis approach. Springer, 2000.

[73] S. Ikeda and N. Murata. A method of ICA in time frequency domain. In *Proccedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pages 365–370, Aussois, France, 1999.

[74] A.A. Ioannides. Comparison of magnetoencephalography with other functional techniques. *Clinical Physics and Physiological Measurement*, 12 Suppl A:23–28, 1991.

[75] J.E. Jackson. *A user's guide to principal components.* Wiley, New York, 1991.

[76] B.W. Jervis, M. Coelho, and M. Morgan. Effect on EEG responses of removing ocular artifacts by proportional EOG subtraction. *Medical & Biological Engineering & Computing*, 27:484–490, 1988.

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

[78] T.P. Jung, S. Makeig, C. Humphries, T.W. Lee, M.J. McKeown, V. Iragui, and T.J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

[79] T.P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T.J. Sejnowski. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, 111(10):1745–1758, 2000.

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

[81] O. Kallenberg. *Foundations of modern probability.* Springer–Verlag, New York, 1997.

[82] J.F. Kenney. *Mathematics of statistics Part 1.* Van Nostrand, London, 3rd edition, 1954.

[83] S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 4598(220):671–680, 1983.

[84] K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *Proceedings of the Fifth International Conference on Neural Information Processing (ICONIP'98)*, volume 2, pages 895–898, Japan, 1998.

[85] K. Knuth. A Bayesian approach to source separation. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 283–288, Aussois, France, 1999.

[86] K. Kobayashi, C.J. James, T. Nakahori, T. Akiyama, and J. Gotman. Isolation of epileptiform discharges from unaveraged EEG by independent component analysis. *Clinical Neurophysiology*, 110(10):1755–1763, 1999.

[87] Z.J. Koles. Trends in EEG source localisation. *Electroencephalography and Clinical Neurophysiology*, 106:127–137, 1998.

[88] M. Könönen and J.V. Partanen. Blocking of EEG alpha activity during visual performance in healthy adults - a quantitative study. *Electroencephalography and Clinical Neurophysiology*, 87(3):164–166, 1993.

[89] L. De Lathauwer, B. De Moor, and J. Vandewalle. Fetal electrocardiogram extraction by blind source subspace separation. *IEEE Transactions on Biomedical Engineering*, 47(5):567–572, 2000.

[90] D.N. Lawley and A.E. Maxwell. *Factor analysis as a statistical method.* Butterworths, London, 2nd edition, 1971.

[91] J.S. Lee, D.S. Lee, J.Y. Ahn, G.J. Cheon, S.K. Kim, J.S. Yeo, K. Seo, K.S. Park, J.K. Chung, and M.C. Lee. Blind separation of cardiac components and extraction of input function from $H_2\,^{15}O$ dynamic myocardial PET using independent component analysis. *Journal of Nuclear Medicine*, 42(6):938–943, 2001.

[92] T.W. Lee. *Independent component analysis: theory and applications.* Kluwer Academic Publishers, 1998.

[93] T.W. Lee, M. Girolami, and T.J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[94] H. Liang. Adaptive independent component analysis of multichannel electrogastrograms. *Medical Engineering and Physics*, 23(2):91–97, 2001.

[95] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[96] A.K. Liu, J.W. Belliveau, and A.M. Dale. Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 95:8945–8950, 1998.

[97] O.V. Lounasmaa, M. Hämäläinen, R. Hari, and R. Salmelin. Information processing in the human brain: magnetoencephalographic approach. *Proceedings of the National Academy of Sciences of the United States of America*, 93(17):8809–8815, 1996.

[98] W. Lu and J.C. Rajapakse. Constrained independent component analysis. In *Advances in Neural Information Processing Systems*, volume 13, pages 570–576, 2000.

[99] D. Luenberger. *Optimization by vector space methods.* John Wiley & Sons, 1969.

[100] S. Maeda, S. Inagaki, H. Kawaguchi, and W.J. Song. Separation of signal and noise from in vivo optical recording in guinea pigs using independent component analysis. *Neuroscience Letters*, 302(2–3):137–140, 2001.

[101] D. Maino, A. Farusi, C. Baccigalupi, F. Perrotta, A.J. Banday, L. Bedini, C. Burigana, G. De Zotti, K.A. Gorski, and E. Salerno. All-sky astrophysical component separation with fast independent component analysis (FASTICA). *Monthly Notices of the Royal Astronomical Society*, 334(1):53–68, 2002.

[102] S. Makeig, A. Bell, T.P. Jung, and T.J. Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, 1996.

[103] A.N. Mamelak, N. Lopez, M. Akhtari, and W.W. Sutherling. Magnetoencephalography-directed surgery in patients with neocortical epilepsy. *Journal of Neurosurgery*, 97(4):865–873, 2002.

[104] A.M. Martoglio, J.W. Miskin, S.K. Smith, and D.J. MacKay. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, 18(12):1617–1624, 2002.

[105] J. Mathews and K. Fink. *Numerical methods using MATLAB*. Prentice–Hall, 3rd edition, 1999.

[106] K. Matsuura and Y. Okabe. A robust reconstruction of sparse biomagnetic sources. *IEEE Transactions on Biomedical Engineering*, 44(8):720–726, 1997.

[107] M.J. McKeown, S. Makeig, G.G. Brown, T.P. Jung, S.S. Kindermann, A.J. Bell, and T.J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998.

[108] M.J. McKeown and T.J. Sejnowski. Independent component analysis of fMRI data: examining the assumptions. *Human Brain Mapping*, 6(5–6):368–372, 1998.

[109] L. Medsker, S. Unadkat, S. Guruswami, and M. Ciocoiu. ICA applications in data mining. In *Artificial Intelligence and Soft Computing. Proceedings of the IASTED International Conference*, pages 294–298, Anaheim, USA, 2000.

[110] J. Meijs and M. Peters. The EEG and MEG: using a model of eccentric spheres to describe the head. *IEEE Transactions on Biomedical Engineering*, 34(12):913–920, 1987.

[111] A. Mohammad-Djafari. A Bayesian approach to source separation. In *The 19th International Workshop on Bayesian Inference and Maximum Entropy Methods (MaxEnt 99)*, Boise, Idaho, USA, 1999.

[112] J.E. Moody and L. Wu. What is the "true price"? – State space models for high frequency FX data. In *Proceedings of the Fourth International Conference on Neural Networks in the Capital Markets (NNCM'96), Decision Technologies for Financial Engineering*, pages 346–358, Singapore, 1997.

[113] J. Mosher and R. Leahy. Recursive MUSIC: a framework for EEG and MEG source localization. *IEEE Transactions on Biomedical Engineering*, 45(11):1342–1354, 1998.

[114] J. Mosher, R. Leahy, and P.S. Lewis. EEG and MEG: forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46(3):245–259, 1999.

[115] J. Mosher, P. Lewis, and R. Leahy. Multiple dipole modeling and localization from spatio-temporal MEG data. *IEEE Transactions on Biomedical Engineering*, 39(6):541–557, 1992.

[116] E. Moulines, J.F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume 5, pages 3617–3620, Munich, Germany, 1997.

[117] G. Müller and H.W. Gärtner. Imaging techniques in the analysis of brain function and behaviour. *Trends in Biotechnology*, 16(3):122–130, 1998.

[118] R. Näätänen, R.J. Ilmoniemi, and K. Alho. Magnetoencephalography in studies of human cognitive brain function. *Trends in Neurosciences*, 17(9):389–395, 1994.

[119] J.A. Nedler and R. Mead. A simplex method for function minimisation. *Computer Journal*, 7:308–313, 1965.

[120] G. Nolte and G. Curio. The effect of artifact rejection by signal-space projection on source localization accuracy in MEG measurements. *IEEE Transactions on Biomedical Engineering*, 46(4):400–408, 1999.

[121] P.L. Nunez. The brain's magnetic field: some effects of multiple sources on localization methods. *Electroencephalography and Clinical Neurophysiology*, 63(1):75–82, 1986.

[122] R. Ornstein and R.F. Thompson. *The amazing brain.* Houghton Mifflin, Boston, 1984.

[123] W.W. Jr Orrison. 3M Mayneord memorial lecture: functional brain imaging – an overview. *British Journal of Radiology*, 69(822):493–501, 1996.

[124] H. Otsubo and O.C. Snead. Magnetoencephalography and magnetic source imaging in children. *Journal of Child Neurology*, 16(4):227–235, 2001.

[125] A. Papoulis. *Probability, random variables, and stochastic processes.* McGraw-Hill, 1991.

[126] H.J. Park, J.J Kim, T. Youn, D.S. Lee, M.C. Lee, and J.S. Kwon. Independent component model for cognitive functions of multiple subjects using $[^{15}O]H_2O$ PET images. *Human Brain Mapping*, 18(4):284–295, 2003.

[127] H.M. Park, H.Y. Jeong, T.W. Lee, and S.Y. Lee. Sub-band-based blind signal separation for noisy speech recognition. *Electronic Letters*, 35(23):2011–2012, 1999.

[128] E. Pataraia, C. Baumgartner, G. Lindinger, and L. Deecke. Magnetoencephalography in presurgical epilepsy evaluation. *Neurosurgical Review*, 25(3):141–159, 2002.

[129] D.T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proceedings of the European Signal Processing Conference (EUSIPCO'92)*, pages 771–774, Brussels, Belgium, 1992.

[130] D.T. Pham and D. Karaboga. *Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks.* Springer, London, 2000.

[131] J. Phillips, R. Leahy, J. Mosher, and B. Timsari. Imaging neural activity using MEG and EEG. *IEEE Engineering in Medicine and Biology Magazine*, 16(3):34–42, 1997.

[132] Pitié Salpêtrière Hospital - Centre MEG–EEG. LENA UPR 640, 47 Bd de l'Hôpital, 75651 Paris CEDEX 13, France. <http://web.ccr.jussieu.fr/meg-center/>.

[133] Pitié Salpêtrière Hospital - MEG sensor noise. <http://web.ccr.jussieu.fr/meg-center/public/moyensag.html>.

[134] M.E. Raichle. Visualizing the mind. *Scientific American*, 270(4):58–64, 1994.

[135] M. Reite, P. Teale, and D.C. Rojas. Magnetoencephalography: applications in psychiatry. *Biological Psychiatry*, 12(1553–1563):1999, 45.

[136] T. Ristaniemi and J. Joutsensalo. On the performance of blind source separation in CDMA downlink. In *Proceedings of International Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 437–441, Aussois, France, 1999.

[137] D. Rose, E. Ducla-Soares, and S. Sato. Improved accuracy of MEG localization in the temporal region with inclusion of volume effects. *Brain Topography*, 1(3):175–181, 1989.

[138] M. Samonas, M. Petrou, and A.A. Ioannides. Identification and elimination of cardiac contribution in single-trial magnetoencephalographic signals. *IEEE Transactions on Biomedical Engineering*, 44(5):386–393, 1997.

[139] J. Sarvas. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Physics in Medicine and Biology*, 32(1):11–22, 1987.

[140] E. Schleussner, U. Schneider U, S. Kausch, C. Kahler, J. Haueisen, and H.J. Seewald. Fetal magnetoencephalography: a non-invasive method for the assessment of fetal neuronal maturation. *BJOG: an International Journal of Obstetrics and Gynaecology*, 108(12):1291–1294, 2001.

[141] T.A. Severini. *Likelihood methods in statistics.* Oxford University Press, 2000.

[142] E. Stern and D.A. Silbersweig. Advances in functional neuroimaging methodology for the study of brain systems underlying human neuropsychological function and dysfunction. *Journal of Clinical and Experimental Neuropsychology*, 23(1):3–18, 2001.

[143] M. Svensen, F. Kruggel, and H. Benali. ICA of fMRI group study data. *Neuroimage*, 16(3:1):551–563, 2002.

[144] S. Swithenby. SQUID magnetometers: uses in medicine. *Physics in Technology*, 18(1):17–24, 1987.

[145] R.F. Thompson. *The brain: a neuroscience primer.* Worth Publishers, New York, 3rd edition, 2000.

[146] J. Tiihonen, H. Katile, E. Pekkonen, I.P. Jääskeläinen, M. Huotilainen, H.J. Aronen, R.J. Ilmoniemi, P. Räsänen, J. Virtanen, E. Salli, and J. Karhu. Reversal of cerebral asymmetry in schizophrenia measured with magnetoencephalography. *Schizophrenia Research*, 30(3):209–219, 1998.

[147] S. Tong, A. Bezerianos, J. Paul, Y. Zhu, and N. Thakor. Removal of ECG interference from the EEG recordings in small animals using independent component analysis. *Journal of Neuroscience Methods*, 108(1):11–17, 2001.

[148] K. Torkkola. Blind separation of radio signals in fading channels. In *Advances in Neural Information Processing Systems*, volume 10, pages 756–762, 1997.

[149] T.M. Tu. Unsupervised signature extraction and separation in hyperspectral images: a noise-adjusted fast independent component analysis approach. *Optical Engineering*, 39(4):897–906, 2000.

[150] S. Ueno. Biomagnetic approaches to studying the brain. *IEEE Engineering in Medicine & Biology Magazine*, 18(3):108–120, 1999.

[151] M.A. Uusitalo and R.J. Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical & Biological Engineering & Computing*, 35(2):135–140, 1997.

[152] R. Vigário. Extraction of ocular artifacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology*, 103(3):395–404, 1997.

[153] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems*, volume 10, pages 229–235, 1998.

[154] R. Vigário and E. Oja. Independence: a new criterion for the analysis of the electromagnetic fields in the global brain? *Neural Networks*, 13(8–9):891–907, 2000.

[155] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, 2000.

[156] R. Vigário, J. Särelä, and E. Oja. Independent component analysis in wave decomposition of auditory evoked fields. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'98)*, pages 287–292, Skövde, Sweden, 1998.

[157] Z.S. Wang, J.Y. Cheung, and J.D. Chen. Blind separation of multichannel electrogastrograms using independent component analysis based on a neural network. *Medical and Biological Engineering and Computing*, 37:80–86, 1999.

[158] S.J. Williamson, G.L. Romani, L. Kaufman, and I. Modena, editors. *Biomagnetism: an interdisciplinary approach*, chapter Neurogenesis of evoked magnetic fields, pages 399–408. Plenum Press, New York, 1993.

[159] K.C. Yen and Y. Zhang. Adaptive co-channel speech separation and recognition. *IEEE Transactions in Speech and Audio Processing*, 7(2):138–151, 1999.

[160] F. Zappasodi, F. Tecchio, V. Pizzella, E. Cassetta, G.V. Romano, G. Filligoi, and P.M. Rossini. Detection of fetal auditory evoked responses by means of magnetoencephalography. *Brain Research*, 917(2):167–173, 2001.

[161] V. Zarzoso and A.K. Nandi. Noninvasive fetal electrocardiogram extraction: blind separation versus adaptive noise cancellation. *IEEE Transactions on Biomedical Engineering*, 48(1):12–18, 2001.

[162] X. Zhang and C.H. Chen. New independent component analysis method using higher order statistics with application to remote sensing images. *Optical Engineering*, 41(7):1717–1728, 2002.

[163] L. Zhukov, D. Weinstein, and C. Johnson. Independent component analysis for EEG source localization. *IEEE Engineering in Medicine and Biology*, 19:87–96, 2000.