UNIVERSITÀ DEGLI STUDI DI ROMA "LA SAPIENZA"
FACOLTÀ DI INGEGNERIA

Tesi di Dottorato di Ricerca in Ingegneria Elettronica

Blind Source Separation in real-world environments: new algorithms, applications and implementations

Separazione cieca di sorgenti in ambienti reali: nuovi algoritmi, applicazioni e implementazioni

Candidato: Giancarlo Valente

Tutor: Ing. Marco Balsi
Co-Tutor: Prof. Stefano Pisa

Ciclo XVIII - 2002/2005

To my parents
Giovanni and Anna Maria
and to my brother Fabio

Acknowledgements

I would like to thank my thesis advisor, Ing. Marco Balsi, for guiding me throughout all these years, with his insights, suggestions and invaluable advice. For this and for his patience and support, I will always be grateful. It has been a privilege to work with him.

I would like to express my sincere thanks to Prof. Valerio Cimagalli, for his detailed and constructive comments, for the fruitful discussions on my research and for his important support throughout this work.

My sincere thanks go to Prof. Elia Formisano and Prof. Rainer Goebel, for making my work in Maastricht University so pleasant and fruitful, for providing me with instruments, comments, suggestions and help in the amazing research on fMRI. My next thanks go to the people I have been working with during my stay in Maastricht; in particular, I would like to give special thanks to the Italian people up there: Federico De Martino, Francesco Gentile and Angela Manginelli.

I am grateful to the research group at "Ospedale Fatebenefratelli" in Rome, for the research made on MEG signals under the supervision of Dr. Franca Tecchio. It has been so interesting and stimulating working with all of them. In particular, my thanks go to Giulia Barbati, Sara Graziadio, Camillo Porcaro, Franca Tecchio and Filippo Zappasodi.

I am indebted to Prof. Patrizia Pantano and to the Department of Neuroscience of University "La Sapienza", for providing help, suggestions and instruments in fMRI data analysis.

My thanks go to people at "Centro Studi Giorgio Barzilai", and in particular to Prof. Alessandro Trifiletti and Dr. Pasquale Tommasino for providing help and support.

Working these years in the Electronic Engineering Department, I came across people that I consider friends rather than colleagues: Adriano, Enrico, Francesco, Giuliano, Lorenzo, Marta, Mauro, Nando.

Last, but not least, many many thanks to my friends, for supporting and helping me during these years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The present work is focused on the development of new Blind Source Separation (BSS) techniques for specific problems and on their implementation in embedded systems. The aim of such techniques is to extract in a "blind" fashion (i.e. without making specific assumptions) meaningful signals that have been mixed linearly, without knowing the original signals or the mixing coefficients. However, when we deal with real-world signals, we are never completely "blind", in that we do know (in a more or less detailed and quantitative way) some or their characteristic features. The "optimized" ICA algorithms presented in this work aim at enhancing separation of relevant signals by exploiting such *a priori* knowledge without renouncing the advantages of a substantially blind approach.

Independent Component Analysis (ICA) is a recently developed technique whose aim is to recover *statistically independent* signals that have been mixed linearly. Loosely speaking, consider a set of sources, i.e. independent signals, $\mathbf{s}$ and a mixing process that can be described in terms of a mixing matrix $\mathbf{A}$. The aim of ICA is to recover both $\mathbf{s}$ and $\mathbf{A}$ starting from the observation of the linear mixture

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

without making any particular assumptions other than statistical independence of the sources. In Chapter 2 the principles underlying ICA will be revised, in order to provide an inner sight to the problems that have been dealt with. In particular, theorems regarding identifiability, uniqueness and separability of the linear ICA model will be provided, together with the most common approaches to solve the problem, that are based on information theory or on non-linear decorrelation.

Since the need for fast independent component separation is crucial in some areas (e.g. telecommunications), in Chapter 4 the problem of fast independent component computation has been addressed, by designing optimized solutions on Digital Signal Processor (DSP) developing boards. This investigation has been done also to evaluate the possibility of embedding the ICA extraction unit into a dedicated processing block. In particular, floating and fixed point architecture DSP implementations have been explored, that may serve as the basis for a parallel (multi-threading) architecture.

Due to the generality of its formulation, ICA is applied in many and diverse research fields. In particular, in this work functional brain imaging and statistical electronic device modeling have been considered.. In Chapter 3 a set of selected ICA applications will be described in detail.

Statistical modeling of transistor is a crucial issue for yield oriented design and usually equivalent-circuit-based models (ECM) are employed. Principal Component Analysis has been proposed as suitable preprocessing for decorrelation and simplification of the representation before a Monte Carlo simulation of the model. However, ICA may be more suitable in representing data since it guarantees uncorrelatedness and *independence* of the parameters, that are not equivalent unless parameters have a Gaussian probability density. Therefore a population of High Electron Mobility Transistor (HEMT) Monolithic Microwave Integrated Circuit (MMIC) has been generated starting from the physical parameters of the active devices, and subsequently the model has been validated by means of statistical testing and of PCA and ICA. The results are described in Chapter 5.

Another investigation has involved brain imaging techniques. In particular functional Magnetic Resonance Imaging (fMRI) and MagnetoEncephaloGraphy MEG, that investigate inner brain activitiy by means of magnetic fields, together with analysis techniques, are discussed. Both fMRI and MEG experiments in cognitive neuroscience aim at individuating areas of the brain related to specific activities of the brain. Usually a subject performs a task (designed to investigate particular functionalities of cerebral areas), usually consisting of a block design of activity and rest, and subsequently data are analyzed in several ways in order to point out the functional areas involved in the execution of the task.

Since meaningful signals are embedded in unstructured noise, and other physiological signals, it is troublesome to extract them by simple inspection. Techniques employed for this purpose can be divided into *hypothesis-driven* and *data-driven.* While the first are employed in confirmatory analysis, the latter are mainly used for exploratory analysis; the differences between these two approaches are discussed in detail in Chapter 3. One of the most employed data-driven techniques is ICA, due to the assumption of independence among brain related activities, that therefore can be separated blindly by maximization their independence. In particular, it is applied in two different ways to fMRI data analysis, exploiting spatial or temporal independence.

The work of this thesis is focused on exploiting intrinsic structure of the sources for optimizing Blind Source Separation. In fact, in many applications where real-world data are involved, and in particular with respect to physiological signals, we do know that data display a regular structure in space and time, that is indeed used (most often heuristically by physicians) to judge about relevance of results of processing. Starting from these considerations, a new ICA approach has been developed to take prior information into account. The proposed methodology is based on the maximization of a modified (with respect to standard ICA) contrast function

$$F = J + \lambda H$$

where J is the contrast function whose maximization leads to independent components decomposition, while H accounts for the prior information of the sources. According to the weighting parameter λ it is possible to perform two different kinds of optimization: constrained (when it makes the additional term much higher than independence term) and multi-objective (when the two terms are comparable). Moreover, since prior information on the sources may also be described by a non-differentiable function, or even in a procedural way, the new contrast function F was optimized by means of simulated annealing, that does not require the use of derivatives, and performs *global optimization*, while gradient-based algorithms usually employed in independent component analysis only guarantee local optimization. This new methodology will be described in Chapter 6.

The proposed methodology has been applied successfully to fMRI time-series and MEG recordings, with different kinds of prior information. In particular, several constraints have been added to independent component separation for fMRI data analysis, regarding precise information about spatial and temporal features of the target independent sources, in a semi-blind fashion that allows selecting relevant sources first. This is an important advantage to save computation (and therefore time), and to reduce expert work in assessment of results. Moreover, spatio-temporal regularities have been successfully exploited using a multi-objective approach considering an additional contrast function related to spatial and temporal autocorrelation. This application is discussed in Chapter 7. Precise prior information on “interesting” signals have been also pointed out in MEG recordings, where a reactivity index has been considered to constrain the extracted sources. A new procedure (Functional Component Analysis) has been developed in order to extract more plausible sources by means of functional constraints. Moreover, a new technique not based on orthogonalization (Functional Source Separation) has been developed to deal with spatially overlapping sources. These techniques will be described in Chapter 8.

Chapter 2

Independent Component Analysis: General Principles

2.1 Introduction

A goal of many statistical techniques is to find a suitable representation of multivariate data. The new representation should allow to retrieve some information that is hidden in original data. Blind Source Separation techniques aim at retrieving some of this hidden information without making “strong” hypotheses on its nature.
In particular, Independent Component Analysis (**ICA**) looks for a linear representation of a collection of data such that the new signals are *maximally statistically independent*. This can be seen as an extension of Principal Components Analysis (**PCA**), that will be illustrated in detail in the following, as PCA is based only on *second order* statistics, while ICA uses information from all the statistics.
To give an explicit formulation of the problem, let us consider the *observed* data $\mathbf{X} \in \Re^{m \times n}$, where m is the dimension of the multidimensional vector we are observing, while n is the number of *realizations* we have of it. In other words, $\mathbf{X}$ is an m-dimensional vector and we have n observations.
In the case of a *linear* transformation of this dataset, we are considering a matrix $\mathbf{W} \in \Re^{q \times m}$, with $q \leq m$, such that the signals $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ have some desired property. While performing an ICA extraction, in particular, the q rows of $\mathbf{Y}$ are meant to be statistically independent.
The aim of these techniques is therefore to find a suitable linear transformation $\mathbf{W}$ according to the problem we are dealing with. In particular, if $q < n$ we are per-

forming dimension reduction, that may be useful in applications where the number of observed signals is greater than the "expected dimension" of the problem.
Several solutions to an ICA problem have been proposed in literature, based on different criteria. In the following of the chapter some of the proposed algorithms will be illustrated. In the next paragraph a well known example of an ICA problem will be shown.

2.2 An example: the Cocktail Party Problem

While dealing with Blind Source Separation (**BSS**), it is usual to address the problem by means of a famous example called the "Cocktail Party Problem". Let us assume we are in a room where a party is going on and people are having some conversations. Let us assume that two different conversations are held in different points of the room. Two microphones are placed in the room, in two different positions. Each of the two microphones will record both conversations, but, according to the distance between the microphone and the speaker, the two recordings will be different. The signal from each microphone will consist of a mixture of both conversations, but the two "observations", if the microphones are placed in different positions, will be different mixtures of *independent* signals.
The aim of ICA is to retrieve the original signals (i.e. the original conversations) using only the two recordings from the microphones, without any further assumption either on the nature of the sources or on the position of the microphones (i.e. without knowing the mixing process).
This simple example is particularly helpful in illustrating the nature of ICA problems, and furthermore helps addressing some of the main principles underlying the use of blind separation techniques. It is to be noted, in fact, that:

- The number of recordings is equal to the number of *independent* signals we want to retrieve (this is a case of *square mixing*).
- The microphones are supposed to be placed in different positions, i.e. the mixing matrix is not singular.
- The original signals are supposed to be independent. We have made the assumption that conversations that are held in *different* places are statistically *independent*.

Some of the issues raised by this examples are absolutely non trivial (like the assumption on independence), and will be addressed in the following sections, and a rigorous formulation of the problem will be given in the next sections.

2.3 Mathematical Background

Before giving a precise definition of Independent Component Analysis it may be useful to provide definitions of independence, uncorrelatedness and some properties of linear transformations.
In particular, as independence is defined by means of probability density functions, relations between second order and higher order statistics will be stressed. Moreover, some aspects and definitions from information theory, that will be used in the chapter, will be recalled.

2.3.1 Statistical Independence

Given m random variables $s_1, s_2, \ldots, s_m$, those variables are said to be independent *if and only if*:

$$p_{s_1,s_2,\ldots,s_m}(s_1, s_2, \ldots, s_m) = p_{s_1}(s_1) \cdot p_{s_2}(s_2) \cdot \ldots \cdot p_{s_m}(s_m) \tag{2.1}$$

that means that the *joint* probability distribution function (pdf) of those variables can be expressed as the product of their marginal densities. In other words, those m random variables are independent if the knowledge of any of them does not affect in any way the knowledge of any other. For instance,

$$p(s_2, \ldots, s_m \mid s_1) = p(s_2, \ldots, s_m). \tag{2.2}$$

Considering a transformation of a random vector $\mathbf{x}$ such that $\mathbf{y} = \mathbf{g}(\mathbf{x})$, and such that the inverse $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y})$ exists and is unique, it is possible to express the pdf of $\mathbf{y}$ in terms of the pdf of $\mathbf{x}$:

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\mid \det \mathbf{J}_{\mathbf{g}}(\mathbf{g}^{-1}(\mathbf{y})) \mid} p_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{y})) \tag{2.3}$$

where $\mathbf{J}_{\mathbf{g}}$ is the *Jacobian* matrix.
In the special case of a linear and nonsingular transformation such that $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, eq. (2.3) becomes

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\mid \det\mathbf{A} \mid} p_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y}) \tag{2.4}$$

2.3.2 Uncorrelatedness

A less strict requirement for a multivariate, that is limited to second order statistics, is uncorrelatedness. Starting from the *scalar* case, two random variables are uncorrelated if

$$\mathrm{E}\{(x - m_x)(y - m_y)\} = 0 \tag{2.5}$$

where $m_x = \mathrm{E}\{x\}$ and $m_y = \mathrm{E}\{y\}$.

Considering the multivariate case, the components of the vector $\mathbf{x} \in \Re^m$ are said to be uncorrelated if

$$\mathrm{E}\left\{(\mathbf{x} - \mathbf{m}_{\mathbf{x}})(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^T\right\} = \mathbf{D} \tag{2.6}$$

where $\mathbf{D}$ is a diagonal matrix whose nonzero elements are the variance of the corresponding components. As shown in equation (2.6), uncorrelatedness is related only to second order statistics, whereas independence is related to the entire pdf, as shown in (2.1). This difference is extremely important and will be pointed out in the following sections.

If two signals are independent, they will be uncorrelated. On the other hand, it is not true in general that two uncorrelated signals are independent. Uncorrelatedness is a necessary but not sufficient condition for independence, and decorrelation is often used as a first step during an ICA extraction procedure, together with variance normalization (*whitening*), as will be explained in section 2.6.

2.3.3 Multivariate Gaussian Variables

All the information on a multivariate random variable is “stored” in its probability density function, that, in many cases, cannot be evaluated easily. If the kind of distribution (e.g. Gaussian, Laplace, etc.) is known in advance, this evaluation is considerably faster and reliable, as the estimation is performed only on the parameters of the distribution. Among all “known” distributions, the Gaussian one plays a central role, and needs to be illustrated in detail.

Consider a Gaussian multivariate $\mathbf{x} \in \Re^m$. Let us consider its mean vector and covariance matrix, defined as:

$$\mathbf{m}_{\mathbf{x}} = \mathrm{E}\{\mathbf{x}\}, \qquad \mathbf{C}_{\mathbf{x}} = \mathrm{E}\left\{(\mathbf{x} - \mathbf{m}_{\mathbf{x}})(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^T\right\} \tag{2.7}$$

It is possible to express the probability density function in terms of $\mathbf{m}_{\mathbf{x}}$ and $\mathbf{C}_{\mathbf{x}}$:

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{C}_{\mathbf{x}})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_{\mathbf{x}})\, {\mathbf{C}_{\mathbf{x}}}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^T\right) \tag{2.8}$$

This means that higher order statistics (i.e. all the moments higher than 2) are unnecessary in describing a Gaussian variable, as it is fully "fully" explained by means of its mean and covariance matrix. All the other information contained in higher order statistics is already in the first two moments.

Another important property is that *a linear transformation of a Gaussian multivariate is still a Gaussian multivariate.* Let us consider a linear transformation $\mathbf{A}$ of $\mathbf{x}$: $\mathbf{y} = \mathbf{A}\mathbf{x}$, then $\mathbf{y}$ will still be a Gaussian multivariate, with mean $\mathbf{m}_{\mathbf{y}} = \mathbf{A}\mathbf{m}_{\mathbf{x}}$ and covariance matrix $\mathbf{C}_{\mathbf{y}} = \mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{A}^T$.

Moreover, Gaussian multivariate has an additional property: independence is *equivalent* to uncorrelatedness. This can be seen also from the fact that a Gaussian multivariate is described *only* by its mean vector and its covariance matrix, therefore there is no further information in higher order statistics.

2.3.4 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a well-known technique used for feature extraction and compression in multivariate data analysis since the early work of Pearson in 1901 [142]. Given a set of multivariate data, the aim is to find a smaller set of variables, with less redundancy, that however gives a good representation of data, under some suitable criterion. There is a close relationship between PCA and ICA, as the first accounts for second order statistics, while the second involves also higher order statistics.

PCA can be defined in several ways; the first formalization of a PCA criterion is from Hotelling in 1933 [75], and it is related to variance maximization.

Consider a multidimensional vector $x \in \Re^m$ and a linear transformation $\mathbf{w}_1$ such that

$$y_1 = \sum_{k=1}^{m} w_{k1} x_k = \mathbf{w}_1 \mathbf{x} \tag{2.9}$$

The term y_1 is called a *principal component* of $\mathbf{x}$ if its variance is *maximally* large. As the variance depends both on the norm and the orientation of the vector $\mathbf{w}$ and it increases unlimited as the norm increases, the constraint that the norm of $\mathbf{w}$ should be properly normalized (usually equal to 1). Therefore the problem becomes a constrained maximization:

$$\begin{aligned} J_1^{PCA}(\mathbf{w}_1) &= \mathrm{E}\{y_1^2\} = \mathrm{E}\{(\mathbf{w}_1^T\mathbf{x})^2\} = \\ &= \mathbf{w}_1^T \mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{w}_1 = \mathbf{w}_1 \mathbf{C}_{\mathbf{x}} \mathbf{w}_1 \qquad (2.10) \\ \text{with } \|\mathbf{w}_1\| &= 1 \qquad (2.11) \end{aligned}$$

where the norm operator $\|\cdot\|$ is the Euclidean norm operator:

$$\|\mathbf{w}_1\| = (\mathbf{w}_1^T\mathbf{w}_1)^{1/2} = (\sum_{k=1}^{m} w_{k1}^2)^{1/2} \qquad (2.12)$$

and the matrix $\mathbf{C}_{\mathbf{x}}$ denotes the *correlation* matrix of $\mathbf{x}$ (that corresponds to the *covariance* matrix in case of zero-mean). It is known from linear algebra ([48]) that the solution of the PCA problem is given in terms of unit-length eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ of the matrix $\mathbf{C}_{\mathbf{x}}$, that correspond to the eigenvalues $d_1, d_2, \ldots, d_m$, ordered such that $d_1 \geq d_2 \geq \ldots \geq d_m$. The solution that maximizes eq. (2.10) is the eigenvector $\mathbf{e}_1$ associated to the first eigenvalue, that is the one that "explains" the most variance.

It is possible to generalize the formulation of (2.10) and (2.11) to more dimensions, and in this case there is the further constraint of uncorrelatedness. In general, considering the n-th principal component $y_n = \mathbf{w}_n^T\mathbf{x}$, with $1 \leq n \leq m$, we have that its variance is maximized under the constraint that y_n is uncorrelated with all the previously found principal components:

$$\mathrm{E}\{y_n y_k\} = 0, \quad k < n \qquad (2.13)$$

Condition (2.13) can be expressed in terms of $\mathbf{w}_i$ and $\mathbf{x}$ as follows:

$$\mathrm{E}\{y_n y_k\} = \mathrm{E}\{(\mathbf{w}_n^T\mathbf{x})(\mathbf{w}_k^T\mathbf{x})\} = \mathbf{w}_n^T \mathbf{C}_{\mathbf{x}} \mathbf{w}_k = 0, \quad k < n \qquad (2.14)$$

Consider the *second* component: as it is already known that $\mathbf{w}_1 = \mathbf{e}_1$, the vector that explains the maximum variance must fulfill the following constraint:

$$\mathrm{E}\{y_2 y_1\} = \mathbf{w}_2^T \mathbf{C}_{\mathbf{x}} \mathbf{w}_1 = d_1 \mathbf{w}_2^T \mathbf{e}_1 = 0 \qquad (2.15)$$

therefore the solution is the *second* eigenvector. Recursively, it is easy to show that $\mathbf{w}_k = \mathbf{e}_k$; for further details see [80, 48].

It is possible to see PCA from the point of view of minimum mean-square error compression of original data $\mathbf{x} \in \Re^m$. Consider an orthonormal basis $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ (meaning that $\mathbf{w}_i^T \mathbf{w}_j = \delta_{ij}$, where δ_{ij} is the Kronecker's delta) with $n \leq m$. The projection of $\mathbf{x}$ on the subspace generated by $\mathbf{w}_i$ is $\sum_{i=1}^{n} (\mathbf{w}_i^T \mathbf{x}) \mathbf{w}_i$. The mean square error (MSE) criterion in this case becomes:

$$J_{MSE}^{PCA} = \mathrm{E}\{\|\mathbf{x} - \sum_{i=1}^{n} (\mathbf{w}_i^T \mathbf{x}) \mathbf{w}_i \|^2\} \tag{2.16}$$

The criterion, since the set of $\mathbf{w}_i$ is orthonormal, can be further written as:

$$\begin{aligned} J_{MSE}^{PCA} &= \mathrm{E}\{\|\mathbf{x}\|^2\} - \mathrm{E}\{\sum_{i=1}^{n} (\mathbf{w}_i^T \mathbf{x})^2\} = \\ &= \mathrm{trace}(\mathbf{C_x}) - \sum_{i=1}^{n} \mathbf{w}_j^T \mathbf{C_x} \mathbf{w}_j \end{aligned} \tag{2.17}$$

It can be shown (see [80]) that the minimum of the reconstruction error in (2.16) can be obtained considering *any* orthonormal basis of the subspace spanned by the *first* n eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ of the covariance matrix $\mathbf{C_x}$. It can also be shown that the value of the mean square error of (2.16) is:

$$J_{MSE}^{PCA} = \sum_{i=n+1}^{m} d_i \tag{2.18}$$

that means that the *minimum* achievable reconstruction error is related to the sum of the last eigenvalues that have not been considered in reconstruction: the more eigenvalues (i.e. the more dimensions for the compression subspace) one considers, the better the reconstruction becomes. However, as the eigenvalues are ordered according to their value, in most cases a large percentage (e.g. 95%) of the variance will be explained by a limited number of eigenvalues. The principle of dimension reduction lies in the previous observation: it is possible (of course, according to the redundancy of data) to reduce the dimensionality of the problem without affecting the effectiveness of the representation, within a given percentage. For this reason PCA is often used as a pre-processing step of independent component extraction, usually preserving a large percentage of original variance while reducing the number of dimensions of the problem. Consider the case where measured data $\mathbf{x} \in \Re^m$

are reduced by means of PCA to a dimension $n < m$, with a reconstruction error ϵ. If subsequently ICA is performed on this reduced data set, the reconstruction error of the new representation will be the *same*, as ICA perfoms a linear *invertible* transformation and therefore data are in the *same* subspace spanned by the first n principal components.
There is an intimate connection between PCA and ICA: while PCA is based on second order statistics, and is optimal in term of variance explained, ICA accounts for higher order statistics, thus it enriches the representation in terms of information theory. Consider now a Gaussian multivariate $\mathbf{x} \in \Re^m$: as seen in section 2.3.3, in this case uncorrelatedness is *equivalent* independence. Then, PCA will give a independent representation of data, as it gives uncorrelated components. There will be no improvement in further rotating the coordinates (i.e. looking for a linear transformation), as independence is already ensured by the uncorrelatedness constraint. This suggests that for a Gaussian multivariate it may be troublesome to look for independence by means of high order criteria. In sections 2.4 and 2.5 it will be clearer why Gaussian distributions among the sources are in contrast with the uniqueness and the identifiability of the ICA model.

2.3.5 Entropy of a random variable

Entropy $H(x)$ (often called “Differential Entropy”, in the continuous case) of a random variable $x \in \Re$ is defined as follows:

$$H(x) = -\int_{-\infty}^{+\infty} p(x) \ln p(x)\, dx \tag{2.19}$$

Differential entropy of a random variable can be interpreted as the degree of information that the observation of that variable may give. Consider for instance a variable whose probability density function is concentrated mainly in one point; it is evident to see that in this case its entropy will be rather small if compared with another one that has a pdf spread across a wider area. It is easy to generalize the definition of entropy to the multidimensional case, where, considering random vector $\mathbf{x} \in \Re^n$, we have:

$$H(\mathbf{x}) = -\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \mathbf{p}_{\mathbf{x}}(\boldsymbol{\xi}) \ln \mathbf{p}_{\mathbf{x}}(\boldsymbol{\xi}) d\boldsymbol{\xi}. \tag{2.20}$$

A remarkable property of entropy, that makes it extremely useful while dealing with ICA problems, is that a Gaussian variable has the *largest* entropy among *all* the random variables having the same mean and variance. The entropy of a Gaussian random vector $\mathbf{x}_{Gauss} \in \Re^m$ with covariance matrix $\mathbf{C}_\mathbf{x}$ can be evaluated as:

$$H(\mathbf{x}_{Gauss}) = \frac{1}{2} \log \mid \det(\mathbf{C}_\mathbf{x}) \mid + \frac{m}{2}[1 + \log 2\pi] \tag{2.21}$$

This is a particular result of the maximum entropy method [46], and its central role in some of the most widely used ICA method will be seen later, since non-Gaussianity leads to independence. Consider now a linear transformation of a random vector $\mathbf{x} \in \Re^m$:

$$\mathbf{y} = \mathbf{W}\mathbf{x}. \tag{2.22}$$

It is possible to characterize the entropy of the transformed variable $\mathbf{y}$ in terms of the entropy of $\mathbf{x}$

$$H(\mathbf{y}) = H(\mathbf{x}) + \log \mid \det(\mathbf{W}) \mid \tag{2.23}$$

For a rigorous proof, see [80].
Eq. (2.23) also shows that entropy is not *scale-invariant*. Consider a random variable $\mathbf{x}$ and multiply it by a scalar α. The entropy will change into:

$$H(\alpha x) = H(x) + \log \mid \alpha \mid \tag{2.24}$$

To overcome this problems, and using the property of maximum entropy of a Gaussian random vector, a modified version of entropy is often used. *Negentropy* $J(\mathbf{x})$ of random vector $\mathbf{x}$ is defined as the difference between the entropy of a Gaussian random vector with the same covariance matrix $\mathbf{C}_\mathbf{x}$ as $\mathbf{x}$ and the entropy of $\mathbf{x}$

$$J(\mathbf{x}) = H(\mathbf{x}_{Gauss}) - H(\mathbf{x}) \tag{2.25}$$

It is evident that negentropy is always nonnegative, and, due to the maximal entropy property, it is zero only when $\mathbf{x}$ has a Gaussian distribution. Moreover it is scale-invariant, and this makes negentropy particularly useful when dealing with ICA.

2.3.6 Mutual Information

Mutual Information is closely connected with entropy, and it can be defined as a measure of the information that some members of a set of random variables have on the other random variables in the set. According to [80], there are two ways of computing Mutual Information: by means of Entropy or by means of Kullback-Leibler divergence

Definition using Entropy

The mutual information between m scalar random variables $x_i, i = 1, 2, \ldots, n$ is defined as:

$$I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} H(x_i) \ - \ H(\mathbf{x}) \tag{2.26}$$

If the x_i are independent, it is straightforward to see that mutual information becomes zero.

Definition using Kullback-Leibler Divergence

Kullback-Leibler divergence is a sort of "distance" (even if it lacks the symmetry property, mandatory for any kind of distance) between two probability density functions. It is defined as:

$$\delta(p_1, p_2) = \int p_1(\xi) \log \frac{p_1(\xi)}{p_2(\xi)} d\xi \tag{2.27}$$

It has to be noted that K-L divergence is always nonnegative, due to the convexity of the negative logarithm and to Jensen's inequality [46]. Mutual information can be defined by means of K-L divergence, evaluating the "distance" in the following way:

$$I(x_1, x_2, \ldots, x_n) = \delta(p_{x_1, x_2, \ldots, x_n}, \prod_i p_{x_i}) \tag{2.28}$$

2.3.7 Maximum Likelihood Estimation

Maximum Likelihood (ML) estimation assumes that data are generated according to a predefined model with unknown parameters θ, and is based on the maximization of the likelihood function. In particular, the ML estimate θ_{ML} is the parameter vector θ that maximizes the probability of data, given the parameters:

$$\theta_{ML} = \arg\ \max_{\theta} p(\mathbf{x} \mid \theta) \tag{2.29}$$

As usually many probability densities contain exponential functions, usually a *log likelihood function* $\ln p(\mathbf{x} \mid \theta)$ is used. Due to the fact that natural logarithm is monotonic, all the maxima and minima of eq (2.29) will be preserved. Moreover, if different observation of random vector $\mathbf{x}$ are independent, then likelihood function factorizes into the product:

$$p(\mathbf{x} \mid \theta) = \prod_t p(x_t \mid \theta) \tag{2.30}$$

where t denotes the single realization of $\mathbf{x}$.
Of course, considering the logarithm of the likelihood function, the factorization in eq. (2.30) becomes a sum of marginal conditional probabilities. The maximum of a ML estimation, will be usually found from the solutions of the *likelihood equation*:

$$\left.\frac{\partial}{\partial\theta} \ln p(\mathbf{x} \mid \theta)\right|_{\theta=\theta_{ML}} = 0 \tag{2.31}$$

In the case of log-likelihood function, eq. (2.31) becomes a set of scalar equations. Unfortunately, the computational load for solving a ML estimation can be prohibitive, therefore various approximations are employed.

2.4 Linear ICA definition

Let us consider an observed multivariate random variable $\mathbf{x} \in \Re^m$. Assume the following statistical model:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon} \tag{2.32}$$

where $\mathbf{A} \in \Re^{m \times n}$, $\mathbf{s} \in \Re^n$ and $\boldsymbol{\epsilon} \in \Re^m$. Vector $\mathbf{s}$ is such that its components are statistically independent, while vector ϵ represents noise. The goal of ICA is to estimate both *mixing matrix* $\mathbf{A}$ and *independent components* $\mathbf{s}$ "blindly", meaning that no prior information on both sources and mixing, other than independence, is available (or is employed).
Since noise ϵ is assumed to have an unknown distribution, it can only be treated as a nuisance, and the ICA model we will consider becomes:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.33}$$

where noise is considered together with components, and $\mathbf{s}$ is the vector that *maximizes* a suitable measure of independence. According to m and n we have three different kind of ICA:

- If $m = n$, it is the case of *square ICA*

- If $m < n$ we are performing an *undercomplete ICA*
- If $m > n$ we are performing an *overcomplete*, or *underdetermined ICA*

It is possible now to give a definition of Independent Component Analysis of a random vector in *the square case*:

Definition 1. *The ICA of a random vector* $\mathbf{x} \in \Re^m$ *with finite covariance* $\mathbf{C_x}$ *is a pair* $\{\mathbf{F}, \mathbf{D}\}$ *of matrices such that:*

1. *The covariance matrix factorizes into:*
$$\mathbf{C_s} = \mathbf{F}\,\mathbf{D}^2\mathbf{F}^T \tag{2.34}$$
where $\mathbf{D}$ *is diagonal real positive and* $\mathbf{F}$ *is full column rank* m*;*
2. *The observation* $\mathbf{x}$ *can be written as*
$$\mathbf{x} = \mathbf{F}\mathbf{z} \tag{2.35}$$
where $\mathbf{z} \in \Re^m$ *has covariance matrix* $\mathbf{D}^2$ *and its component are* as much independent as possible*, in the sense that maximize a suitable contrast function that indicates independence.*

Since multiplying components by scalar values or changing their order does not affect their independence, it is evident that ICA defines an *equivalence class*, more than a *unique* solution. Therefore a suitable contrast function *must* account for this indeterminacy.

Property 2. *If a pair* $\{\mathbf{F}, \mathbf{D}\}$ *is a solution of an ICA problem, then so is the pair* $\{\mathbf{F}', \mathbf{D}'\}$*, where*
$$\mathbf{F}' = \mathbf{FSUP}, \qquad \mathbf{D}' = \mathbf{P}^T\mathbf{S}^{-1}\mathbf{DP} \tag{2.36}$$
where $\mathbf{S} \in \Re^{m\times m}$ *is a diagonal real positive scaling matrix,* $\mathbf{U} \in \Re^{m\times m}$ *is a diagonal matrix with entries of unit modulus and* $\mathbf{P} \in \Re^{m\times m}$ *is a permutation matrix.*

As clearly stated in Property 2, it is never possible to have a *unique* solution to an ICA problem: a permutation and a scaling of the components will not affect independence, nor it will be possible to define an ordering for the components (unlike in PCA, where components are ordered according to the variance explained). However, within the equivalence class, if at most one component has a Gaussian distribution, the solution is unique. To show this property, it is necessary to define a contrast function:

Definition 3. *A contrast is a mapping H from a set of densities $\{p_{\mathbf{y}}, \mathbf{y} \in \Re^m\}$ to $\Re$ satisfying the following requirements:*

- *$H(p_{\mathbf{y}})$ does not change if the components y_i are permuted:*

$$H(P_{p_{\mathbf{y}}}) = H(p_{\mathbf{y}}), \qquad \forall P \, permutation$$

- *H is invariant by scale change:*

$$H(p_{\mathbf{Sy}}) = H(p_{\mathbf{y}}), \qquad \forall S \, diagonal\ invertible.$$

If $\mathbf{y}$ has independent components, then

$$H(p_{\mathbf{Ay}}) \leq H(p_{\mathbf{y}}), \qquad \forall A \, invertible.$$

Definition 4. *A contrast H will be said to be discriminating over a set $\mathbf{E}$ if equality $H(p_{\mathbf{A}y}) = H(p_y)$ holds only when $\mathbf{A}$ is of the form $\mathbf{A} = \mathbf{SP}$, as defined in Property 2, when $\mathbf{y}$ is a random vector of $\mathbf{E}$ having independent components.*

To show that if at most one component must be Gaussian in order to have uniqueness of the solution, we will use the following theorem:

Theorem 1. *Let $\mathbf{y}$ be a vector with independent components, of which at most one is Gaussian, and whose densities are not reduced to a point-like mass. Let $\mathbf{C} \in \Re^{m \times m}$ be an orthogonal matrix and $\mathbf{z}$ the vector $\mathbf{z} = \mathbf{Cy}$. Then the following three properties are equivalent:*

(i) *The components y_i are pairwise independent*

(ii) *The components y_i are mutually independent*

(iii) *$\mathbf{C} = \mathbf{SP}$, with $\mathbf{S}$ diagonal and $\mathbf{P}$ permutation.*

For a proof of this theorem, see [45].
In the following sections it will be clearer why Gaussianity of more than one component makes it impossible to separate univocally sources. As shown in section 2.3.3, a linear combination of Gaussian random variables is still Gaussian, therefore a contrast based on the pdf will not be able to discriminate effectively between two different linear combinations of Gaussian signals. This is why, in the classical ICA

model, it is assumed that no more than one source has a Gaussian pdf, even if in real problems this limitation is not so strict.
To reduce indeterminacy, at least in scale, it is usually imposed that the columns of $\mathbf{F}$ have unit norm. This is not such a strict constraint, and will be extremely useful while dealing with the optimization algorithms employed.

To sum up, independent component analysis aims at finding a linear decomposition of an observed vector $\mathbf{x}$ into a set of independent components, without prior knowledge on the mixing matrix or on the original signals. Recalling eq. (2.33), the problem becomes finding a suitable $\mathbf{W} \in \Re^{m \times n}$ such that vector $\mathbf{y}$:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{2.37}$$

has maximally mutually independent components .
From now on, the following notation will be considered:

- $\mathbf{x}$: observed signals

- $\mathbf{s}$: independent components (to estimate)

- $\mathbf{A}$ and $\mathbf{W}$: respectively *mixing* and *unmixing* matrix

- $\mathbf{y}$: estimate of the independent components

- $\mathbf{z}$: whitened observed signals (will be explained in section 2.6)

2.5 Conditions for linear ICA model

In this section the concepts of indentifiability, separability and uniqueness for the ICA model will be investigated, both for the case where the mixing is *square* (same sources as observations) and for the *overcomplete* case (more sources than observations).
Let us recall linear ICA model presented in eq. (2.33), that is:

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

where $\mathbf{x} \in \Re^m$, $\mathbf{s} \in \Re^n$ and $\mathbf{A} \in \Re^{m \times n}$. We define a *representation* of $\mathbf{x}$ as the couple $(\mathbf{A}, \mathbf{s})$. In the following the section we will refer to *reduced* representations, that are those where columns of the mixing matrix are not *pairwise* linearly dependent (this is done to ensure uniqueness, and it means also that, if we have two different

representations, any column of the first mixing matrix will be linearly dependent on only *one* column of the other mixing matrix, or on no one at all [51]). It is easy to show that the noisy ICA model (2.32), where the noise is a multivariate Gaussian, can be seen as special case of model presented in (2.33), where we have additional sources (related to noise) in the model [51, 160].
It is possible now to define the concepts of identifiability, uniqueness and separability for an ICA model, without reducing to the square mixing case:

Definition 5. *The ICA model in* (2.33) *is said to be* ***identifiable*** *if in every reduced representations* $(\mathbf{A}, \mathbf{s})$ *and* $(\mathbf{A}_1, \mathbf{s}_1)$ *of* $\mathbf{x}$*, every column of* $\mathbf{A}$ *is linearly dependent on some column of* $\mathbf{A}_1$ *and vice versa.*

Definition 6. *The model in* (2.33) *is* ***unique*** *if it is identifiable* and *sources* $\mathbf{s}$ *and* $\mathbf{s}_1$ *have the same distribution for some permutation and changes of scale or position.*

Definition 7. *The model in* (2.33) *is said to be* ***separable*** *if for every* $\mathbf{W} \in \Re^{m \times m}$ *such that* $\mathbf{W}\mathbf{x}$ *has* m *independent components, we have that* $\mathbf{\Lambda}_m \mathbf{P}_m \mathbf{s} = \mathbf{W}\mathbf{x}$*, for some block diagonal matrix* $\mathbf{\Lambda}_m$ *with nonzero elements on main diagonal and permutation matrix* $\mathbf{P}_m$*. Moreover, matrix* $\mathbf{W}$ *has to always exist.*

In the next sections some characterizing theorems will be provided in a general approach, including the overcomplete case. In section 2.4 it has been shown that, if there is at most one source with Gaussian probability density then the model is identifiable. The generalization to the overcomplete case will give some further conditions for those properties. It may be useful, however, to recall some theorems presented in [51] and [160], that will be useful in the following.

Theorem 2. *Let* $(\mathbf{A}, \mathbf{s})$ *and* $(\mathbf{A}_1, \mathbf{s}_1)$ *be two representations of an* n*-dimensional vector* $\mathbf{x}$*, where* $\mathbf{A}$ *and* $\mathbf{A}_1$ *are constant matrices of dimensions* $n \times k_1$ *and* $n \times k_2$*, and* $\mathbf{s} \in \Re^{k_1}$ *and* $\mathbf{s}_1 \in \Re^{k_2}$ *are vectors with independent components. Then the following properties hold:*

i) If the i*-th column of* $\mathbf{A}$ *is not linearly dependent on any column of* $\mathbf{A}_1$*, then* $\mathbf{s}_i$ *has a Gaussian pdf.*

ii) If the i*-th column of* $\mathbf{A}$ *is linearly dependent on the* j*-th column of* $\mathbf{A}_1$*, then the logarithm of the characteristic function of* $\mathbf{s}$ *and* $\mathbf{s}_1$ *differ by a polynomial in a neighborhood of the origin.*

The *characteristic function* $\psi_{\mathbf{x}}(\mathbf{u})$ of a random vector $\mathbf{x}$ (also known as *first characteristic function* or *moment generating function*) is defined as follows:

$$\psi_{\mathbf{x}}(\mathbf{u}) = \mathrm{E}\{e^{(j\mathbf{u}^T\mathbf{x})}\} \tag{2.38}$$

and in the case of ICA generative model (2.33), it can be expressed as the product of the characteristic functions of the sources [160]:

$$\psi_{\mathbf{x}}(\mathbf{u}) = \prod_{i=1}^{k} \psi_{s_i}(\mathbf{a}_i^T\mathbf{u}) \tag{2.39}$$

It has to be noted that the characteristic function of a Gaussian random variable is a polynomial of degree 2 [151].
To reduce the ambiguity presented in Theorem 2, it is possible to introduce some structural constraints on representations. We introduce the (columnwise) Kathri-Rao product $\odot$ on matrices defined as matrix *columnwise* Kronecker product $\otimes$. Therefore $\mathbf{A} \odot \mathbf{B} = (\alpha_1 \otimes \beta_1, \ldots, \alpha_n \otimes \beta_n)$, where α_i are the columns of $\mathbf{A}$ and β_i are the columns of $\mathbf{B}$, with the notation $(\mathbf{A}\odot)^q\mathbf{A}$ meaning $\mathbf{A} \odot \ldots \odot \mathbf{A}$, including $\odot$ q times. Using this notation it is possible to state the following theorem:

Theorem 3. *Let random vector $\mathbf{x} \in \Re^n$ with nonvanishing characteristic function have a representation $(\mathbf{A}, \mathbf{s})$, where $\mathbf{A}$ is a known $n \times m$ matrix and let q be the integer such that the* $\mathrm{rank}[(\mathbf{A}\odot)^q\mathbf{A}] = m > \mathrm{rank}[(\mathbf{A}\odot)^{q-1}\mathbf{A}]$. *Then the characteristic function of each component s_i of $\mathbf{s}$ is determined up to a factor* $\exp(\mathcal{P}_{i,q}(t))$, *where $\mathcal{P}_{i,q}(t)$ is a polynomial of degree at most q.*

In the following sections some theorems regarding indentifiability, separability and uniqueness will be provided. For a proof of these theorems, see [51] and its references.

2.5.1 Identifiability

As seen in previous section, the concept of identifiability is related to the uniqueness of the mixing matrix, obviously up to permutation and scaling. We have seen in section 2.4 that for the square case, we have identifiability if at most one independent component has a Gaussian probability density function. It is possible to generalize Theorem 1 to the overcomplete case: Theorem 4 states some conditions for identifiability:

Theorem 4. *The model* $\mathbf{x} = \mathbf{A}\mathbf{s}$ *is identifiable among all representations* $(\mathbf{A}_1, \mathbf{s}_1)$ *of* $\mathbf{x}$ *that:*

i) do not contain any Gaussian source

ii) $\mathbf{A}_1$ *is of full column rank and at most one source has a Gaussian pdf.*

This theorem can be seen as a generalization of Theorem 1, and states once more that Gaussianity of sources can be in most cases an obstacle for blind source separation. Consider for instance a Gaussian multivariate with independent components. As every orthogonal rotation applied to data does not change independence of data, there will be no unique mixing matrix (up to permutation or scaling), and therefore the model will not be identifiable (and subsequently not unique, as seen in Definition 6).
As example of application of the theorem, consider two non-Gaussian signals s_1 and s_2 and two Gaussian signals n_1 and n_2. Consider the following mixing model:

$$\begin{aligned}\mathbf{x} &= \begin{pmatrix}1 & 1 & 0\\ 1 & 0 & 1\end{pmatrix}\begin{pmatrix}s_1\\ s_2+2n_1\\ 2n_2\end{pmatrix} = \begin{pmatrix}s_1+s_2+2n_1\\ s_1+2n_2\end{pmatrix} = \\ &= \begin{pmatrix}1 & 1 & 1\\ 1 & 0 & -1\end{pmatrix}\begin{pmatrix}s_1+n_1+n_2\\ s_2\\ n_1-n_2\end{pmatrix} \qquad (2.40)\end{aligned}$$

Here the model is clearly unidentifiable, since the mixing matrix is not unique. This is due to the fact that the representation has a Gaussian component, thus condition (i) of theorem 4 is not fulfilled, and the mixing matrix is not full column rank, meaning that also condition (ii) is not verified.
This shows also that, while in the square case, one component with Gaussian pdf is allowed, in the *overcomplete* case it is *mandatory* not to have a Gaussian source for identifiability to hold, as it will never be possible to have a full-column-mixing matrix.
It should be clear that if an ICA model is identifiable, this does not mean that the other two characteristics (uniqueness and separability) will automatically hold for the model. An interesting example on this considerations is the following: consider four *non Gaussian* signals $s_i, i = 1 \dots 4$ and consider two standard *Gaussian* and

independent signals n_1 and n_2. Now consider the following mixing model:

$$\begin{aligned}\mathbf{x} &= \begin{pmatrix}1 & 0 & 1 & 1\\ 0 & 1 & 1 & -1\end{pmatrix}\begin{pmatrix}s_1\\ s_2\\ s_3+n_1\\ s_4+n_2\end{pmatrix} = \begin{pmatrix}s_1+s_3+s_4+n_1+n_2\\ s_2+s_3-s_4+n_1-n_2\end{pmatrix} = \\ &= \begin{pmatrix}1 & 0 & 1 & 1\\ 0 & 1 & 1 & -1\end{pmatrix}\begin{pmatrix}s_1+n_1+n_2\\ s_2+n_1-n_2\\ s_3\\ s_4\end{pmatrix} \qquad (2.41)\end{aligned}$$

As n_1 and n_2 are Gaussian and independent, also n_1+n_2 and n_1-n_2 are independent. The example shows that the model is identifiable, as the mixing matrix is unique up to permutation and scaling (This is guaranteed by Theorem 4 (i), because no source has a Gaussian pdf, even if some of the sources have a *normal component*). However, it is clear that the model is not unique, as there are different sources for the same mixing model. A source for the mixing model in (2.33) is said to *have* a normal component if it is of the form $s+n$, where only n has a Gaussian pdf.

2.5.2 Separability

As seen in Definition 7, separability concept is related to the recovery of the sources. The following theorem is based on what presented in [45] and generalized in [51].

Theorem 5. *An ICA model* $\mathbf{x} = \mathbf{As}$ *is separable if and only if the mixing matrix* $\mathbf{A}$ *is of full column rank and at most one source variable has a Gaussian pdf.*

Useful considerations can be taken from the Theorem 5. Consider the two aspects presented for separability to hold: mixing matrix being full-column rank and no more than one source with Gaussian pdf. Consider now the two cases of square and overcomplete ICA. In the first we know that, for the model to be identifiable, non Gaussianity of *all* the sources is required, and, in the case of Gaussianity of only one source, the model is still identifiable if the mixing matrix is of full column rank. Therefore, in the square case identifiability and separability are both present in the case of non-singular mixing, while in the case of singular mixing, the model is still identifiable but not separable.

Consider now the overcomplete case: in this case the model will *never* be separable, as it is not possible for the mixing matrix to be of full column rank. Therefore there is still chance for an overcomplete model to be identifiable, but not separable (that means that there in no *unique* way of decomposing data, up to permutation and scaling, in order to recover the sources).

2.5.3 Uniqueness

In sections 2.5.1 and 2.5.2 we have seen how to characterize a linear model in terms of uniqueness of the mixing matrix and of possibility of recovering the sources. Moreover, it has been shown that in the overcomplete case it is not possible to recover the sources. However it could be possible to determine the distribution of the sources, that means that the model is still unique (see Definition 6). Then it would be possible to recover the sources in probabilistic sense (that is, using a Maximum Likelihood approach, used in [103]). To ensure the uniqueness of an overcomplete ICA model, the following Theorem proposed in [160] and generalized in [51] will be presented.

Theorem 6. *The ICA model* $\mathbf{x} = \mathbf{A}\mathbf{s}$ *is unique if any of the following properties hold.*

i) The model is separable.

ii) All the cumulative functions (c.f.) of the sources are analytic (or are nonvanishing) and none of the c.f.'s has an exponential factor with a polynomial of degree at least 2.

iii) All the sources have non-Gaussian probability distribution with nonvanishing c.f.'s and $\operatorname{rank}[\mathbf{A} \odot \mathbf{A}] = m$.

iv) All the sources have nonvanishing c.f.'s ϕ *such that they cannot be expressed as* $\phi = \varphi \exp(\mathcal{P})$*, where* $\mathcal{P}$ *is a polynomial of degree* n*, with* $1 < n \leq q$ *and* $\operatorname{rank}[(\mathbf{A}\odot)^q\mathbf{A}] = m > \operatorname{rank}[(\mathbf{A}\odot)^{q-1}\mathbf{A}]$.

A characteristic function is analytic when the moment generating function exists (that is, when all the moments exist). Moreover, part (*ii*) of the Theorem could be reformulated by substituting the requirement for the degree of the polynomial with the one of non-Gaussianity of the sources. For requirement (*ii*) the number of the sources is unlimited, while for parts (*iii*) and (*iv*) the number is limited up to a value related to the number of observations p ([51]).

2.6 Standard ICA preprocessing

Before performing an independent component analysis, usually some steps are taken in order to improve performances of extraction. They are basically two: *centering* and *whitening*.

As in the model data are assumed to be zero-mean, it is necessary to remove the input data mean, that can be added at the end of the extraction procedure. Consider the (centered) vector $\tilde{\mathbf{x}} = \mathbf{x} - \mathrm{E}\{\mathbf{x}\}$ and the eq. (2.37). The independent components of the new vector $\tilde{\mathbf{x}}$ are $\tilde{\mathbf{s}} = \mathbf{W}\tilde{\mathbf{x}}$. The relation between $\mathbf{s}$ and $\tilde{\mathbf{s}}$ is straightforward: $\mathbf{s} = \tilde{\mathbf{s}} + \mathbf{W}\mathrm{E}\{\mathbf{x}\}$, thus to perform ICA data mean is removed, and after extraction it is added back (only in the cases where the mean matters). The other step of preprocessing, namely whitening, is performed in most algorithms, and even in those that do not require sphered data it may me useful to improve performances.

Whitening, or sphering, data vector $\mathbf{x}$ means linearly transforming it into a new vector $\mathbf{z}$ such that its covariance matrix equals unity matrix.

$$\mathbf{z} = \mathbf{V}\mathbf{x} \tag{2.42}$$

where $\mathbf{V}$ is called the *whitening* matrix.

Since the covariance matrix $\mathbf{C}_{\mathbf{x}} = \mathrm{E}\{\mathbf{x}\mathbf{x}^T\}$ is positive semi-definite (and for most cases where the components are greater or equal to the number of sources, positive), it can always be decomposed by means of its eigenvalues, namely:

$$\mathbf{C}_{\mathbf{x}} = \mathbf{E}\mathbf{D}\mathbf{E}^T \tag{2.43}$$

where $\mathbf{D} = \mathrm{diag}(d_1, d_2, \cdots, d_m)$ is a diagonal matrix with the eigenvalues on the main diagonal, and $\mathbf{E}$ contains on its columns the eigenvectors related to those eigenvalues. Then a linear whitening transform is given by:

$$\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T \tag{2.44}$$

and it is easy to show why. Consider the transformed data $\mathbf{z}$ and compute its covariance matrix $\mathbf{C}_{\mathbf{z}}$:

$$\mathrm{E}\{\mathbf{z}\mathbf{z}^T\} = \mathbf{V}\mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{V}^T = \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{E}\mathbf{D}\mathbf{E}^T\mathbf{E}\mathbf{D}^{-1/2} = \mathbf{I} \tag{2.45}$$

as matrix $\mathbf{E}$ is orthonormal by definition.

It has to be noted that any orthonormal transformation of matrix $\mathbf{V}$ is still a whitening matrix. It can be shown easily considering the matrix $\mathbf{U}\mathbf{V}$ and evaluating the

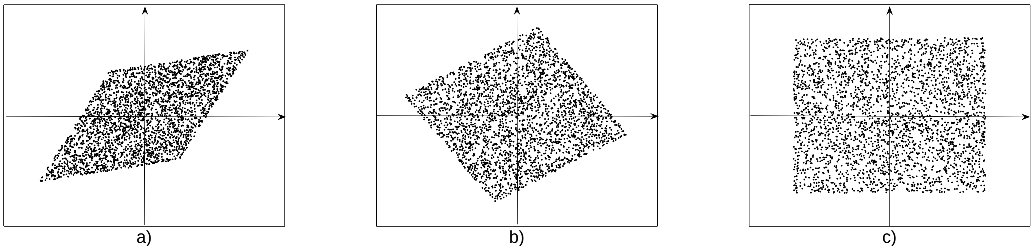


Figure 2.1: Example of whitening and ICA of a bidimensional random vector. **a**) Raw bidimensional data, **b**) whitened data, **c**) independent components.

covariance matrix of $\mathbf{z} = \mathbf{UVx}$

$$\mathrm{E}\{\mathbf{z}\mathbf{z}^T\} = \mathbf{UV}\mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{V}^T\mathbf{U}^T = \mathbf{UIU}^T = \mathbf{I} \tag{2.46}$$

It is evident that, at the end of whitening process, vector $\mathbf{z}$ will have uncorrelated components. As shown in section 2.3.2, this does not mean that its components will be independent. However the search for independent components, while starting from whitened data, will consist just of a suitable orthogonal transformation of whitened data, as we have assumed data to have unit norm.

To better clarify this concept, it may be useful considering a simple example involving a two dimensional random variable, whose marginal densities are uniform. In left panel of fig. 2.1 raw data are plotted. Marginal densities of the bidimensional vector are obtained considering data on X and Y axes. It is evident that the joint density is not the product of the marginal densities (this can be also seen e.g. by observing that knowing that x has a value in the left part of the two-dimensional plot, gives a lot of information on possible y value). In the center panel the whitened signals $\mathbf{z}$ are depicted. It is to be noted that in this case variance is normalized for each signal, and they are uncorrelated, but still they are not independent. In the last panel independent decomposition of original data is depicted. Now it is evident how the knowledge on one variable cannot any more influence the knowledge on the other one.

To sum up, whitening performs a rotation and a scaling of the original signals, such that they are uncorrelated and have unit variance. There are infinite ways of performing whitening, as shown in eq. (2.46), but *only one* is the overall transformation that leads to independence.

There are some cases, especially while dealing with structured data like time-series, where some kind of filtering before independent component separation may be useful. As a linear filter in the discrete domain can be obtained by multiplying data by a suitable matrix, it is always possible to pre-filter data without altering the model properties. In fact, considering the data matrix $\mathbf{X}$ whose columns are different time points of the observation (and thus different realizations of the multidimensional random vector), the ICA model can be expressed as follows:

$$\mathbf{X} = \mathbf{AS} \tag{2.47}$$

Filtering of $\mathbf{X}$ corresponds to multiplying $\mathbf{X}$ *from the right* by a matrix, let us call it $\mathbf{M}$. This gives:

$$\mathbf{X}^* = \mathbf{XM} = \mathbf{ASM} = \mathbf{AS}^* \tag{2.48}$$

that means that the components are filtered by the same filtering that was applied on the mixtures. They are not mixed up in $\mathbf{S}$, however, because the matrix $\mathbf{M}$ is, by definition, a component-wise filtering matrix. Since the mixing matrix does not change, it is possible to use the filtered data to perform ICA estimation, and then use the *same* mixing matrix on the original data to obtain the components.

2.7 Principles for solving an ICA problem

As shown in section 2.4, solving an ICA problem basically means finding a matrix $\mathbf{W}$ such that $\mathbf{y} = \mathbf{Wx}$ has components maximally statistically independent. As usually it is rather difficult to estimate the joint probability density function, several ways to estimate independence are employed. In general, to find an independent decomposition of original data, an *objective function* that denotes independence is defined, and an optimization procedure is chosen to maximize or minimize it. Both the objective function and the optimization procedure account for the overall performances of the ICA separation. In particular:

- The statistical properties, like consistency, asymptotic variance and robustness, depend on the choice of the objective function
- the algorithmic properties, like convergence speed, memory requirement and numerical stability, depend on the optimization algorithm

These two classes of properties are in most cases independent, meaning that different optimization procedures may be employed in order to optimize a given objective function.

Focusing now on the objective functions, it is to be noted that there are mainly two classes of functions, those who estimate all the components together, and those that estimate one component at a time. Even if the first class seems to be more "related" to the problem of estimating independence (as, obviously, independence is defined *among* all the components), nonetheless the one-unit contrast functions have some appealing properties. There are some advantages in using one-unit contrast functions:

- In many applications, it is not necessary to estimate all the components. In particular, if the maximization (or minimization) is carried out by means of some global optimization procedure, extraction can be halted at a certain point without considering *all* the components.
- It is not necessary to have a prior knowledge on the number of independent sources in the problem, since the independent components can be estimated one at a time.
- In some cases evaluation of the contrast function may be rather less computationally demanding, in terms of memory usage.

As previously stated, the intrinsic difficulty in directly estimating independence, leads to the use of different criteria to achieve independence. Many of the objective functions proposed in literature start from information theoretic criteria, or from non-linear decorrelation principles. In the following sections, several objective functions will be shown, starting from the multi-unit ones in sections from 2.7.1, to 2.7.5 and moving to one-unit ones in section 2.7.6.

2.7.1 Information Maximization (INFOMAX)

One of the most used approach, is based on a neural network implementation. As show in 2.3.6, minimization of mutual information is a very "natural" way of looking for independence. However the estimate of mutual information has the drawback of needing the estimation of probability densities functions, that may be sometimes too complex. To overcome this problem, a neural network approach has been proposed in [24] where, by maximizing the differential entropy of the output, it is possible to minimize mutual information, and thus to achieve independence. To simplify, consider only a bi-dimensional vector $\mathbf{y} = (y_1, y_2)$, that is the output of a neural network with input $\mathbf{x}$. The joint differential entropy (section 2.3.5) of the output $\mathbf{y}$, can be expressed in the following way:

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2) \tag{2.49}$$

As seen in 2.3.6, if two random variables y_1 and y_2 are independent, their mutual information is zero, otherwise it is positive. Therefore, by minimizing mutual information between the output of the neural network one maximizes network output entropy. According to [24], output entropy of the network, whose nonlinearity is $g(x)$, is obtained when the high density parts of probability density function of x are aligned with the high sloping parts of nonlinearity $g(x)$. In particular, a logistic transfer function is chosen:

$$g(x) = \frac{1}{1+e^{-u}}, \quad u = wx + w_0 \tag{2.50}$$

Thus output entropy maximization is performed by means a stochastic gradient optimization, which leads to the learning rules for the two parameters of the transfer function w and w_0:

$$\begin{aligned} \Delta w &\propto \frac{1}{w} + x(1-2y), & (2.51) \\ \Delta w_0 &\propto x(1-2y) & (2.52) \end{aligned}$$

Generalization to the multidimensional case is straightforward, and considering the input $\mathbf{x}$ and the output $\mathbf{y}$ of a network whose transfer function is a multidimensional sigmoid $\mathbf{y} = \mathbf{g}(\mathbf{u}), \mathbf{u} = \mathbf{W}\mathbf{x} + \mathbf{w_0}$, we have the following learning rules:

$$\begin{aligned} \Delta \mathbf{W} &\propto [\mathbf{W}^T]^{-1} + (\mathbf{1} - 2\mathbf{y})\mathbf{x}^T & (2.53) \\ \Delta \mathbf{w_0} &\propto \mathbf{1} - 2\mathbf{y} & (2.54) \end{aligned}$$

There are some cases, however, where INOFMAX algorithm, in the version presented in (2.53) and (2.54), fails to recover the sources. This happens especially in the case of presence of *sub-Gaussian* sources. To overcome this drawback, in [102] an extension of the algorithm has been proposed, based on maximum likelihood estimation, that will be treated in section 2.7.2

2.7.2 Maximum Likelihood Methods

Maximum Likelihood (**ML**) estimation is a powerful tool widely used in statistical estimation, and ML approaches have been employed in ICA literature, as well. The first approaches using ML principles can be found in [66, 145]. ML estimation has some desirable properties in terms of consistency and asymptotic efficiency, but its effectiveness rely on the choice of the parameters of the estimate. To illustrate this aspect, it may be useful to remind eq. (2.33), and, by means of eq. (2.4), to derive the likelihood for an ICA model:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.55}$$

Considering *unmixing* matrix $\mathbf{W}$ and eq. (2.4), it is possible to express the pdf of $\mathbf{x}$ in terms of the pdf of $\mathbf{y}$:

$$p_{\mathbf{x}}(\mathbf{x}) = \mid \det \mathbf{W} \mid p_{\mathbf{s}}(\mathbf{s}) = \mid \det \mathbf{W} \mid \prod_i p_i(s_i) \tag{2.56}$$

where p_i denotes the densities of the independent components. Eq. (2.56) can be expressed in terms of $\mathbf{x}$ and $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_m)^T$, giving:

$$p_{\mathbf{x}}(\mathbf{x}) = \mid \det \mathbf{W} \mid \prod_i p_i(\mathbf{w}_i^T \mathbf{x}) \tag{2.57}$$

Assuming that we there are T observations available, then using eq. (2.30), the likelihood function becomes:

$$L(\mathbf{W}) = \prod_{t=1}^{T} \prod_{i=1}^{m} p_i(\mathbf{w}_i^T \mathbf{x}(t)) \mid \det \mathbf{W} \mid \tag{2.58}$$

As shown in section 2.3.7, it is possible to use a logarithm of likelihood, as monotonicity of logarithm preserves maxima of the likelihood function. Moreover, the log-likelihood can be expressed in terms of sums rather than products:

$$\log L(\mathbf{W}) = \sum_{t=1}^{T} \sum_{i=1}^{m} \log p_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log \mid \det \mathbf{W} \mid \tag{2.59}$$

To simplify notation, considering the expected value over time, it holds:

$$\frac{1}{T} \log L(\mathbf{W}) = \mathrm{E}\{\sum_{i=1}^{m} \log p_i(\mathbf{w}_i^T \mathbf{x}(t))\} + T \log \mid \det \mathbf{W} \mid \tag{2.60}$$

One non trivial aspect of this approach is that the p_i indicated in (2.56)- (2.60) is the probability density function of the *unknown* sources. Therefore the pdf of the sources has to be estimated and this may be a hard problem, since it is, in general, a *nonparametric problem.* However there are some cases where the distribution of sources is known *a priori*, and in this case the log-likelihood would be a function of the only unmixing matrix $\mathbf{W}$. If this is not the case, it is possible to overcome this problem by approximating the densities of the independent components by means of a family of densities that are specified by a limited number of parameters. If the number of parameters is considerably high, there is no gain in such approach. However, it is possible to approximate almost *any* function by means of an extremely simple family of densities, as stated in the following theorem.

Theorem 7. *Denote by $\tilde{p}_i$ the* assumed *densities of the independent components and*

$$g_i(s_i) = \frac{\partial}{\partial s_i} \log \tilde{p_s}(s_i) = \frac{\tilde{p_i}'(s_i)}{\tilde{p_i}(s_i)} \tag{2.61}$$

Constrain the estimates of the independent components $y_i = \mathbf{b}_i^T \mathbf{x}$ to be uncorrelated and to have unit variance. Then the ML estimator is locally consistent, if the assumed densities $\tilde{p}_i$ fulfill:

$$\mathrm{E}\{s_i g_i(s_i) - g'(s_i)\} > 0 \tag{2.62}$$

for all i.

This theorem has been presented in different forms in [4], [39] and [80], and here we are referring to the formulation presented in [80], where there is the constraint of norm unit, that simplifies the theorem. The theorem shows also that small misspecifications in the densities p_i do not affect the consistency of the estimator, since sufficiently small changes do not change the sign in (2.62).
Moreover, the Theorem shows also how to construct families consisting of only two densities, so that the condition (2.62) is true for one of those densities. Consider the following two log-densities:

$$\log \tilde{p_i}^{+}(s) = \alpha_1 - 2 \log \cosh(s) \tag{2.63}$$

$$\log \tilde{p_i}^{-}(s) = \alpha_2 - [s^2/2 - \log \cosh(s)] \tag{2.64}$$

where α_1 and α_2 are positive parameters that are fixed so that the two functions are actually logarithms of probability densities.
It is easy to show that one of the two densities fulfills the condition of Theorem 7. In fact, it can be shown that, for density in (2.63), condition (2.62) becomes:

$$\mathrm{E}\{s_i g_i(s_i) - g'(s_i)\} = 2E\{-\tanh(s_i)s_i + (1 - \tanh(s_i)^2)\} \tag{2.65}$$

while for density in (2.64), condition (2.62) is:

$$\mathrm{E}\{s_i g_i(s_i) - g'(s_i)\} = 2E\{\tanh(s_i)s_i - (1 - \tanh(s_i)^2)\} \tag{2.66}$$

As the signs of the two expressions are always opposite, for *almost* any distribution, one of the two densities will fulfill Theorem 7. Of course, for some distributions the sign of (2.65) and (2.66) may be zero , but such cases are considered very rare (and correspond to the case of *perfect* Gaussianity, that is reasonably unfrequent).
Performing a ML estimation of the ICA model consists of maximizing the likelihood or log-likelihood function, as seen in this section. As guaranteed by Theorem 7, it is possible to approximate almost any kind of density of the sources by means of relatively simple density family. Moreover, it is possible to see the INFOMAX algorithm as a ML estimation (see [38],[129] and [141]). In the following sections, a ML formulation of the INFOMAX algorithm, presented in section 2.7.1, will be presented, together with the natural gradient method.

INFOMAX Algorithm

It is possible to derive a simple form of a ML estimation by means of a gradient method. Considering eq. (2.60) and differentiating with respect to $\mathbf{W}$, we obtain:

$$\frac{1}{T}\frac{\partial \log L}{\partial \mathbf{W}} = [\mathbf{W}^T]^{-1} + \mathrm{E}\{\mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T\} \tag{2.67}$$

where $\mathbf{g}(\mathbf{y}) = (g_1(y_1), \ldots, g_m(y_m))$ is a component-wise vector function that consists of the negative *score functions* g_i of the distribution of s_i, defined as:

$$g_i = (\log p_i)' = \frac{p_i{}'}{p_i} \tag{2.68}$$

It is possible to derive an optimization algorithm based on a gradient optimization, using (2.67):

$$\Delta\mathbf{W} \propto [\mathbf{W}^T]^{-1} + \mathrm{E}\{\mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T\} \tag{2.69}$$

and, in case of stochastic optimization, where the expected value is omitted, we have:

$$\Delta\mathbf{W} \propto [\mathbf{W}^T]^{-1} + \mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T \tag{2.70}$$

The eq. (2.70) is a generalization of eq. (2.51), where in this case g can be one chosen using the two functions in (2.63) and (2.64), according to super-Gaussianity

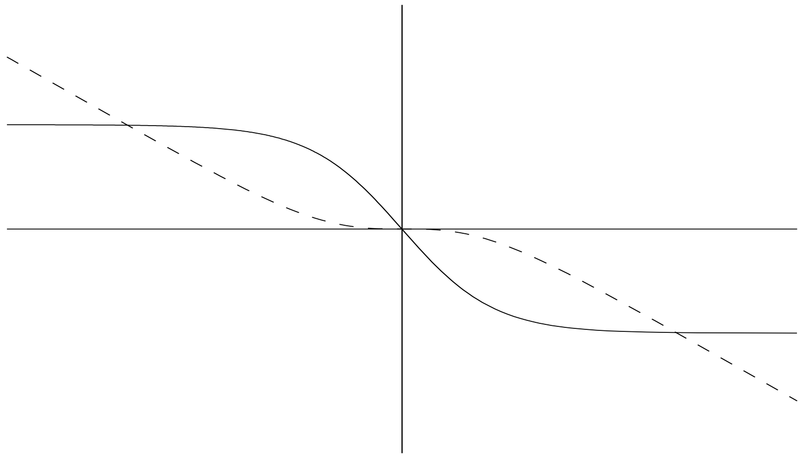


Figure 2.2: Score functions for Maximum Likelihood estimation: g^{+} in eq. (2.71) in solid line and g^{-} in eq.(2.72) in dashed line.

or sub-Gaussianity.
In case of super-Gaussian sources, using (2.63), we have

$$g^{+}(y) = -2\tanh(y) \tag{2.71}$$

while for sub-Gaussian sources, using (2.64), the score function is:

$$g^{+}(y) = \tanh(y) - y \tag{2.72}$$

The two nonlinearity are depicted in Fig. 2.2 The choice between the two nonlinearities in (2.71) and (2.72) can be made by computing the nonpolynomial moment:

$$\mathrm{E}\{-\tanh(s_i)s_i + (1 - \tanh(s_i)^2)\} \tag{2.73}$$

using some estimates of the components: if this polynomial moment is positive, the nonlinearity in (2.71) should be used, otherwise the nonlinearity in (2.72) should be used. This procedure has to be employed *while* performing iterations, and one may switch between the two nonlinearities during the extraction.
This algorithm, however, converges very slowly, especially due to the inversion of matrix $\mathbf{W}$, that is needed in every step. To improve the convergence, one may use whitened data (section 2.6), and especially use the natual gradient, that is explained in next section.

Natural Gradient

The Natural Gradient method simplifies the maximization of the likelihood considerably, and makes it better conditioned. It is based on the observation that, in an *Euclidean* orthogonal coordinate system, the steepest direction is the one of the *gradient*, while if the parameter space has a *Riemannian* metric structure, the steepest direction is given by the so called *natural gradient*. For further details, see [3]. In our case, the use of this principle amounts to multiplying the right hand side of (2.69) by $\mathbf{W}^T\mathbf{W}$, obtaining:

$$\Delta\mathbf{W} \propto \mathbf{W} + (\mathrm{E}\{\mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T\mathbf{W}^T\})\mathbf{W} = (\mathbf{I} + \mathrm{E}\{\mathbf{g}(\mathbf{y})\mathbf{y}^T\})\mathbf{W} \tag{2.74}$$

Interestingly, this algorithm can be interpreted as *nonlinear decorrelation*, that will be treated in section 2.7.3

Bayesian formulation of the INFOMAX algorithm

It is possible to look at the INFOMAX algorithm from a different point of view. In fact, in [93], Knuth proposed a Bayesian framework for blind source separation that leads to the INFOMAX update rule. The natural starting point of a Bayesian approach is, of course, the Bayes theorem, that in our case allows to express the probability of the model in terms of the likelihood of the data and the prior probability of the model and the data:

$$P(model\,|data, I\,) = \frac{P(data\,|model, I\,)P(model\,|I\,)}{P(data\,|I\,)} \tag{2.75}$$

where I represents the prior information on data or model. For a source separation problem, the model in eq. (2.75) becomes:

$$P(\mathbf{A}, \mathbf{s}\,|\mathbf{x}, I\,) = \frac{P(\mathbf{x}\,|\mathbf{A}, \mathbf{s}, I\,)P(\mathbf{A}, \mathbf{s}\,|I\,)}{P(\mathbf{x}\,|I\,)} \tag{2.76}$$

With the assumptions of linear mixing, of independence of the sources s_i, expressed in form of priors, the optimization process that leads to the solution is:

$$\Delta\mathbf{W} = \mathbf{W} + \left(\frac{p_i'(u_i)}{p_i(u_i)}\right)\mathbf{u}^T\mathbf{W} \tag{2.77}$$

where, as usual, p_i denotes an estimate of the pdf of the source s_i.

2.7.3 Nonlinear decorrelation

As shown in section 2.3.2, if two random variables are uncorrelated, this do not necessarily imply that they are independent. On the contrary, if two random variables are independent, they will surely be uncorrelated. In the following of the section we will consider zero-mean random variables, whose correlation and covariance are equal. Extending the concept of correlation, it is possible to define the *nonlinear correlation* of two random variables y_1 and y_2 in the following way:

$$E_{fg} = \mathrm{E}\{f(y_1)g(y_2)\} \tag{2.78}$$

where f and g are two functions of which at least one is nonlinear. The two random variables y_1 and y_2 are said to be nonlinearly uncorrelated, if their nonlinear correlation is zero, for a given couple of functions f and g:

$$\mathrm{E}\{f(y_1)g(y_2)\} = 0 \tag{2.79}$$

Considering two random variables, it is possible to infer something about their independence starting from their *nonlinear uncorrelatedness*? There is a general theorem (see [80]) that states that two random variables y_1 and y_2 are statistically independent *if and only if*

$$\mathrm{E}\{f(y_1)g(y_2)\} = \mathrm{E}\{f(y_1)\}\mathrm{E}\{g(y_2)\} \tag{2.80}$$

for *all* continuous functions f and g that are zero ouside a finite interval. To use the theorem in order to look for independence, it is necessary to make some approximations, as f and g are completely arbitrary. In particular, in [88] it has been proposed an approach to achieve independence by means of nonlinear decorrelation. To overcome some of the drawbacks of this algorithm, in [43] a modified version has been proposed. In the following sections, the two algorithms will be illustrated. It has to be noted that the Jutten-Hérault algorithm is highly inefficient, if compared with other algorithms, but is one of the first works in the ICA topic (together with the work of Comon [45]), and thus it is worth explaining with some details its properties

The Jutten-Hérault algorithm

Consider the two functions f and g considered in (2.78): assume that both have derivatives of all order in the neighborhood of the origin. They can be expanded in Taylor series:

$$f(y_1) \;=\; f(0) + f'(0)y_1 + \frac{1}{2}f''(0)y_1^2 + \ldots = \sum_{i=0}^{\infty} f_i y_1^i \tag{2.81}$$

$$f(y_2) \;=\; g(0) + g'(0)y_2 + \frac{1}{2}g''(0)y_2^2 + \ldots = \sum_{i=0}^{\infty} g_i y_2^i \tag{2.82}$$

where f_i, g_i is shorthand for the coefficients of the ith powers in the series. The product of the functions is then

$$f(y_1)g(y_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} f_i g_j y_1^i y_2^j \tag{2.83}$$

Condition (2.79) now becomes:

$$\mathrm{E}\{f(y_1)g(y_2)\} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} f_i g_j \mathrm{E}\{y_1^i y_2^j\} \tag{2.84}$$

A sufficient condition for (2.84) can be:

$$\mathrm{E}\{y_1^i y_2^j\} = 0 \tag{2.85}$$

for all indices i, j appearing in the series expansion.

There may be, however, some situations where higher order correlations are not zero, but the coefficients f_i and g_i happen to be suitable to cancel the terms and make the sum in (2.84) exactly zero. However, for nonpolynomial functions that have infinite Taylor expansions, such spurious solutions can be be considered unlikely.

A sufficient condition for (2.85) is that y_1 and y_2 are *independent* and one of $\mathrm{E}\{y_1^i\}$, $\mathrm{E}\{y_2^j\}$ is zero ([80]). Let us require that $\mathrm{E}\{y_1^i\} = 0$ for all powers i appearing in its series expansion: this is only possible when f is an odd function (in fact, unless y_1 is a constant, the moment of second order, i.e. its variance, cannot be zero, therefore in order to satisfy the condition f must have a Taylor expansion such that even coefficients f_{2k} are zero). To conclude, a sufficient, but not necessary condition for the nonlinear uncorrelatedness condition (2.79) is that y_1 and y_2 are independent, and for one of them the nonlinearity is an odd function.

Guided by this principles, Jutten , Herault proposed in [88] a neural network architecture with feedback presented in the bidimensional case in Fig. 2.3.

The model of the ICA problem, in the bidimensional case, is the following:

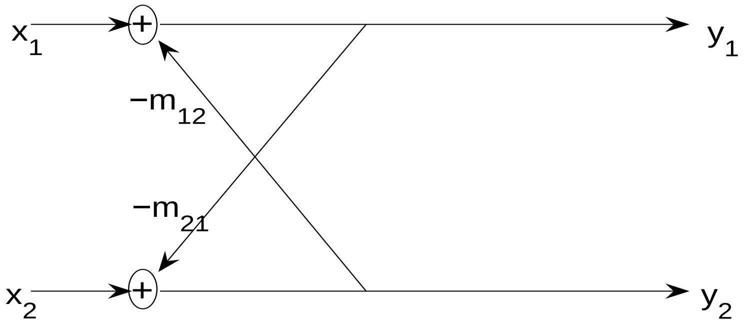


Figure 2.3: Jutten-Hérault feedback architecture for source separation

$$x_1 = a_{11}s_1 + a_{12}s_2 \tag{2.86}$$

$$x_2 = a_{21}s_1 + a_{22}s_2 \tag{2.87}$$

From Fig 2.3 we have:

$$y_1 = x_1 - m_{12}y_2 \tag{2.88}$$

$$y_2 = x_2 - m_{21}y_1 \tag{2.89}$$

Defining a matrix $\mathbf{M}$ with off-diagonal elements m_{12} and m_{21} and diagonal elements equal to zero, equations (2.88) and (2.89) can be expressed as:

$$\mathbf{y} = \mathbf{x} - \mathbf{M}\mathbf{y} \tag{2.90}$$

therefore the input-output mapping of the network can be expressed as:

$$\mathbf{y} = (\mathbf{I} + \mathbf{M})^{-1}\mathbf{x} \tag{2.91}$$

It has to be noted that if $\mathbf{I}+\mathbf{M} = \mathbf{A}$, then the output of the neural network becomes equal to the independent components. To obtain independence of the output, the authors have proposed the nonlinear correlations criterion, and used the following learning rules:

$$\Delta m_{12} \propto \mu f(y_1)\, g(y_2) \tag{2.92}$$

$$\Delta m_{21} \propto \mu f(y_2)\, g(y_1) \tag{2.93}$$

where μ is the learning rate and $f(y) = y^3$ and $g(y) = \arctan y$, according to the previous considerations in this section. The algorithm has several drawbacks, and does not work efficiently if the sources are bad scaled or if the mixing matrix is ill-conditioned. Moreover, the inversion of matrix $(\mathbf{I}+\mathbf{M})$ at each iteration is rahter computationally demanding. To overcome these problems, in [43] has been proposed an extension, and it will be presented in the next section

The Cichocki-Unbehauen algorithm

To overcome the problems of the feedback architecture proposed by Jutten and Hérault presented in Fig. 2.3, Cichocki and Unbehauen proposed a *feedforward* architecture with weight matrix $\mathbf{B}$, with the mixture input vector $\mathbf{x}$ and the output $\mathbf{y} = \mathbf{Bx}$. The aim of the optimization is to adapt the matrix $\mathbf{B}$ so that the output elements are independent. The learning algorithm proposed is:

$$\Delta\mathbf{B} \propto \mu[\mathbf{\Lambda} - \mathbf{f}(\mathbf{y})\mathbf{g}(\mathbf{y}^T)]\mathbf{B} \tag{2.94}$$

where μ is, as usual, the learning rate, $\mathbf{\Lambda}$ is a diagonal matrix whose elements determine the amplitude scaling for the output, and f and g are two *nonlinear* functions, where $\mathbf{f}(\mathbf{x})$ indicates a column vector whose elements are $f(x_1), f(x_2), \ldots, f(x_n)$. The authors proposed a polynomial and the hyperbolic tangent as a choice for the nonlinearities.

Once we have convergence for this algorithm, the output will be nonlinearly uncorrelated, as it will hold:

$$\mathbf{\Lambda} - \mathrm{E}\{\mathbf{f}(\mathbf{y})\mathbf{g}(\mathbf{y}^T)\} \tag{2.95}$$

and, as $\mathbf{\Lambda}$ is a *diagonal* matrix, it will hold:

$$\mathrm{E}\{f(y_i)g(y_j)\} = 0, \quad \forall\, i \neq j \tag{2.96}$$

As shown in section 2.7.2, this algorithm has a close relationship with the Maximum Likelihood method. In fact, choosing nonlinearities according to what stated in 2.7.2, and using the *natural gradient* to optimize the log-likelihood, one obtains the *same* algorithm. In general, many of the ICA methods presented throughout this chapter have some kind of theoretical connection, and it will be presented more systematically in section 2.7.7.

2.7.4 Nonlinear PCA criterion

As seen in section 2.3.4, PCA is a powerful tool to have a representation of a dataset in terms of minimum square compression error. PCA is based on second order statistics, while ICA uses all the statistics. Therefore it is straightforward the extension of the PCA criterion to a nonlinear version, such that higher order statistics can be accounted for. It is useful to remind the *linear* PCA criterion for minimum square compression error presented in eq. (2.16):

$$J_{MSE}^{PCA} = \mathrm{E}\{\|\mathbf{x} - \sum_{i=1}^{n}(\mathbf{w}_i^T\mathbf{x})\mathbf{w}_i\|^2\} \tag{2.97}$$

The aim of the nonlinear criterion is to use a *suitable* nonlinear function such that higher order statistics are considered while performing minimization. The nonlinear PCA criterion ([133, 90]) can be formulated as the problem of minimizing a new cost function related to data as in (2.16) but with an additional nonlinearity:

$$J_{NLPCA}(\mathbf{W}) = \mathrm{E}\{\|\mathbf{x} - \mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{x})\|^2\} \tag{2.98}$$

where $\mathbf{g}(\mathbf{x})$ is the vector obtained by applying a nonlinear function $g(x)$ to the vector $\mathbf{x}$. In [133], it was proposed to use as nonlinearities some odd functions as $g(t) = \tanh(t)$ or $g(t) = t^3$, and it was proposed a minimization of the criterion by means of a stochastic gradient descent, giving the following update rule:

$$\boldsymbol{\Delta}\mathbf{W} \propto [\mathbf{x} - \mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{x})]\mathbf{g}(\mathbf{x}^T\mathbf{W}). \tag{2.99}$$

For this particular criterion, it is not necessary to perform whitening of data, as the optimum solution $\mathbf{W}_{opt}$ is an orthogonal matrix, regardless of the whitening constraint.

As the optimization criterion proposed in (2.99) is demanding in terms of memory and time, in [89] it has been proposed a Least Square Method (LSM) to optimize the contrast function. In section 2.7.7, it will be shown that the nonlinear PCA criterion is intimately connected with Maximum Likelihood estimation.

2.7.5 Use of Time-structure

The algorithms presented in this section are considerably different from the others, and some of the considerations that lead to use this kind of algorithms will be done also in Chapter 6. The basic idea behind these algorithm is that, while dealing with structured signals, classical ICA techniques tend to neglect the structure of data, therefore it could be useful to use the information about such structure during extraction. These algorithms, actually, can perform extraction also in the cases where classical ICA fails, but have some more limitation not present in other formulations. Consider a random vector $\mathbf{x} \in \Re^m$ and consider T realizations of such vector. Data can be stored in a matrix $\mathbf{X} \in \Re^{m \times T}$, where each row represents an observed signal, and each column is a realization of the random vector at a given time point. Suppose that data have a intrinsic structure (like, for instance, audio signals), and suppose to shuffle all the columns such that this information is lost. Classic ICA techniques will not be affected by this procedure, as the statistics will be preserved, however the structure of data is lost in the shuffling process, and so some information available will be unused. To overcome this drawback, some techniques have been proposed to extract independent components with time structure.

The simplest form of time structure is given by linear autocovariances, that are the covariances between the values of the signal at different time points: $\mathrm{cov}(x_i(t)x_i(t-\tau))$, where τ is some *lag* constant. In addition to autocovariances, it is possible to consider also covariances between two different signals: $\mathrm{cov}(x_i(t)x_j(t-\tau))$ with $i \neq j$. To consider all these statistics in compact form, the time-lagged covariance matrix can be considered:

$$\mathbf{C}_\tau^\mathbf{x} = \mathrm{E}\{\mathbf{x}(t)\mathbf{x}(t-\tau)^T\} \tag{2.100}$$

As seen in sections 2.3.4 and 2.4, performing a whitening of data is not enough to achieve independence. There is an infinite number of linear transformations that give decorrelated components, therefore higher order statistics are used to add some information to the separation to achieve also independence. The second order statistics approaches, instead, do not take this additional information from higher order statistics, but from the lagged covariance matrix $\mathbf{C}_\tau^\mathbf{x}$, starting from the consideration that if x_1 and x_2 are independent, not only their cross-covariance will be zero, but also the lagged cross-covariances for any lag. This means that we are looking for a linear transformation $\mathbf{W}$ of data $\mathbf{x}$ such that the vector $\mathbf{y}$ has these two properties:

$$\mathrm{E}\{y_i(t)y_j(t)\} = 0 \qquad \forall i \neq j \tag{2.101}$$

$$\mathrm{E}\{y_i(t)y_j(t-\tau)\} = 0 \qquad \forall i \neq j, \forall \tau. \tag{2.102}$$

While (2.101) alone would lead to uncorrelatedness, the conbination of (2.101) and (2.102) leads to independence, without using high order statistics. Two approaches have been proposed, using these considerations, and are based on the use of a single lag or a set of several lags.

Techniques with one lag

In the simplest case, one may consider only one lag in the lagged covariance matrix, and find a linear transformation of data $\mathbf{x}$ such that (2.101) and (2.102) hold. Suppose data have been whitened, so we have, instead of $\mathbf{x}$, a data vector $\mathbf{z}$ whose components are uncorrelated (at lag zero), as seen in (2.42). The target separating matrix $\mathbf{W}$ is such that:

$$\mathbf{Wz}(t) = \mathbf{s}(t) \tag{2.103}$$

$$\mathbf{Wz}(t-\tau) = \mathbf{s}(t-\tau) \tag{2.104}$$

A modified version of the lagged covariance matrix seen in eq. (2.100) is used in these algorithms, and precisely:

$$\overline{\mathbf{C}}^{\mathbf{z}}_{\tau} = \frac{1}{2}[\mathbf{C}^{\mathbf{z}}_{\tau} + (\mathbf{C}^{\mathbf{z}}_{\tau})^T] \tag{2.105}$$

and, by linearity and orthogonality and eqq. (2.103)–(2.104) we have:

$$\overline{\mathbf{C}}^{\mathbf{z}}_{\tau} = \frac{1}{2}\mathbf{W}^T \left[\mathrm{E}\{\mathbf{s}(t)\mathbf{s}(t-\tau)^T\} + \mathrm{E}\{\mathbf{s}(t-\tau)\mathbf{s}(t)^T\}\right] \mathbf{W} = \mathbf{W}^T\overline{\mathbf{C}}^{\mathbf{s}}_{\tau}\mathbf{W} \tag{2.106}$$

As $\mathbf{s}$ is the vector of the independent components, its lagged covariance matrix $\overline{\mathbf{C}}^{\mathbf{s}}_{\tau}$ will be diagonal, therefore we have that the lagged covariance matrix $\overline{\mathbf{C}}^{\mathbf{z}}_{\tau}$ of whitened data $\mathbf{z}$ can be decomposed by means of its eigenvalues and eigenvectors $\overline{\mathbf{C}}^{\mathbf{z}}_{\tau} = \mathbf{W}^T\mathbf{D}\mathbf{W}$ and its eigenvalues are the lagged autocovariances of the sources. Some algorithms based on these criterion have been proposed in [124] and [165] (AMUSE).

Although this approach is very simple and fast to compute, it only works if the eigenvalues of the lagged covariance matrix are uniquely defined. If some of the eigenvalues are identical, the corresponding eigenvectors cannot be uniquely defined, therefore the corresponding IC cannot be estimated. To overcome this problem, it is possible to search for a suitable lag such that all the eigenvalues are distinct, but it could be not possible in the case signals have the same power spectra (and, subsequently, the same autocovariance).

Techniques with several lags

To overcome some of the problems presented in previous section, algorithms that consider multiple lags have been developed, performing *simultaneous diagonalization* of lagged covariance matrices. As it is highly unlikely that the eigenvectors of different lagged covariance matrices are the same, a measure of how good the diagonalization is must be introduced. One suitable choice is:

$$\mathrm{off}(\mathbf{M}) = \sum_{i \neq j} m_{ij}^2 \tag{2.107}$$

that is the sum of the squares of the off-diagonal elements of matrix $\mathbf{M}$.
Multiple lags algorithms are based on a set S of chosen lags, and are based on the minimization of a contrast function related to eq. (2.107)

$$\mathcal{J}_1(\mathbf{W}) = \sum_{\tau \in S} \mathrm{off}(\mathbf{W}\overline{\mathbf{C}}^{\mathbf{z}}_{\tau}\mathbf{W}^T) \tag{2.108}$$

Minimization can be performed by means of gradient descent, or by adapting the existing methods for eigenvalue decomposition to this simultaneous approximate diagonalization of several matrices. The algorithm called SOBI (Second-order blind identification), that is based in these considerations, has been proposed in [26], and also the algorithm called TDSEP (Temporal Decorrelation source SEParation) has been proposed in [180]. Different improvements to these techniques have been proposed, in particular in [175] an optimal way of weighting the different lags have been proposed.
If compared with classical ICA techniques, second order methods have the advantage of dealing also with Gaussian sources. However, if the sources have all the same power spectra (and thus autocovariance), second order methods will fail in recovering the sources, while higher order techniques will not suffer from this limitation.

2.7.6 Maximization of Non-Gaussianity

As seen in sections 2.4 and 2.5, Gaussianity is "forbidden" while looking for independent components. As seen before, if more than one component has a Gaussian probability density, the model will not be identifiable and far from separable. This information, combined with the Central Limit Theorem (CLT), can provide a useful criterion to achieve independence in a linear mixture.

Consider n *independent* random variables $\{x_1, \dots, x_n\}$ with mean and variance $\{\mu_1, \dots, \mu_n\}$ and $\{\sigma_1^2, \dots, \sigma_n^2\}$ respectively. Consider now the random variable obtained as a sum of x_i:

$$x = \sum_{i=1}^{n} x_i \tag{2.109}$$

Due to the independence of the x_i, the mean of x will be $\mu = \sum_{i=1}^{n} \mu_i$ and the variance will be $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$. The Central Limit Theorem states that, under certain general conditions ([151, 137]), the probability density function of x tends to the normal density as the number of components n increases. The theorem can be stated as a limit considering the new variable $z = (x - \mu)/\sigma$, giving:

$$\lim_{n \to \infty} f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \tag{2.110}$$

For a proof, see [137]. The Theorem guarantees that, if the number of signals tends to infinity, the probability density will be described as a Gaussian. However, if the variables have the same distribution, around 30 signals will turn to be enough such that the distribution of the sum is Gaussian. Moreover, in the case of smooth densities, a value of n as low as 5 can be used ([137]).

Starting from the CLT, it is possible to consider a way of maximizing independence by means of non-Gaussianity. Loosely speaking, consider a set of independent signals $\mathbf{s}$, and consider the observation $\mathbf{x}$ after the linear combination by means of mixing coefficients $\mathbf{A}$. To recover the sources one has to look for the combination of the original data $\mathbf{y} = \mathbf{W}\mathbf{x}$ such that the components of $\mathbf{y}$ are maximally statistically independent. It is possible to see the problem of finding the best transformations of data in terms of Gaussianity. Starting from what stated in 2.4 and 2.5, it is necessary, for the ICA model to hold, that there are no Gaussian sources (at most there can be only one). Therefore we can assume that the original sources have a non-Gaussian distribution, and, by means of the Central Limit Theorem, state that a linear combination of the sources will have a pdf that is closer to a Gaussian or at most equal to one of the sources. Consider a particular combination y of the observations $\mathbf{x}$ by means of vector $\mathbf{w}^T$: $y = \mathbf{w}^T\mathbf{x}$ and consider its non-Gaussianity (measured by some suitable criterion, that will be explained later). Knowing that there is a generative model behing the observations $\mathbf{x} = \mathbf{A}\mathbf{s}$, it is possible to express *any* linear combination of the observed (known) signals in terms of sources: $\mathbf{y} = \mathbf{w}^T\mathbf{A}\mathbf{s} = \mathbf{b}^T\mathbf{s}$ where $\mathbf{b} = \mathbf{A}^T\mathbf{w}$. The combination $\mathbf{b}$ that guarantees the maximum non-Gaussianity is the one that considers a single source: in fact, a combination of two or more non-Gaussian sources will have a pdf more Gaussian than the original terms in the sum.

All that is needed to point out non Gaussianity is a function of random variable that has a value f for a Gaussian pdf and a value f' for *any* other distribution, such that it always holds $f \geq f'$ (or $f \leq f'$), where the equality holds only when f' has a Gaussian density.

One candidate as a discriminating function over Gaussianity is kurtosis, that is defined as the fourth order cumulant, and has the property of being always null for Gaussian distribution and nonzero for (almost) all non-Gaussian distributions. Another more information theory based way of looking for non-Gaussianity is to maximize negentropy. Those two approaches will be discussed in the two sections.

Non-Gaussianity using cumulants

To define the concept of cumulant, recall the first *characteristic function* (c.f.) defined in section 2.5, eq. (2.38). By considering the terms of the Taylor expansion around 0 of the c.f. it is possible to estimate all the moments of a random variable (in fact, the characteristic function is also called *moment generating function*).

The *second characteristic function* $\phi(\omega)$ of a random variable x, called also *cumulant generating function*, is given by the natural logarithm of the first c.f.:

$$\phi(\omega) = \ln(\psi(\omega)) = \ln(\mathrm{E}\{(e^{j\omega x})\}) \tag{2.111}$$

and, expanding $\phi(\omega)$ by means of Taylor series, we have:

$$\phi(\omega) = \sum_{k=0}^{n} \kappa_k \frac{(j\omega)^k}{k!} \tag{2.112}$$

where the k-th cumulant is obtained as the derivative

$$\kappa_k = (-j)^k \frac{d^k\phi(\omega)}{d\omega^k}\bigg|_{w=0} \tag{2.113}$$

In general, the cumulants of all orders can be expressed as a combination of moments, but usually it is preferable to work directly with cumulants , because they present in a clearer way the additional information provider by higher order statistics. Moreover, cumulants show additional properties not shared by moments ([80]):

- Given two *independent* random vectors $\mathbf{x}$ and $\mathbf{y}$ having the same dimension, then the cumulant of their sum $\mathbf{z} = \mathbf{x}+\mathbf{y}$ is equal to the sum of the cumulants of $\mathbf{x}$ and $\mathbf{y}$. The property also holds for the sum of more than two independent random vectors.

- If the distribution of the random vector $\mathbf{x}$ is multivariate Gaussian, then all its cumulants of order three and higher are zero.

The fourth order cumulant, as defined before, is also called kurtosis, and for a *zero mean* random variable it holds:

$$\mathrm{kurt}(x) = \kappa_4 = \mathrm{E}\{x^4\} - 3[\mathrm{E}\{x^2\}]^2 \tag{2.114}$$

According to the sign of kurtosis, it is possible to classify the distribution of a random vector x. Consider $\kappa = \mathrm{kurt}(x)$, then:

- If $\kappa = 0$: x has a Gaussian distribution

- If $\kappa > 0$: x has a supergaussian distribution (*Leptokurtic*)

- If $\kappa < 0$: x has a subgaussian distribution (*Platikurtic*)

Supergaussian distributions, depicted in fig. 2.4.(b), have tipically a “spiky” pdf with heavy tails, having larger values at zero and at at the tails, while being smaller in the central values, if compared with a Normal density. On the other hand, subgaussian pdfs (2.4.(c)) are rather constant near zero and very small for larger values of the variable. A typical example of supergaussian variable is given by Laplace density, while a uniform density is subgaussian. There are also some nongaussian random variables that have zero kurtosis, but they can be considered very rare ([80]).

The computation of kurtosis, in the case of zero mean and unit standard deviation variables (that happens often while dealing with ICA after preprocessing), reduces to the computation of the fourth order moment, and also all the analysis are simplified

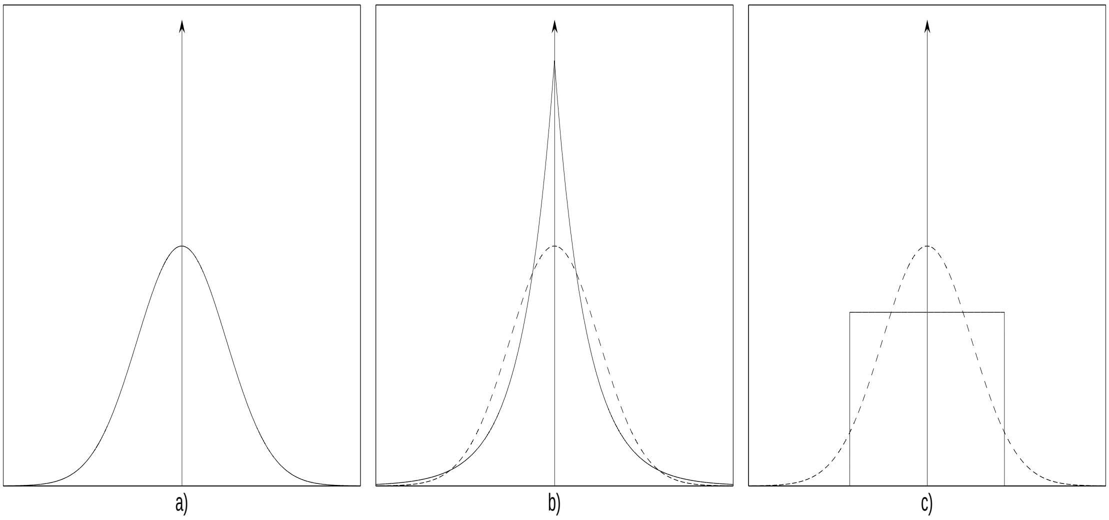


Figure 2.4: Probability distributions at different kurtosis. **a**) Gaussian distribution ($\kappa = 0$), **b**) Supergaussian distribution ($\kappa > 0$), compared with a Gaussian (dashed line) **c**) Subgaussian distribution ($\kappa < 0$), compared with a Gaussian (dashed line).

by the fact that the sum of independent signals has a kurtosis that is the sum of the kurtosis of the single variables. In general, for any linear combination $y = b_1 s_1 + b_2 s_2$ of two independent random variables, it holds:

$$\mathrm{kurt}(y) = \mathrm{kurt}(b_1 s_1 + b_2 s_2) = b_1^4 \mathrm{kurt}(s_1) + b_2^4 \mathrm{kurt}(s_2) \tag{2.115}$$

Therefore, the linear transformation that points out a *local* maximum in the absolute value of kurtosis (i.e. considering both supergaussian and subgaussian sources), leads to an independent component.

However, it is clear that a measure based on the fourth order moment can be rather sensitive to outliers. In fact, considering a pdf centered around zero, if there is a measurement error that leads to a value extremely outside the range of values of the variable, the kurtosis, that is basically based on the fourth power of data, will be heavily affected by this error. Therefore kurtosis is not a robust measure for independence, and this could be a problem in real world problems, where usually noise heavily affects measurements. Therefore a different approach has been proposed in literature to overcome this problem, and it is the one based on Negentropy, presented in the following section.

Non-Gaussianity using Negentropy

Negentropy has been defined in section 2.3.5, eq. (2.25), and it has been shown how it could be used as a measure of "distance" between a random variable density and a Gaussian density with the same mean and variance. Moreover, negentropy is invariant for linear invertible transformation, unlike classical entropy.
This properties make negentropy a suitable function for pointing out independence, with a rigorous theoretical background and with more robustness with respect to noise and outliers, if compared with cumulants. Of course, to properly estimate the negentropy of a random variable, one needs to know exactly the probability density of the variable, that in the majority of the cases is not feasible. Thus some approximations of negentropy $J(y)$ of a random variable y with zero mean and unit variace have been provided.
The classical method of approximating negentropy is by means of cumulants, giving:

$$J(y) \approx \frac{1}{12}(\mathrm{E}\{y^3\})^2 + \frac{1}{48}(\mathrm{kurt}(y))^2 \tag{2.116}$$

Of course, the approximation in eq. (2.116) does not give any help in terms of robustness to outliers, as it is defined by means of moments and cumulants. In [80] a different approximation of negentropy can be found, and it is based on expectations on nonquadratic functions:

$$J(y) \approx k_1(\mathrm{E}\{G_a(y)\})^2 + k_2(\mathrm{E}\{G_b(y)\} - \mathrm{E}\{G_b(\nu)\})^2 \tag{2.117}$$

where G_a and G_b are two nonquadratic functions, with G_a odd and G_b even, ν is a Gaussian variable with zero mean and unit variance (as the variable y), and k_1 and k_2 are positive constants.
The previous approximation can be done also by means of a *single* function G in the following way:

$$J(y) \approx k[\mathrm{E}\{G(y)\} - \mathrm{E}\{G(\nu)\}]^2 \tag{2.118}$$

The quality of the approximation heavily rely on the choose of G function, and choosing a G that does not grow too fast helps making a more robust estimation. Two good choices for G are the following:

$$\begin{aligned} G_1(y) &= \frac{1}{a}\log\cosh ay && (2.119) \\ G_2(y) &= -\exp(-y^2/2) && (2.120) \end{aligned}$$

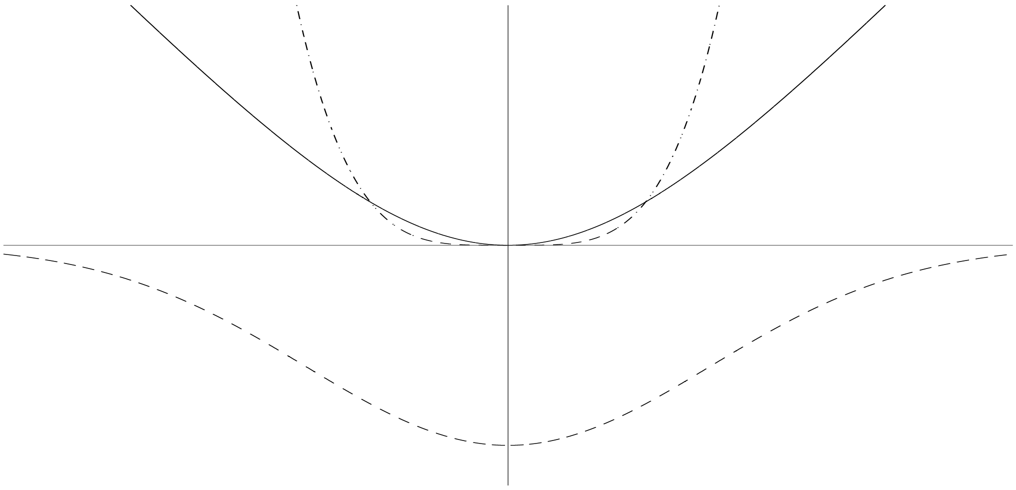


Figure 2.5: Nonlinearities for negentropy approximation: $G_1(y)$ in solid line, G_2 in dashed line and $G_3(y)$ (kurtosis) in dash-dotted line

where $1 \leq a_1 \leq 2$ is some suitable constant, often taken as 1. Also kurtosis can be expressed in the same form, choosing $G_3(y) = y^4$. These three approximating functions are depicted in Fig. 2.5.

2.7.7 Connection between independence estimation principles

It is possible to find some connections between the different solutions to the ICA problem presented from section 2.7.1 to 2.7.6. In fact many of the algorithms presented are related to the mutual information maximization approach.

Mutual information (see section 2.3.6) is related to the degree of independence of a set of random variables. For a linear transformation $\mathbf{y} = \mathbf{B}\mathbf{x}$, equation (2.26) becomes:

$$I(y_1, y_2, \ldots, y_n) = \sum_i H(y_i) - H(\mathbf{x}) - \log |\det \mathbf{B}| \tag{2.121}$$

If we constrain the y_i to be uncorrelated and of unit variance, the the last term on the right hand side is constant. In fact it does not depend on $\mathbf{B}$: consider the covariance matrix $C_{\mathbf{y}} = \mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{B}\mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{B}^T$, and for independence and unit variance it holds: $\mathbf{C}_{\mathbf{y}} = \mathbf{I}$. It is possible to show therefore that $\det \mathbf{B}$ must be constant:

$$\det \mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \det\big(\mathbf{B}\mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{B}^T\big) = (\det \mathbf{B})\left(\det \mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\right)(\det \mathbf{B}^T) = 1 \tag{2.122}$$

and, since det E$\{\mathbf{xx}^T\}$ does not depend on the transformation $\mathbf{B}$, det$\mathbf{B}$ must be constant. Moreover, if y_i has unit variance, its entropy and its negentropy will differ only by a constant and the sign (see eq. (2.25)). Therefore, we obtain that:

$$I(y_1, y_2, \ldots, y_n) = \text{const.} - \sum_i J(j_i) \tag{2.123}$$

that shows the relation between mutual information and negentropy: for a linear transformation $\mathbf{B}$, the minimization of mutual information leads to the maximization of negentropy, and therefore to the maximization of non-Gaussianity of the independent components. There is also a connection between mutual information methods and ML based ones. In fact, consider eq. (2.60), that is reported for clarity (inverting the position of the expected value and of the sum, always possible due to the linearity of the two operators):

$$\frac{1}{T} \log L(\mathbf{B}) = \sum_{i=1}^{m} \mathrm{E}\{\log p_i(\mathbf{b}_i^T \mathbf{x}(t))\} + T \log \mid \det\mathbf{B} \mid \tag{2.124}$$

If the estimates of the probability density of the sources p_i was equal to the actual densities, then the first term in the right hand side of the equation would be equal to the entropy of those random variables (up to an additive constant and sign). On the other side, it is possible to express the mutual information by means of an approximation of the densities of the sources y_i, and use this approximated densities to evaluate the entropy. In this case, mutual information can be expressed as:

$$I(y_1, y_2, \ldots, y_n) = -\sum_i \mathrm{E}\left\{G_i(y_i)\right\} - \log \mid \det \mathbf{B} \mid - H(\mathbf{x}) \tag{2.125}$$

leading to almost the same algorithm as in the ML case.

It has been show in section 2.7.3 that the equations of the algorithms based on the nonlinear decorrelation principle are of the same form of the ones of the ML case. Thus, ML estimation gives a principle for choosing those nonlinearities.

It is possible to say, therefore, that almost all of the criteria here presented (excluding second order methods) can be seen as different aspect of the same information theoretic principles of mutual information or of maximum likelihood. Moreover also non-Gaussianity search, derived from different principles (Central Limit Theorem) is intimately connected with mutual information approach, and its validity is further enforced by the rigorousness of the mutual information approach.

2.8 FastICA algorithm

The algorithm known as FastICA [79] is one of the most popular ICA algorithms, and both its speed and accuracy make it one of the most used in applications of several kinds. The principle of FastICA is the maximization of non Gaussianity, that is estimated by means of a negentropy approximation. As seen in section 2.7.6, the maximization of non-Gaussianity leads to independence, as long as the original sources do not have a Gaussian distribution (but it is not possible that more than one component has a Gaussian pdf). It has been shown also that it is possible to approximate negentropy in a robust way by means of nonlinear functions G.

Consider observed data vector $\mathbf{x} \in \Re^m$, and suppose that for $\mathbf{x}$ it holds the ICA model (2.33): $\mathbf{x} = \mathbf{As}$, where $\mathbf{s} \in \Re^m$ is the independent components vector, and $\mathbf{A} \in \Re^{m\times m}$ is the mixing matrix. After performing the standard preprocessing (mean removal and whitening, section 2.6), data will be represented by vector $\mathbf{z} \in \Re^m$, with zero mean and with covariance matrix equal to identity matrix: $\mathbf{C_z} = \mathbf{I}$, where $\mathbf{z} = \mathbf{Vs}$, as in eq. (2.42). In section 2.4 it has been shown that, due to double indeterminacy both in sources and in the mixing matrix, sources are constrained to be unit-norm, that is

$$\|\mathbf{s}\|^2 = 1. \tag{2.126}$$

The aim of the algorithm is to find a linear transformation of data $\mathbf{W}$ such that $\mathbf{s} = \mathbf{Wz}$. Considering eq. (2.126) and the fact that whitened data $\mathbf{z}$ are such that $\|\mathbf{z}\|^2 = 1$ by definition, it holds for a single component $s = \mathbf{w}^T\mathbf{z}$:

$$\|s\|^2 = 1 = \|\mathbf{w}^T\mathbf{z}\|^2 = \|\mathbf{w}\|^2\|\mathbf{z}\| = \|\mathbf{w}\|^2 \tag{2.127}$$

therefore, equation (2.126) leads, for whitened data, to the constraint $\|\mathbf{w}\|^2 = 1$.

Now it is possible to formulate the FastICA optimization in terms of whitened data $\mathbf{z}$ and unmixing coefficients $\mathbf{w}$ for a *single* unit. The problem of maximizing an estimate of negentropy $J(\mathbf{w}^T\mathbf{z})$ becomes a problem of constrained optimization; in fact, the problem can be formulated as

$$\begin{array}{rl} \text{maximize} & J_G(\mathbf{w}^T\mathbf{z}) \\ \text{under constraint} & \|\mathbf{w}\|^2 = 1 \end{array} \tag{2.128}$$

where $J_G(\mathbf{w}^T\mathbf{z}) = k\left[\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\} - \mathrm{E}\{(\nu)\}\right]$ as seen in eq. (2.118).

The problem has $2n$ local constrained maxima. In fact if $\mathbf{w}$ is a solution, also $-\mathbf{w}$

will be a solution. In case of n component extraction, the additional constraint of decorrelation for vectors $\mathbf{w}_i$ must be added. Therefore, the complete problem can be formalized as follows:

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{n} J_G(\mathbf{w}_i^T \mathbf{z}) \\ \text{under constraint} & \mathrm{E}\{(\mathbf{w}_j^T \mathbf{z})(\mathbf{w}_k^T \mathbf{z})\} = \delta_{j\,k} \end{array} \tag{2.129}$$

The convergence of the algorithm to the independent components may not be ensured, since a "rough" approximation of negentropy is being used. However, it is possible to prove that the approximation of negentropy of eq. (2.118) is good enough for a wide class of functions G to guarantee convergence, as stated in the following Theorem.

Theorem 8. *Assume that the input data follows the ICA model with whitened data:* $\mathbf{z} = \mathbf{VAs}$*, where* $\mathbf{V}$ *is the whitening matrix, and assume that* G *is a sufficiently smooth even function. Then the local maxima (resp. minima) of* $\mathrm{E}\{G(\mathbf{w}^T\mathbf{x})\}$ *under the constraint* $\|\mathbf{w}\| = 1$ *include those rows of the mixing matrix* $\mathbf{VA}$ *such that the corresponding independent component* s_i *satisfy*

$$\mathrm{E}\{s_i g(s_i) - g'(s_i)\} > 0 \quad (\text{resp.} < 0) \tag{2.130}$$

where g *is the derivative of* G *and* g' *is the derivative of* g*.*

For a proof, see [80, 79]. The Theorem shows that practically any nonquadratic function G may be used, as G divides the space of probability distributions into two half-spaces: independent components whose distribution is in one of the half spaces can be estimated by maximizing $\mathrm{E}\{G(\mathbf{w}^T\mathbf{x})\}$, while those whose distribution is in the other are estimated by minimizing the same function. The nonlinearities g usually employed are the derivatives of the ones presented in eqq. (2.119) and (2.120):

$$g_1(y) = \tanh(y) \tag{2.131}$$

$$g_2(y) = -y\exp(-y^2/2) \tag{2.132}$$

$$g_3(y) = y^3 \tag{2.133}$$

Those nonlinearities are depicted in Fig. 2.6.
Moreover, the previous theorem implies the following:

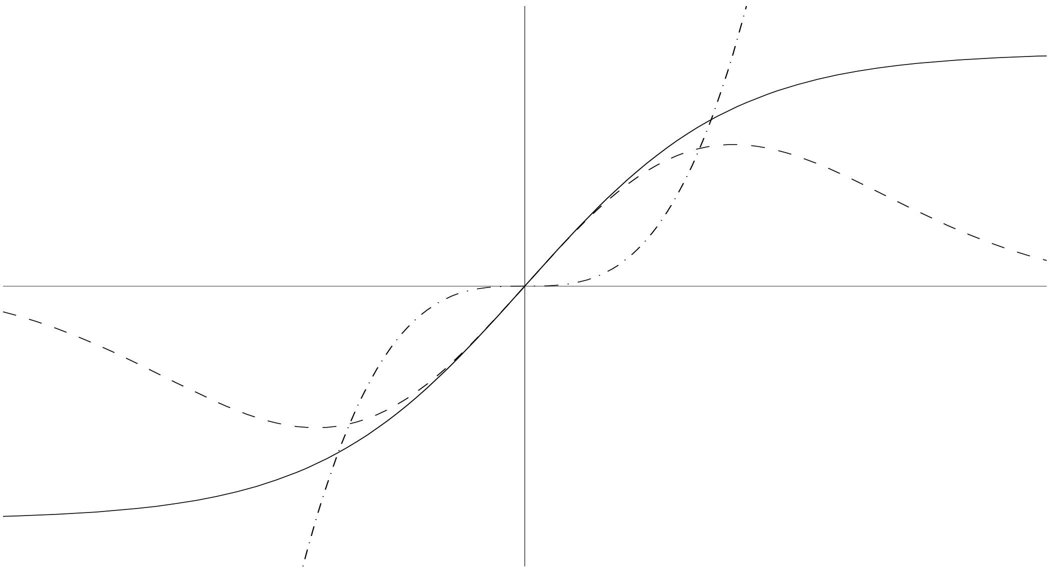


Figure 2.6: Nonlinearities for FastICA algorithm: $g_1(y) = \tanh(y)$ in solid line, $g_2(y) = -y\exp(-y^2/2)$ in dashed line and $g_3(y) = y^3$ in dash-dotted line

Theorem 9. *Assume that data follows the ICA model as in Theorem 8, and that G is a sufficiently smooth even function. then the asymptotically stable points of a gradient algorithm maximizing eq.* (2.118) *include the i-th row of the inverse of the whitened mixing matrix* **VA** *such that the corresponding independent component s_i fulfills:*

$$\mathrm{E}\{s_i g(s_i) - g'(s_i)\}\left[\mathrm{E}\{G(s_i)\} - \mathrm{E}\{G(\nu)\}\right] \tag{2.134}$$

where g is the derivative of G and ν is a standardized Gaussian variable

For a proof, see [80]. It has to be noted that if $\mathbf{w}$ equals the i-th row of $(\mathbf{VA})^{-1}$, the linear combination $\mathbf{w}^T\mathbf{z}$ equals the i-th independent component.
It is possible to formulate, now, a gradient algorithm to optimize (2.118) under the constraint of $\|\mathbf{w}\| = 1$:

$$\Delta\mathbf{w} \propto \gamma\mathrm{E}\left\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\right\} \tag{2.135}$$

$$\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\| \tag{2.136}$$

where the self adaptation constant γ can be chosen, according to Theorem 9 in the following way:

$$\gamma = \mathrm{sign}(\mathrm{E}\{yg(y) - g'(y)\}) \tag{2.137}$$

However there is a faster way of performing this optimization, using a fixed point algorithm. In fact, Newton methods are known to be faster than gradient based techniques, as long as the computations performed at each iteration are not excessively demanding, and this is the case. For the negentropy optimization, in fact, it is possible to approximate the Hessian matrix in such a way that it is not necessary to perform matrix inversion to reach the solution. Moreover, Newton methods do not require the setting of learning parameters as gradient techniques, and this makes the algorithms more "easy" to use. The fast fixed optimization of negentropy starts from the use of Lagrange multipliers. Consider the problem of looking for a maximum (in this case we are considering the one-unit case, so there are $2n$ local maxima) of the function $J_G(\mathbf{w}^T\mathbf{z}) = \left[\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\} - \mathrm{E}\{G(\nu)\}\right]^2$, then the maxima will be in certain optima of $\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\}$. In fact, as stated by Theorem 8, both local maxima and minima of $\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\}$ are related to a solution of the ICA problem, so the problem of maximizing independence becomes the problem of finding the stable point of $\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\}$. It is possible to look for these optima in a very fast way by means of Lagrange multipliers. Consider the Lagrangian function of the problem in eq. (2.128):

$$L(\mathbf{w}, \lambda) = \left[\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\} - \mathrm{E}\{G(\nu)\}\right] + \lambda(\|\mathbf{w}\|^2 - 1) \tag{2.138}$$

where the square sign in negentropy approximation has been removed due to the fact that it does not alter the local optima. The desired points $\mathbf{w}^*$ and λ^* are those where the gradient of the Lagrangian with respect to $\mathbf{w}$ and to λ is zero:

$$\begin{cases} \left.\dfrac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}}\right|_{(\mathbf{w},\lambda)=(\mathbf{w}^*,\lambda^*)} = \mathrm{E}\{\mathbf{z}g(\mathbf{w}^{*T}\mathbf{z})\} + 2\lambda^*\mathbf{w}^* = \mathbf{0} \\ \left.\dfrac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda}\right|_{(\mathbf{w},\lambda)=(\mathbf{w}^*,\lambda^*)} = \|\mathbf{w}^*\|^2 - 1 = 0 \end{cases} \tag{2.139}$$

To obtain the points where the gradient is zero, it is possible to implement a Newton iteration considering the gradient of the Lagrangian in equation (2.139) and the Hessian matrix computed as:

$$\frac{\partial^2 L(\mathbf{w}, \lambda)}{\partial \mathbf{w}^2} = \mathrm{E}\{\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\} + 2\lambda\mathbf{I} \tag{2.140}$$

The update rule for $\mathbf{w}$ is:

$$\mathbf{w} \leftarrow \mathbf{w} - \left(\frac{\partial^2 L(\mathbf{w}, \lambda)}{\partial \mathbf{w}^2}\right)^{-1} \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} \tag{2.141}$$

that in this case becomes:

$$\mathbf{w} \leftarrow \mathbf{w} - \left[\mathrm{E}\{\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\} + 2\lambda\mathbf{I}\right]^{-1} \cdot \left(\mathrm{E}\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} + 2\lambda\mathbf{w}\right) \tag{2.142}$$

Since data are spehered, it is possible to do a key approximation in the algorithm, that allows inverting easily the Hessian matrix (2.140):

$$\mathrm{E}\{\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T\mathbf{z})\} \approx \mathrm{E}\{\mathbf{z}\mathbf{z}^T\}\mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\} = \mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\}\mathbf{I} \tag{2.143}$$

that allows eq. (2.142) to be written as:

$$\mathbf{w} \leftarrow \mathbf{w} - \left[\mathrm{E}\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} + 2\lambda\mathbf{w}\right] / \left[\mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\} + 2\lambda\right] \tag{2.144}$$

Remembering that vector $\mathbf{w}$ is constrained to be unit norm, at each iteration, its standard deviation can be removed:

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|} \tag{2.145}$$

It is possible to further simplify the algorithm, considering the standard deviation removal at each iteration, by multiplying at each iteration both sides of (2.144) by the scalar $\mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\} + 2\lambda$, obtaining:

$$\mathbf{w}\left[\mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\} + 2\lambda\right] = \mathbf{w}\mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\} + 2\mathbf{w}\lambda - \mathrm{E}\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} - 2\lambda\mathbf{w} \tag{2.146}$$

and, considering that at each iteration eq. (2.145) will be applied to $\mathbf{w}$, eq. (2.146) becomes:

$$\mathbf{w} \leftarrow \mathbf{w}\mathrm{E}\{g'(\mathbf{w}^T\mathbf{z})\} - \mathrm{E}\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} \tag{2.147}$$

Therefore the optimization procedure for a *single* unit can be done by means of a very fast fixed point algorithm with normalization of vector $\mathbf{w}$ at each iteration. In the case of multi-unit, two approaches have been proposed in [79]: the deflation approach (where the components are extracted one at a time) and the symmetric approach (where the components are extracted all together). For both approaches it is necessary, in addition, to decorrelate the set of independent components from each other, to avoid being trapped in the same maxima. To do this, in the deflationary case, each component at each iteration is decorrelated by means of Gram-Schmidt orthogonalization from the subspace generated by the components already found, while in the symmetric approach, decorrelation is performed for all components at the same time. It is possible to formulate now the algorithm in both the deflation and symmetric approaches.

For what concerns the deflation approach, the main iteration for pth component consists of the following steps:

$$\begin{aligned}
\mathbf{w}_p &\leftarrow \mathbf{w}_p\mathrm{E}\{g'(\mathbf{w}_p^T\mathbf{z})\} - \mathrm{E}\{\mathbf{z}g(\mathbf{w}_p^T\mathbf{z})\} \\
\mathbf{w}_p &\leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1}(\mathbf{w}_p^T\mathbf{w}_j)\mathbf{w}_j \\
\mathbf{w}_p &= \frac{\mathbf{w}_p}{\|\mathbf{w}_p\|} \qquad (2.148)
\end{aligned}$$

until convergence for the p-th component, and then moving to component $p+1$. Convergence can be decided when the norm of the difference or of the sum of two consecutive $\mathbf{w}_p$ is below a suitable value ϵ (usually $\epsilon = 10^{-4}$). The advantage of using a deflation approach are in terms of computational load, but a drawback is that errors propagate from an extracted component to subsequent ones, making the last extracted components a bit less reliable than in the symmetric case.

In the symmetric case decorrelation is done at each iteration and for all components in a different way. The simplest way is to update unmixing matrix estimate $\mathbf{W}$ in the following way ([79]):

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W} \qquad (2.149)$$

where the inverse square root $(\mathbf{W}\mathbf{W}^T)^{-1/2}$ may be computed by means of eigenvalue decomposition of $(\mathbf{W}\mathbf{W}^T)$. Other more sophisticated ways of on-line decorrelating matrix $\mathbf{W}$ may be found in [80]. The symmetric extraction procedure consists then of the following steps:

$$\begin{aligned}
\mathbf{w}_i &\leftarrow \mathbf{w}_i\mathrm{E}\{g'(\mathbf{w}_i^T\mathbf{z})\} - \mathrm{E}\{\mathbf{z}g(\mathbf{w}_i^T\mathbf{z})\} \quad \text{for} \quad i = 1, \dots, n \\
\mathbf{W} &\leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W} \qquad (2.150)
\end{aligned}$$

It has the advantage of not “privileging” any component during extraction, but in this approach it is necessary to know the exact number of sources.

Chapter 3

Independent Component Analysis Applications

3.1 Introduction

The strength of Blind Source Separation approach comes from the possibility of dealing with problem of completely different nature. Given a set of measurements or signals, BSS techniques reveal the underlying hidden factors and therefore it is possible to design a statistical generative model of the observations. This model can be employed in such tasks as compression, denoising and pattern recognition.

In many cases the measurements are given as a set of parallel signals or images, like mixtures of simultaneous sounds that have been picked up by several microphones, brain images obtained by MRI, or several radio signals arriving at a portable phone. Due to the generality and flexibility of the source separation framework, BSS techniques have been employed in several research fields. In biomedical signal processing, and in particular in brain imaging, BSS provides helpful insight to hidden independent phenomena without making particular assumptions on their nature. BSS has also been applied in telecommunication research. In particular by means of BSS it is possible to enhance performances of code division multiple access (CDMA) receivers. In fact, the CDMA model can be interpreted as a noisy ICA model, and by properly taking the effects of fading channels and delays into account it has been shown in [149] how to enhance CDMA receiver capability of detecting desired user's symbol.

By means of BSS it is possible also to reveal hidden trends, that cannot be observed

by direct investigation, in financial time series [92].
Due to the considerable number of BSS applications, for comprehensive reviews see [80, 5, 37, 135].
Blind techniques have been applied to a statistical model of electronic devices to perform yield-oriented design [146]. In this case, the use of Independent Component Analysis comes as a natural evolution of the Principal Component Analysis approach, that is commonly employed in statistical modeling, due to the fact that parameters of the model do not have, in general, a Gaussian probability density.
ICA has been applied also to brain imaging techniques like functional Magnetic Resonance Imaging (fMRI) and MagnetoEncephaloGraphy (MEG) to extract meaningful spatial and temporal patterns related to cerebral activity in a blind fashion. In fact physiological considerations allow to expect that brain activations related to different activities are independent and observations yield linear mixing [118]. By means of independence, therefore, it is possible to extract and interpret signals without making particular assumptions on the nature of the experiment. In fact, it is known that assumptions used in hypothesis-driven methods are rather sensitive, and inaccuracy in stating them results in artifacts, poor results, or even failure in extracting the relevant activities.
In section 3.2 the statistical non-linear modeling of transistors in particular related to High Electron Mobility Transistor (HEMT) Monolithic Microwave Integrated Circuit (MMIC) will be briefly revised, while the results of modeling based on BSS techniques like PCA and ICA will be discussed in chapter 5. In sections 3.3 and 3.4 two brain imaging techniques will be described, together with a review of the signal processing strategies that are commonly employed to extract relevant information from both fMRI time-series and MEG recordings. Enhancements of such techniques, aided at "tailoring" Independent Component Analysis to specific problems, exploiting both specific and loose prior information on the sources are proposed and evaluated in Chapters 6, 7 and 8.

3.2 Statistical Non-Linear Modeling of Transistors

In recent years, several statistical models have been developed in order to perform yield-oriented design [146] of MMIC's by means of short-length GaAs and InP monolithic technological processes. In fact, their enhanced performances, obtained thanks to their very high f_T values (> 100 GHz), are joined to strong process parameter dispersion, and therefore an increased accuracy in yield evaluation and optimization is usually required.

Recently several kinds of statistical models have been proposed in literature to explain the behavior of active devices both in linear and non linear operation, based on physical description of HEMT and MESFET behavior and employing measurement databases or empiric equivalent circuits. The proposed models can be divided in two groups:

- Physics-based model (**PBM**).
- Equivalent-circuit-based model (**ECM**).

PBM models describe the behavior of the device by means of equations accounting for physical parameters (like gate length, doping density, geometrical dimensions) whose variability can be related to the production process [16]. To move from a physical model to a statistical one, it is necessary to know the probability distribution of the parameters. Due to the closeness to the physical model of the transistor, these methods are quite accurate, but they are in general excessively computationally demanding.

ECM models, on the other hand, simulate the behavior of a device from a circuital point of view [7]. The equivalent circuit representation may account for linear and non-linear components (in this case, their behavior can be modeled by means of empirical equations). To achieve an acceptable model accuracy, usually a measurement database of S parameters is used. ECM models have the advantage of requiring much less time than PBM ones, moreover they are easily implemented in MMIC CAD design tools.

Since parameters of an ECM models are usually correlated, particular care has to be taken while dealing with Monte Carlo simulations based on these models. In fact, if correlation is not properly taken into account, the Monte Carlo method will sample parameter space in a way that is inconsistent with real parameter variations, therefore producing overtly pessimistic yield estimations. Principal Component Analysis and Independent Component Analysis may be therefore employed to have an uncorrelated or even independent representation of data, in order to increase the accuracy of the model.

In fact, since empirical statistical models require the extraction of the statistical distribution of equivalent circuit parameters that are correlated, diagonalization of the correlation matrix is usually done by means of Principal Component Analysis [40, 159]. Since Independent Component Analysis performs correlation matrix diagonalization as well, with the additional characteristic of decomposing the data set into *independent* rather than *uncorrelated* components, it is reasonable to expect that the population obtained by letting IC vary according to a suitable distribution can more accurately describe the real population. In 5 an ICA decomposition of the empirical parameters has been performed, and the results have been compared with Principal Component Analysis.

3.3 Functional Magnetic Resonance Imaging (fMRI): General Concepts

Nuclear Magnetic Resonance imaging, introduced by Lauterbur in 1973 [101], was used, at first almost exclusively, for the study of neuroanatomy and neuropathology, and two more decades passed before its application to the study of brain function. One of the first approaches to MR-based brain mapping, developed at the Massachusetts General Hospital in 1991, was based on the determination of cerebral blood volume (CBV) through the quantification of the signal decrease that is brought about by the dephasing effects induced by the first passage of an intravascular contrast medium (Gd-DTPA) in non-refocused (T2* weighted) MR images [150]. The next step consisted in the determination of CBV changes induced locally by brain activity. Again in 1991, the same group reported the first functional mapping results in the visual cortex obtained during visual stimulation [25]. These findings led to a marked acceleration of the research in the field of the functional applications of MRI in the early 1990s.

The first approaches, which were based on the use of external contrast medium, were soon supplemented by techniques that utilized "endogenous" contrast effects to study activity-related MR changes. The rationale behind this non-invasive approach to fMRI was derived from the much earlier observation of Pauling [139] that a modification of the oxygenation of hemoglobin resulted in an alteration of its magnetic properties.

In 1990, Ogawa and colleagues reported that MRI was sensitive enough to show "blood-oxygenation-level-dependent" (**BOLD**) signal changes in vivo [131, 130]. In 1991, Turner et al. observed a similar effect during experimental anoxia [166]. This effect was not confined to the macroscopic vessels but spread out to the brain tissue itself. Since it was known that changes in neuronal activity were accompanied by local changes in brain oxygen content [58], it became evident that a technique based on the BOLD effect could potentially investigate neuronal activation through changes induced in tissue oxygenation.

In 1992, three groups applied BOLD-based fMRI to the human brain ([132], [98] and [14]). It rapidly became clear that with the new fMRI technique, which provided better spatial and temporal resolution than PET, a new era in functional neuroimaging had started [152]. The success of fMRI techniques depends on the basic assumption of a relation between the BOLD signal and the underlying fundamental neuronal activity. In this sense the study of Logothesis [106] has been of fundamental importance. Before that, in fact, the relation between BOLD and neuronal activity was investigated only performing fMRI studies in humans and single cell recordings in monkeys and then trying to find a relation between the results. In the pioneering study of Logothesis for the first time single cell activity and BOLD signal were recorded simultaneously in monkeys.

3.3.1 Blood Oxygen Level Dependent (BOLD) signals

Hemoglobin (Hb) is the predominant macromolecule in blood. The attachment of oxygen is dependent on the local partial pressure of oxygen (also known as oxygen tension or pO_2) and allows oxygen to be released at the tissue when local metabolic activity causes local oxygen depletion [134].
The binding of oxygen molecules determines the magnetic properties of hemoglobin. Deoxygenated hemoglobin (deoxy-Hb) has the typical properties of paramagnetic materials. Without any external magnetic field the individual magnetic moments interact weakly with one another and, due to thermal motion, are randomly oriented and the deoxy-Hb has not net bulk magnetization. Conversely, when an external magnetic field is applied the dipoles become partially aligned in the direction of the field and act to increase the magnetic field. However, for ordinary external fields strengths, and like the weaker nuclear magnetic properties, only a very small fraction of the magnetic moments will be aligned with the field and, due to the thermal motion, the contribution to the total magnetic field is very small.
The oxygenated hemoglobin (oxy-Hb) behaves like a diamagnetic material without contributing to any applied external magnetic field. [139, 134]. The bulk relaxation of blood can be assimilated to that of hemoglobin solutions, so the effects of cellular membrane of red blood cell and of the proteins in plasma appear to be limited. Deoxy-Hb does not affect the longitudinal relaxation time (T1) but enhances the spin phase dispersion, and thereby affects the transverse relaxation time (T2) and especially the non-refocused transverse relaxation (T2*).
In fact, the presence of the paramagnetic deoxy-Hb in red blood cells confers to them a different magnetic susceptibility from near tissue and plasma. The application of an external magnetic field produces microscopic magnetic inhomogeneities which also depend from the geometry of the cell (Figure 2.1). The presence of this non-uniform magnetic field, within a volume element (voxel), is known to influence both

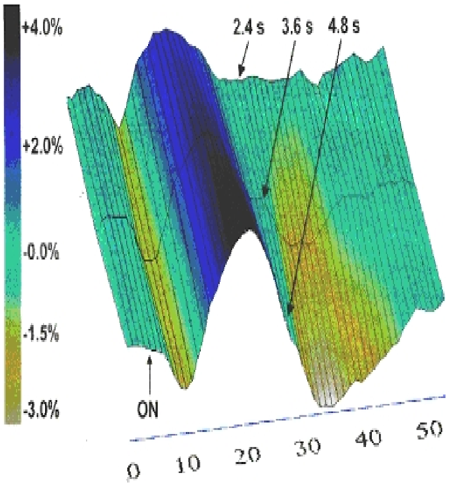


Figure 3.1: BOLD time-courses as detected at 4 T (x -axis: number of scans TR=0.3s). Figure from [77]

T2*- and T2- relaxation times :

- **T2***: the net signal is decreased because the tissue-water protons precess at slightly different Larmor frequencies, depending on the local magnetic field perturbations.
- **T2**: the net signal is decreased because of the irreversible dephasing of tissue-water protons that takes place because of their diffusion in the inhomogeneous magnetic field, during the time of image-acquisition.

Therefore, the class of sequences that do not use re-focusing pulse to cancel the spin phase dispersion induced by the inhomogeneities (Gradient Echo - GRE), are very sensitive to the T2* effect. Conversely, the class of sequences that use re-focusing pulses (Spin Echo - SE) are more sensitive to the T2 effect [15, 13]. A sensory, motor or cognitive stimulus produces a localized increase in neural activity, which in turn induces an increase of oxygen extraction, cerebral blood flow (CBF), and cerebral blood volume (CBV). The mechanism underlying these effects, caused either by the production of metabolites or due to a more direct effect on local blood vessels is poorly understood. Whatever the cause, because CBF (and hence oxygen delivery) changes exceed CBV changes by two to four times, while blood oxygen extraction increases only slightly [97], the total paramagnetic content of deoxy - Hb within

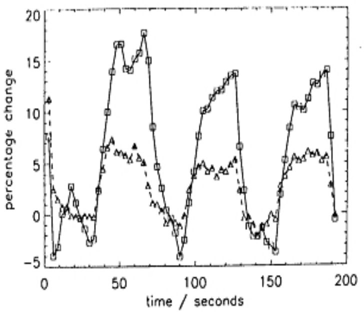


Figure 3.2: BOLD signal changes for a periodic visual stimulation at 1.5 T and at 4.0 T. (Figure from [166])

brain tissue voxels decreases with brain activation. The resulting difference leads to less intravoxel dephasing compared to a resting state and results in increased signal on T2/T2* - weighted images. In BOLD fMRI terms, the initial increase in deoxy-Hb content results in a signal decrease in T2* sensitive images, while the late deoxy-Hb decrease is equivalent to a signal increase. This fits well with the typical fMRI signal time-course, especially if experiments with high temporal resolution are performed at high field strengths (4 T). In such a case [122, 77, 100], two different regions were recognized inside the visual cortex. A smaller region accounted for 30% of the total area of activation and showed a biphasic behaviour with a rapid (0.5-2s) negative response of about 1% of the total signal intensity (the "initial dip") and a later positive change of about 2% (see Figure 3.1). The larger region of activation occupied the remaining 70% of the total area and showed only a late (5-8s) positive signal change, with a time-course of the positive response that was similar in shape to that of the smaller area, but showed a higher amplitude (up to 6% of baseline level). A similar finding of a rapid negative response to activation was described using MR spectroscopy, but it is probably related to other mechanisms than BOLD, since this study reported a decrease with increasing echo times, while the dephasing effects of the deoxy-Hb should increase with echo time [52, 50]. A central question regarding the possibility to perform BOLD - fMRI studies using clinical MR - scanner, is related to the dependence of BOLD signal changes from the static magnetic field intensity (B_0). Several experimental and simulation studies, show that transverse relaxation rate depend more than linearly from B_0. Most of the fMRI studies are, at the date, performed at 1.5 Tesla or 2.0 Tesla; reported mean percent signal changes values highly depend on the type, duration and intensity of the stimulus, ranging between 1% and 8% in the case of sensory stimulation. Groups using 4.0 Tesla magnets report from 5% to 20% mean percent signal changes. At field intensities lower than 1.5 T the stimulus-related inhomogeneity effects may be too low to be detected. The functional spatial resolution of BOLD-fMRI is ultimately limited by the specificity of neurovascular system and of its local changes in response to neural activity. Experimental evidences of optical imaging studies suggest that in the cerebral cortex this specificity is smaller than 1 mm [70, 169, 122, 114]. Fast imaging MR-techniques, such as Echo Planar Imaging, allow the collection of BOLD-sensitive image time-series with a sampling rate sufficient to study the temporal features of task-related haemodynamic response [177, 30]. Figure 3.3 shows the time-course of the fMRI signal change in primary visual cortex (V1) in response to a typical visual stimulus of 10 s duration (shown as box from 2 to 12 s) measured as the average of 54 single trials at a magnetic field strength of 1.5 Tesla [121]. After the stimulus presentation, the signal shows:

- a 2 - 3 s delay (i) that reflects the inertia of the local haemodynamics;
- a signal increase reflecting the stimulus-induced hyperoxygenation with a time-to-peak (ii) that, in this case, is about 13 s.
- a signal decrease, after stimulus cessation, with a return to the baseline level, that is about 9 s (iv-ii).

However, it must be emphasized that the fMRI response is proportional to the local average neural activity, averaged over a small region of the brain and averaged over a period of time and that this spatial and temporal averaging may be different in different brain areas, particularly because the vascular system seems to be specialized in particular brain areas [31]. Therefore, the reported values of the delays are not fixed and may highly change depending on stimulus type, contrast, duration and the stimulated brain area. The limitations of temporal resolution in fMRI, like the spatial resolution, are determined not by the lack of technological capabilities but by the nature of the process underlying the neuronal-induced MR signal. Indeed, the functional temporal resolution (defined as the smallest interval between two detectable separate tasks for the same area) is limited by the duration of the baseline-recovery after the stimulus offset. If individual stimuli are presented more

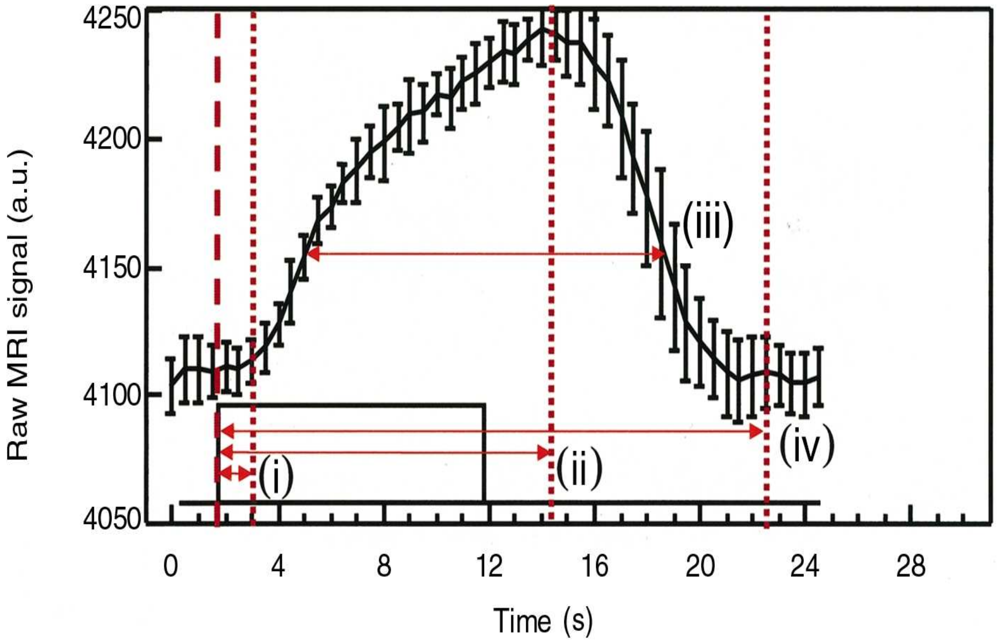


Figure 3.3: The time course of fMRI signal change in visual cortex in response to a visual stimulus. (figure from [121])

rapidly than the width of the hemodynamic response then contrast is reduced because the signal does not have sufficient time to return to the resting level. At 1.5 T, Bandettini et al in [12] were able to separate task-related BOLD time-courses with a stimulus switching frequency of 0.062 Hz (i.e., 8.0 s movement and 8.0 s control) but not with a stimulus switching frequency of 0.125 Hz, suggesting that temporal resolution cannot be better than 8 s. Using high magnetic fields (4 Tesla) with higher SNR and larger BOLD effects, the temporal resolution of the fMRI signal from the motor area during repeated tasks was found to be about 5s [91]. In addition to the activity evoked by the stimulus, both instrumental noise and fluctuations of biological origin influence fMRI time-series. Instrumental noise includes Johnson and thermal noise, which originate on the scanner hardware. This type of noise, that can be correctly considered to be white, effects the background voxels of the images as well as the intra-brain voxels and determines the intrinsic SNR of the MRI images. Physiological noise include both periodic and non periodic signal fluctuations and is the main limiting factor in most task-activation experiments. A detailed understanding of the various sources of biological fluctuations, as would be necessary to get the full range of information from fMRI data, has up to this time been lacking. However, an early report [172] experimentally analyzed the spectral components forming the fMRI noise, using rapid single-slice EPI acquisitions (TR=122 ms) without any stimulation. In background voxels, noise was wide-band and well below the level of the noise in the cortical regions. At low frequencies ($<$ 0.5 Hz), the noise fluctuations were larger in the grey-matter than in the white matter; at higher frequencies ($>$ 2.5 Hz) these differences cancelled. High peaks at the typical frequencies of the cardiac (0.8 - 1.2 Hz) and respiration (0.25 - 0.5 Hz) cycles and at their harmonics were present in the power spectrum in all intra-brain voxels, but mainly in the gray matter voxels. Periodic blood flow and pulsatile bulk motion caused by the heart cycle and a generalized variation in blood oxygenation and bulk displacement of the head in response to breathing are considered to be the main origin of these periodic signal changes [76]. Beside respiratory and cardiac related noise, an intense peak at low frequencies ($\sim$ 0.1 Hz) is usually present in the gray matter. Several interpretation have been proposed about the nature of these 0.1 Hz oscillations. Hyde and Biswal [78] showed that there is a considerable amount of synchrony of low-frequency physiological fluctuations during rest, and that performance of the task increases the overall synchrony. Starting from this observation, the authors correlate the low frequency fluctuations to the functional connectivity in spatially distributed cortical patterns. Mitra et al. [123] also observed that these 0.1 Hz oscillations are characterized by a complex space-time structure, but concluded that their origin was likely to origin on the vasomotor oscillations that exist ubiquitously in blood vessels all over the body, and do not necessarily have any connection with neural activity [123]. Non periodic noise components, typically manifesting as “drifts and shifts” have been often observed in fMRI time series. These fluctuations, at very low frequencies (0.0-0.015 Hz), have been attributed to long term physiological shifts and/or to movement related noise remaining after re-alignment. Finally, it must be pointed out that, if whole brain multislice EPI are used, the temporal sampling rate is typically several seconds. This is above the Nyquist limit for both cardiac and respiratory noise and so aliasing will occur, affecting the spectrum of observed signals.

3.3.2 fMRI Data Analysis

Introduction

Figure 3.4 depicts the flow-diagram of processing steps for the detection and representation of functional and structural information. Input data are the 4D data sets of functional time series that are collected with BOLD sensitive sequences, and 3D data sets that are collected with 2D or 3D conventional sequences and serve as anatomical reference for the visualization of functional information. The left side of the graph represents the steps which are required for extracting the functional information. Some of them (e.g. realignment, spatial and temporal filtering) aim at enhancing the effects of the stimulus-related signal and to reduce the influence of the artifactual signal fluctuations. Others serve to detect localized task-dependent signal changes as visualized by activation maps (3.3.3). Right side of the graph in Figure 3.4 represents the steps which are required for the representation of the functional information with respect to the individual brain anatomy. In fMRI, differently than in PET or MEG, the functional and structural images, are recorded within the same measurement session, therefore the respective data sets are easily co - registered. Spatial normalization refers to the transformation of anatomical and functional data in conventional reference spaces and is essential to compare the results between different subjects and to facilitate communication among laboratories.

Preprocessing

Realignment

Subjects' motion poses a severe problem for the analysis of functional data. Despite the use of physical constraints, head movements cannot be completely excluded during functional scanning. Small head movements (< 1 mm) also produce effects that can mask the relatively small BOLD signal changes and should be corrected using re-alignment algorithms. Let us consider $I_i(\mathbf{x})$ and $I_k(\mathbf{x})$ as two images (2D or 3D) collected at time i and k within a series of T repeated functional measurements. Let us suppose that $I_i(\mathbf{x})$ and $I_k(\mathbf{x})$ are related by a geometric transformation $\mathbf{T}[\mathbf{x}]$, so that:

$$I_k(\mathbf{T}[\mathbf{x}]) \approx I_i(\mathbf{x}) \tag{3.1}$$

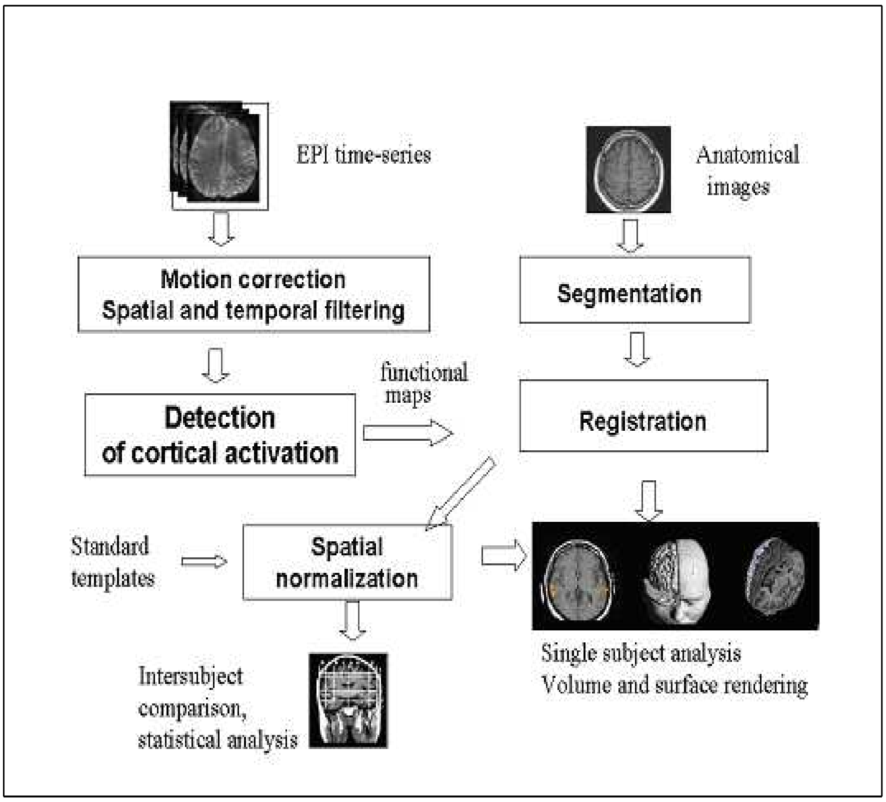


Figure 3.4: Data-flow of processing steps for the detection and the representation of functional and anatomical information

The problem the realignment algorithms deal with is to find the transformation $\mathbf{T}$ that minimizes the differences, due to the subject's motion, between the two images. The most commonly adopted algorithms are based on iterative computation of the rotation-translation parameters that reduce the mismatch between a reference image (e.g. the middle scan of the time-series) and the other images of the time-series [173, 61, 71, 65]. These realignment procedures are based on the following steps:

- measure of the spatial discrepancy between the transformed image $I_k(\mathbf{T}[\mathbf{x}])$ and the reference image $I_{T/2}(\mathbf{x})$;
- evaluation of the parameters that define T;
- evaluation of the new values of I_k after T has been determined (interpolation method).

A robust method, commonly adopted in both PET and fMRI data analysis, considers $\mathbf{T}[\mathbf{x}]$ to be a roto-translation transformation based on the rigid-motion hypothesis [173]. With this hypothesis, transformation $\mathbf{T}[\mathbf{x}]$ is defined by 3 parameters in the case of realignment of 2D images (2 translation offsets and 1 rotation angle) and by 6 parameters, in the case of 3D images (3 translation offsets and 3 rotation angles).

Spatial and Temporal filtering

Spatial and temporal filtering of fMRI time series aim at enhancing the functional contrast-to-noise ratio (see section 7.3.4), reducing the effects of the confounding factors that arise from the instrumentation and from spontaneous physiological activity. High-spatial-frequency noise, mainly from the scanner electronics, can be attenuated by spatially "smoothing" the fMRI time-series with bidimensional low-pass filters. Let us express the acquired data as:

$$\mathbf{I}_i(\mathbf{k}) = \mathbf{S}_i(\mathbf{k}) + \mathbf{E}_i(\mathbf{k}) \tag{3.2}$$

where $\mathbf{k} = (k_x, k_y)$ is the spatial frequency span, $\mathbf{S}_i(\mathbf{k})$ is the frequency domain representation of the desired image at scan i and $\mathbf{E}_i(\mathbf{k})$ is the noise contribution (physiological and electronic). The underlying assumption of spatial smoothing is that $\mathbf{S}_i(\mathbf{k})$ is a monotonically decreasing function of $\mathbf{k}$ and thus there exists some frequency $\mathbf{k}_c$ such that

$$\mathbf{S}_i(\mathbf{k}) \ll \mathbf{E}_i(\mathbf{k}) \quad \text{for} \quad \mathbf{k} > \mathbf{k}_c \tag{3.3}$$

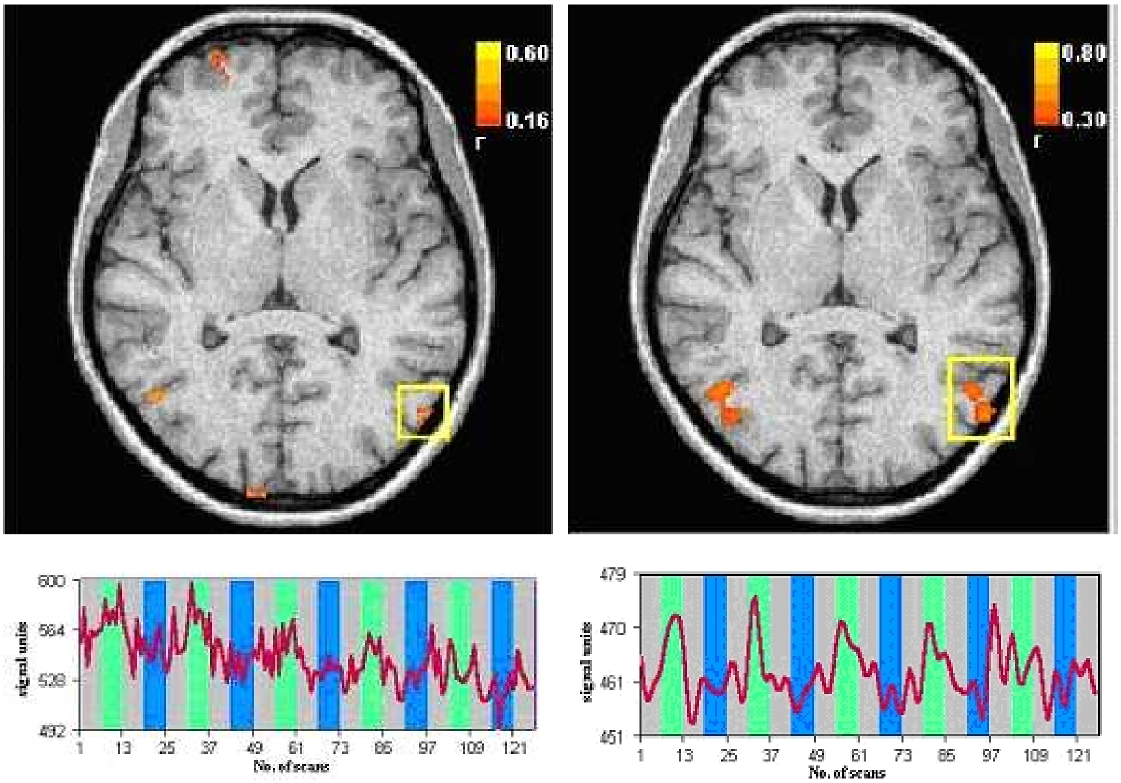


Figure 3.5: Effects of temporal and spatial smoothing: correlation maps and time-courses for unfiltered (left) and for spatially (2D Gaussian filter with FWHM = 2 pixels) and temporally (1D Gaussian filter with FWHM= 3 samples) smoothed (right) EPI time series. Green background indicates stimulation.

Thus, a good filter would be a function $\mathbf{H}(\mathbf{k})$ such that:

$$\mathbf{H}(\mathbf{k}) \approx \begin{cases} 1 & \text{for} \quad \mathbf{k} < \mathbf{k}_c \\ 0 & \text{for} \quad \mathbf{k} > \mathbf{k}_c \end{cases} \tag{3.4}$$

Then, multiplying $\mathbf{H}(\mathbf{k})$ by $\mathbf{I}_i(\mathbf{k})$ would result in noise suppression with minimal effect on the image function $\mathbf{S}_i(\mathbf{k})$. However, the presence of regionally specific activation implies that high-spatial-frequency components of the signal are present in $\mathbf{S}_i(\mathbf{k})$ as well and that, besides reducing the noise contribution, the effect of spatial smoothing will be a decrease in effective spatial resolution. These two contrasting effects will influence the detection of activation regions and have to be balanced. Indeed, when the activated brain regions extend over clusters of several voxels, the spatial smoothing, due to the correlation between the time-courses, will strengthen the signal relative to the noise (Figure 3.5). Conversely, when focal regions of the brain are activated, they might no longer be discernible after spatial smoothing. Furthermore, according to the matched filter theorem, the signal is best detected by smoothing with a filter whose width matches that of the signal. In practical cases, however, since both focal and broad activation regions may be present in the same data set and their real extension cannot be known, the width and type of spatial filter is chosen on the basis of a trade-off between the spatial resolution and the expected enhancement of functional contrast-to-noise ratio [107]. For inter - subject studies, on the other hand, a high degree of spatial smoothing has to be applied in order to reduce the anatomical differences between subjects and to allow the correct use of statistical tools.

3.3.3 Hypothesis-Driven approach: introduction

Historically, the first methods to analyze fMRI datasets were based on correlation analysis and subtraction. Since fMRI signals have no simple quantitative physiological interpretation, usually the signal at a given voxel during the active part of the experiment is compared with its value during a period of rest. The most simple hypothesis-driven technique is a correlation procedure, followed by some statistical test, to point out the voxels in the active areas (3.3.3), while, in the case of more complex experimental paradigms, a General Linear Model (GLM) is used (3.3.4).

Detection of activations using voxel-based methods

In the first fMRI studies a simple method based on image subtraction was used to create descriptive images of the task-dependent brain areas [132, 98, 14]. According to the "pure insertion" hypothesis, voxels with a high gray level in the "difference" image, formed by subtracting the "control" from the "task" condition images, reflect the areas with a task-induced differential activation. With this method, the value of the intensity threshold between activated and non-activated voxels is arbitrarily chosen. Furthermore, image subtraction is very sensitive to movement-related effects and to other unexpected signal changes. More reliable activation maps are produced by using statistical methods. The commonly used Student's *t*-test maps are formed by computing, on a voxel-by-voxel basis, the value of statistical significance for the difference of the means between two conditions. Voxels with a significance value below a given threshold (e.g. $p < 0.001$) are considered activated by the task [20]. The *t*-test method, may be regarded as a weighted subtraction method and the time dependency of signal intensity, which might be potentially useful, is ignored [91]. Furthermore, it might be shown that the *t*-test method is a simple case of correlation analysis, which will be discussed in the following.

Let the discrete random sequence $f_t, (t = 1, \ldots, T; \mathrm{T} = \text{number of scans})$ denote the *observed* fMRI measurements (after pre-processing steps) at a given voxel. The observed sequence is assumed to be of the form:

$$f_t = \mu s_t + e_t \tag{3.5}$$

where s_t is a deterministic (non-random) activation signal at the voxel under consideration, μ is a non-negative constant factor and e_t represents added Gaussian white noise with unknown variance σ^2. Eq. (3.5) can be written in vector notation as follows:

$$\mathbf{f} = \mu\mathbf{s} + \mathbf{e} \tag{3.6}$$

where $\mathbf{f}$, $\mathbf{s}$, and $\mathbf{e}$ are T - dimensional vectors that are represented by $T \times 1$ (column) matrices. When $\mu = 0$, the voxel is not activated and the observed sequence is merely noise ($\mathbf{f} = \mathbf{e}$). When $\mu > 0$, the voxel is said to be activated. Given the observed sequence $\mathbf{f}$ and the model in Eq. (3.6) the problem of deciding whether or not the voxel under consideration is activated reduces to test the null hypothesis :

$$\{H_0 : \mu = 0\} \tag{3.7}$$

versus the alternative hypothesis:

$$\{H_1 : \mu > 0\ \} \tag{3.8}$$

In the so-called correlation maps, hypothesis testing is performed through the computation of the cross-correlation coefficient (cc) between the time-course of the voxel intensity ($\mathbf{f}$) and a model of the BOLD response induced by stimulation $\mathbf{s}$ [12], [64]. The cross-correlation coefficient is given by:

$$cc = \frac{\sum_{t=1}^{T}[(f_t - \mu_f)\cdot(s_t - \mu_s)]}{\sqrt{\sum_{t=1}^{T}(f_t - \mu_f)^2 \cdot \sum_{t=1}^{T}(s_t - \mu_s)^2}} \tag{3.9}$$

where μ_f denotes the average value of vector $\mathbf{f}$ and μ_s the average value of vector $\mathbf{s}$. Equivalently, (3.9) can be written, in vector notation, as follows:

$$cc = \frac{\mathbf{d}_f^T \cdot \mathbf{d}_s}{\mid \mathbf{d}_f \mid\mid \mathbf{d}_s \mid} \tag{3.10}$$

where

$$\begin{aligned}\mathbf{d}_f &= \mathbf{f} - \boldsymbol{\mu}_f\\ \mathbf{d}_s &= \mathbf{s} - \boldsymbol{\mu}_s\end{aligned} \tag{3.11}$$

with μ_f and μ_s vectors of length T. The value of cc, which always varies between -1 and $+1$, represents a measure of the similarity between the shape of the gray level time-course in the functional images time-series of the voxel and the expected BOLD signal enhancement induced by the stimulus. Thus, a high value of cc is expected for voxels related to activated brain areas; conversely a low value of cc is expected for all the other voxels. Separation of activated and non-activated voxels and creation of activation maps are achieved by imposing a threshold value TH for cc. In voxels where $cc < TH$, the null hypothesis is accepted and the corresponding positions are not displayed in the map; in voxels where $cc > TH$, the null hypothesis is rejected and a color code is assigned to the map. The statistical significance for acceptance of signals based on the selected threshold TH is determined by considering the transformation:

$$t = cc\frac{\sqrt{T - T_0}}{\sqrt{1 - cc^2}} \tag{3.12}$$

Eq. (3.12) establishes a one-to-one correspondence between the cross-correlation coefficient cc given in (3.10) and the values of a random variable t. Under the null hypothesis $\{H_0 : \mu = 0\}$, t will have a central Student's t-distribution with $T - T_0$ degrees of freedom. Under the alternative hypothesis $\{H_1 : \mu > 0\}$, t will have a non-central Student's t-distribution with $T - T_0$ degrees of freedom and non centrality parameter $\mu\sqrt{\mathbf{d}_s^T\mathbf{d}_s}/\sigma$. Therefore, for a given value TH, the probability of a type I error (the probability of declaring a voxel activated when it is in fact not), is given by :

$$p = \int_{t_{TH}}^{\infty} f_t(u)du \tag{3.13}$$

where $f_t(u)$ is the Student's t pdf with $T - T_0$ degrees of freedom and t_{TH} is the value obtained by substituting $cc = TH$ in (3.12). The reduction in degrees of freedom (T_0) is due to the number of free parameters in the model being employed. Assuming that, in addition to the signal component and to the mean, a drift component has been estimated and subtracted from the time-series, $T_0 = 3$. [8]. Performance of cross-correlation method strongly depends on the likeness between the reference function s and the *real* shape of the BOLD response. A simple *box-car* (ON/OFF) ideal vector which assumed a value of 0 during the resting period and 1 during the stimulation response, was used in the earlier fMRI studies [12]. The vector obtained by convolution of the ideal ON/OFF vector with an experimentally derived impulse response of the hemodynamics have been shown to be a good approximation of BOLD response, thus enhancing performance of correlation method. It might be shown, using (3.10) and (3.12) that, when an ideal box-car vector is used as reference, the cross-correlation coefficient cc and the t-test used in many fMRI studies (see [20]) lead to coincident results, and there is no advantage from using one or another. When a reference vector nearer to the actual stimulus-related signal changes is used, the advantage of using correlation analysis relies on the possibility of appropriately taking into account the transitions between the ON and OFF state and the shape of the BOLD response. This information is conversely ignored with t -test. With correlation analysis, maps of the hemodynamic delays in different activated areas may also be produced. This is simply achieved by computing additional correlation values between the time-courses and the reference vector shifted to the right of one or more samples.

3.3.4 General Linear Model (GLM)

The correlation analysis described above is adequate for fMRI paradigms with one stimulation and one control condition. For some experiments, however, a more general approach is required.

The General Linear Model (GLM) or Multiple Regression Analysis is a statistical tool that was first introduced to functional imaging data analysis by Friston et al. [62, 63]. A general linear model "explains" or "predicts" the variation of the observed time-course in terms of a linear combination of several *regressor* variables plus an error term:

$$y_t = s_{t_1}\beta_1 + s_{t_2}\beta_2 + \dots s_{t_L}\beta_L + e_j \tag{3.14}$$

where $y_t, (t = 1, \dots, T; \quad T =$ number of measurements) is the observed signal time course at a given voxel, s_{t_l}, $(l = 1, \dots, L; L < N)$ are a set of L *explanatory* variables or *predictors* (functions of measurements), β_l are the unknown parameters (or regressor values), one for each predictor, and the e_j are error terms which are assumed to be independent and identically normally distributed with zero mean and variance σ^2.

Writing (3.14) for each observation t gives the equation system:

$$\begin{cases} y_1 = s_{1_1}\beta_1 + s_{1_2}\beta_2 + \dots s_{1_L}\beta_L + e_1 \\ y_2 = s_{2_1}\beta_1 + s_{2_2}\beta_2 + \dots s_{2_L}\beta_L + e_2 \\ \quad \dots \\ y_T = s_{T_1}\beta_1 + s_{T_2}\beta_2 + \dots s_{T_L}\beta_L + e_T \end{cases} \tag{3.15}$$

or, in matrix notation:

$$\mathbf{y} = \mathbf{S}\boldsymbol{\beta} + \mathbf{e} \tag{3.16}$$

Here, $\mathbf{y}$ is the $T \times 1$ column vector of the observations, $\mathbf{S}$ is the $T \times L$ matrix of the predictors (one row per observation, one column per model parameter); $\boldsymbol{\beta} = [\beta_1, \dots, \beta_L]^T$ is the $L \times 1$ column vector of parameters, and $\mathbf{e}$ is the $T \times 1$ column vector of error terms.

The matrix $\mathbf{S}$ is defined as the *design matrix* of the experiment. fMRI studies which contain a baseline condition as well as several repetitions of one or more experimental conditions, may be easily expressed as a multiple regression problem through defining an appropriate form for $\mathbf{S}$. For instance, for an experimental design with a baseline condition and five different stimulation conditions the design matrix $\mathbf{S}$

has five columns, and one row for each measurement time point. Each predictor is build by convolution of the ideal ON/OFF response with a model of the hemodynamic response. Other effects than expected task-related BOLD enhancement may be modeled in the design matrix. Including one column consisting of 1's for all measurements will account for the mean value of the voxel time course; including one column with linearly increasing values will account for linear trends in voxel time-course. It can be shown that most methods, including simple t-test, ANOVA and correlation can be regarded as a special cases of GLM analysis that, thus, integrates parametric voxel-based analysis into a general frame [74]. Once the design matrix has been defined, next step of the GLM analysis is the estimation of the regression weights β_l such that the predicted values $\mathbf{y}'$ are as close as possible to the measured values $\mathbf{y}$ at each time point. Let us denote with $\mathbf{y}'$ the estimate of the time-course $\mathbf{Y}$ for the regression values $\boldsymbol{\beta}'$:

$$\mathbf{y}' = \mathbf{S}\boldsymbol{\beta}' \tag{3.17}$$

and indicate with

$$\mathbf{e} = [e_1, e_2, \ldots, e_N] = \mathbf{y} - \mathbf{y}' = \mathbf{y} - \mathbf{S}\boldsymbol{\beta}' \tag{3.18}$$

the residuals errors. In the GLM, the Least Squares method is used for estimating the regression weights such that the residual sum of squares

$$\mathbf{S}_r(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left(y_t - s_{t1}\beta_1 - \ldots - s_{tL}\beta_L\right)^2 \tag{3.19}$$

is minimized. This occurs when:

$$\frac{\partial \mathbf{S}_r}{\partial \beta_l} = \sum_{t=1}^{T} (-s_{tl}) \left(y_t - s_{t1}\beta_1 - \ldots - s_{tL}\beta_L\right) = 0 \tag{3.20}$$

If the model is correct and the errors are normal, the least squares estimates are the maximum likelihood estimates and are the Best Linear Unbiased Estimates, i.e. of all linear parameter estimates consisting of linear combinations of the observed data whose expectation is the true value of the parameters, the least squares estimates have the minimum variance. The mean value and the variance of $\boldsymbol{\beta}'$ are respectively :

$$\mathrm{E}\{\boldsymbol{\beta}'\} = \boldsymbol{\beta} \tag{3.21}$$

$$\mathrm{Var}\{\boldsymbol{\beta}'\} = \sigma^2 \left(\mathbf{S}^T\mathbf{S}\right)^{-1} \tag{3.22}$$

Once the regressor coefficients have been estimated, analogously to correlation analysis, a statistical test is required to assess general linear hypotheses. The extra sum of squares principle provides such a test. Let us suppose to have a full model with parameter vector which can be partitioned into two:

$$\boldsymbol{\beta} = \left[\boldsymbol{\beta}_a^T \mid \boldsymbol{\beta}_b^T\right] \tag{3.23}$$

with corresponding partition of the design matrix

$$\mathbf{S} = [\mathbf{S}_a \mid \mathbf{S}_b] \tag{3.24}$$

Let us suppose we wish to test the hypothesis $\{H_b : \boldsymbol{\beta}_b = 0\}$, i.e. we wish to test whether or not the conditions corresponding to the $\mathbf{S}_b$ predictors had some effect on a voxel time-course. When H_b is true, the model reduces to

$$\mathbf{y} = \mathbf{S}_a\boldsymbol{\beta}_a + \mathbf{e} \tag{3.25}$$

Let us denote the residual sum of squares for the full models by $\mathbf{S}_r(\boldsymbol{\beta})$ and for the reduced models by $\mathbf{S}_r(\boldsymbol{\beta}_a)$ respectively. The extra sum of squares due to $\boldsymbol{\beta}_b$ after $\boldsymbol{\beta}_b$ is then defined as

$$\mathbf{S}_r(\boldsymbol{\beta}_a \mid \boldsymbol{\beta}_b) = \mathbf{S}_r(\boldsymbol{\beta}_a) - \mathbf{S}_r(\boldsymbol{\beta}) \tag{3.26}$$

Under H_0, $\mathbf{S}_r(\boldsymbol{\beta}_a \mid \boldsymbol{\beta}_b) \sim \sigma^2\chi^2$ independently of $\mathbf{S}_r(\boldsymbol{\beta})$, with $L_b = \text{rank}(\mathbf{S}) - \text{rank}(\mathbf{S}_a)$ degrees of freedom.

Therefore, under H_0, the ratio:

$$F = \frac{\left(\mathbf{S}_r(\boldsymbol{\beta}_a) - \mathbf{S}_r(\boldsymbol{\beta})\right)/L_b}{\mathbf{S}_r(\boldsymbol{\beta})/(T - L_b - 2)} \tag{3.27}$$

has a central F distribution with $n_1 = L_b$ and $n_2 = T - L_b - 2$ degrees of freedom [8, 74]. Again, the reason for subtracting the two extra degrees of freedom from the denominator is that we are assuming that the mean and linear drift component have been estimated and subtracted from the time- series. Note that when H_0 is not true then $\mathbf{S}_r(\boldsymbol{\beta}_a \mid \boldsymbol{\beta}_b)$ has a non central chi-squared distribution, still independent of $\mathbf{S}_r(\boldsymbol{\beta})$. In summary, for a set of experimental conditions, statistical maps may be produced using the following steps:

1. Calculate, for each voxel, the statistic F of Eq. (3.27)

2. For a fixed value of false alarm rate p determined by :

$$p = \int_{F_0}^{\infty} f_F(u) du \tag{3.28}$$

   where $f_F(u)$ is an F distribution with n_1 and n_2 degrees of freedom, compare F with F_0.

3. Color - code the voxels where $F > F_0$.

3.3.5 Data-Driven approach: introduction

Data-driven approaches are complementary to hypothesis-driven techniques, seen in section 3.3.3. In fact, while the latter look for voxels that "behave" like a predetermined activation model, which means that they are substantially meant for accepting or rejecting currently formulated psychological models, the former are more suited in "exploratory" analysis, meaning that no specific hypotheses are made on spatio-temporal activity patterns.

The information on the activity of the voxels within an experimental paradigm may not be precise and, moreover, a GLM will not be able to point out all those activities in the brain that are not modeled within its design matrix. This means that unanticipated or counterintuitive time courses of activation of localized brain areas are less likely to be found with such a technique. On the contrary, the strength of a data-driven approach comes from the fact that no particular hypothesis is made on the time courses, being able therefore to point out also those activities whose time courses could not be anticipated before the experiment.

The most used exploratory approaches in fMRI are PCA (2.3.4), ICA (chapter 2) and clustering in temporal domain. For what concerns temporal clustering, all the voxel that have similar waveforms are collected together within a cluster, where the "similarity" between two time-courses is estimated by means of some suitable measure [59]. In [144] a comprehensive review of data-driven techniques is provided.

Both PCA and ICA are based on a *linear mixture* model, i.e. the observed data are interpreted as a linear mixture of some spatio-temporal patterns that can be retrieved by means of some criteria. It is known from Chapter 2 that PCA is based on the maximization of the explained variance within an orthogonal basis, while ICA accounts for independence. For what concerns the interpretation of the principal components of an fMRI dataset, it has been pointed out in [144] that their interpretation may be troublesome. In fact, it may be possible that a given principal component represents a mixture of effects, and understanding spatio-temporal patterns that are constrained to be orthogonal may be challenging (unless the observed variability has a natural elliptic structure); the purpose of PCA is, in fact, to find components that explain the maximum variance, and this may not be the best strategy, especially in fMRI where there are multitudes of effects that may be grouped together.

ICA overcomes the limitations inherent to PCA decomposition, and since its first introduction in [119] it has become a widely tool for fMRI data investigation. In section 3.3.6 details about the extraction and the interpretation of independent components of fMRI data will be given.

3.3.6 ICA applied to fMRI data analysis

Before dealing with linear mixtures, the dimensions of an fMRI dataset must be changed, since the original dataset is 4D (three spatial dimensions and a temporal one), while ICA is meant for linear mixture of monodimensional signals. Therefore a whole 3D volume has to be converted to a row of the data matrix $\mathbf{X} \in \Re^{m \times n}$, where m is the number of time points (i.e. scans), while n is the number of volume elements (voxels) for a given scan. Each row of the data matrix represents a whole volume at a given scan, while each column of it contains the time-course of a given voxel.

The linear model considered for the generation of the data matrix $\mathbf{X}$ is, like in (2.32):

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \tag{3.29}$$

where $\mathbf{E}$ is spatially and temporally white noise [144]. Usually the noiseless ICA model is considered, (embedding the noise among sources), obtaining thus:

$$\mathbf{X} = \mathbf{AS} \tag{3.30}$$

The double indeterminacy in both $\mathbf{A}$ and $\mathbf{S}$ gives two possibilities for the formulation of the model. In fact, one may look for a decomposition where the rows of $\mathbf{S}$ are independent, or to another one where the columns of $\mathbf{S}$ are independent. In other words, it is possible to look to data matrix $\mathbf{X} \in \Re^{m \times n}$ in two ways, defining two different techniques:

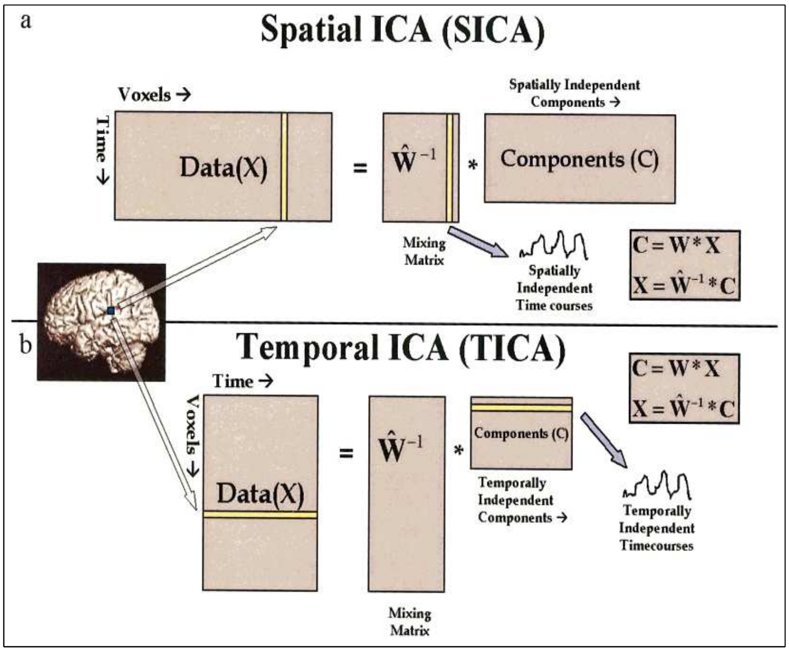


Figure 3.6: Spatial and Temporal ICA of fMRI data. (Figure from [34])

- **Spatial ICA: X** is a made of n realizations of a random variable $\mathbf{x} \in \Re^m$. Looking for independence means therefore looking for some linear combinations of the observed spatial maps that are maximally independent. The associated mixing coefficients **A** can be regarded as "time courses" of the independent maps, and are *unconstrained* to be orthogonal.

- **Temporal ICA: X** is made of m realizations of a random variable $\mathbf{x} \in \Re^n$. ICA decomposition leads to independent time courses with associated *unconstrained* spatial maps.

The two approaches are depicted in Fig. 3.6. It is easy to see that spatial ICA is more efficient from a computation point of view, since there is a relatively limited amount of time points (therefore of independent components) if compared with temporal ICA, where the number of independent components is equal to the number of voxels, that may be too computationally demanding.

The first application of ICA to fMRI data exploration is from McKeown in [119, 118], where Spatial ICA based on INFOMAX algorithm (section 2.7.1) has been applied to different experiments (Stroop color-naming, Brown and Peterson word/number task) to extract independent spatial maps whose interpretation led to the individua-

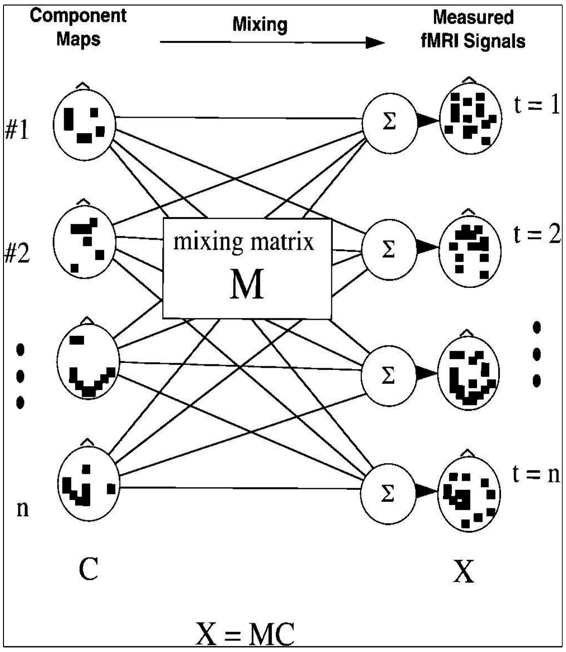


Figure 3.7: Spatial ICA or fMRI data: generative model (Figure from [119]).

tion of task-related maps. The spatial ICA mixture model is depicted in Figure 3.7: each of the n independent maps (on the left) contributes with different weights to the generation of the fMRI measurements (on the right). The weights of this mixing process can be seen as temporal time courses of the *whole* map, and all the maps with their associated time courses sum up linearly to give the measured signal.

The underlying assumptions to the use of ICA in fMRI data analysis have been pointed out in [118]. Basically, for the model to be valid, some requirements have to be made on the mixing model:

1. Maps associated with "independent" activity in the brain are *sparse* and mostly *non overlapping*, although some overlap may occur.
2. The mixing coefficients are *constant* throughout the brain.
3. The components *mix linearly* to form the fMRI measurement.
4. The number of components contained in data are up to the number of time points in the experiment.

These assumptions were examined for a Stroop color-naming task, and the original dataset was decomposed into independent components that were interpreted and identified in the following classes

- **Task-related activation map**: the spatial pattern of a spatial map was physiologically compatible with the results of a GLM and its temporal time course was consistently related to the task of the experiment.
- **Transiently task-related maps**: the time course of these independent components was "locked" to the stimulus only for a part of the stimulus, and therefore they could not be detected with a GLM, that averages over repetitions of the task.
- **Movement related components**: movement of the head of the subject causes the presence of one or more components whose spatial map has a ring-like activation area and whose time course suggests slow or abrupt movements of the head
- **Noise components**: those components whose time course or spatial map had no interpretation and could not be reproduced among repetitions of the ICA extraction on the same dataset.

Moreover in this study, other spatial maps that did not belong to any of the mentioned classes were detected.
In [119] the authors tried to order the components according to the variance contribution in the whole dataset (it has to be noted that both IC and mixing coefficient have no intrinsic value due to the double indetermination, but their product is *univocally* determined), but there was no easy way of detecting the task-related map from the amount of variance it explained. More sophisticated criteria have been therefore used: in [126] a frequency content ordering for a periodic task was proposed, ranking task-related components high. Information theoretic criteria like kurtosis (and therefore sparseness) have demonstrated to be useful but not optimal in terms of components ordering, since components corresponding to local flow were also ranked high with this measure. In [56] more elaborate and more realistic assumptions are invoked for identifying components with autocorrelation or spatial clustering. In [57] an automatic classification algorithm for independent components of fMRI time-series was proposed: by means of a multidimensional feature space, based on spatial, temporal and spectral properties of the estimated sources, the authors have shown how it is possible to discriminate between independent components related to different activities and they have employed a Support Vector

Machines (SVM) based classifier to automatically classify those sources.

From the work of McKeown, more and more studies involving the extraction of independent components of functional images of the brain have been presented, enhancing the capabilities of exploratory approaches (that are *complementary* to the confirmatory ones, [60]). In [147] a comparison of classical hypothesis-driven methods and ICA was made on clinical functional MR images, showing that both techniques are able to identify spatio-temporal patterns of activity, with ICA being more robust in case of data sets corrupted by motion or by incorrect task performance. In [99] it was made a comparison of an ICA extraction (based on INFOMAX algorithm) was compared to several other data analysis methods on simulated and real fMRI data, proving that ICA was able to identify locations of activation not accessible by simple correlation, t-test or general linear model based methods.

More complex experiments, like simulated driving [36] or visual ambiguous stimulations [41] were investigated successfully with ICA, proving that blind separation techniques are also able to deal with experiments where more than one task is performed and where independent sources may partially overlap. Moreover, ICA has been employed effectively in fMRI preprocessing: in [104] a motion correction algorithm based on independent component analysis and entropy criterion was proposed and it proved effective in increasing the informative content of the data in the presence of motion, without requiring the registration of the motion-corrupted volumes to a single reference volume, as this procedure can introduce artifacts because it does not account for variability due to the task-related components. Several ICA algorithms have been employed to extract independent maps; the most used are FastICA (section 2.8) and INFOMAX (section 2.7.1). It has been shown in [53] that whereas both algorithms produce accurate results, FastICA performs better than INFOMAX in terms of spatial and temporal accuracy, while INFOMAX is superior in terms of global estimation of the ICA model and noise reduction capabilities.

Both temporal and spatial ICA have been investigated [34, 29]. As said before, the main problem of temporal ICA is that an extremely large number of components is the result of the extraction, while the number of temporal samples is relatively small compared with the spatial one. This leads to temporal components characterized by spikes, that are almost zero everywhere else. In order to avoid such *overlearning* dimension reduction techniques are employed.

In [116] a hybrid approach was proposed to mix the hypothesis- and data-driven analysis. Time courses of the independent spatial maps extracted with ICA are

used as reference function for a general linear model, with the possibility of giving statistical significance to independent components. In that work, also the issue of how many components to keep in the analysis is addressed. An hybrid approach was also employed in [9], using ICA to remove the cofounds of task-related activation in exploring functional connectivity.
ICA, unlike univariate methods (GLM), does not naturally generalize to a method for drawing inferences about a group of subjects. Neverthless a method for making such group analysis has been developed [33]. This kind of group analysis is based on the main assumption that the data collected from different subjects are statistically independent observations. Each subject is then treated as an observation of the statistics of the population. The first stage of the processing sequence is standard preprocessing and spatial normalization of the data into a standard space (which can be Tailarach space or any other space that is chosen as common). The next step is data reduction, two reduction steps are used to reduce the computational load of simply entering all subjects' data into an ICA analysis. One is taken on the single subjects' data and the second one is done on the aggregate data set. Then the ICA is computed and group maps are formed. The most important step is the data reduction one, in fact an optimal number of components that should be estimated is computed. If the independence between different subjects' data holds, single subject maps can be reconstructed after the group unmixing matrix is obtained. Then there are two possibilities in order to obtain group maps: the first is to compute z-maps over the group ICA maps obtained via the analysis; the second is to compute the single subject maps and then compute the mean and the variance of each component across subjects, where the variance across subjects can be used as an estimate of the population variance. After this hypothesis test can be done to provide a "random effects" inference [33, 158, 110]. A more sophisticated technique based on self-organizing clustering has been recently proposed in [54].

Several improvements to classical ICA tailored to fMRI have been proposed. In [156] a spatio-temporal independent component analysis is carried out by means of a modified cost function, while in [157] a skewed contrast function, that can account for asymmetry in independent components' pdf, has been used to achieve more precise results. However a procedure to extract spatial maps that enforcing spatio-temporal regularities, that a physiologically plausible map must exhibit, has not been addressed in literature.

3.4 MagnetoEncephaloGraphic Signals: General Concepts

MagnetoEncephalography (MEG) is a non invasive technique whose aim is to investigate neuronal currents by means of induced magnetic fields observation. It can be considered, together with ElectroEncephaloGraphy (EEG), a technique complementary to functional Magnetic Resonance Imaging (fMRI). In fact, while the latter has a high spatial resolution and a relatively low temporal resolution (related to the BOLD effect), the former can reach a temporal resolution of the order of magnitude of *msec*, but with somehow troublesome spatial localization.

The first indirect observation of neuronal electrical activity by means of non-invasive measurement on the scalp has been made in 1929 by Berger [27]; in his work, Berger analyzed potential differences among the scalp by means of electrodes, and was able to identify and classify the two main rhythms of the brain (alpha and beta activities). From that work, non-invasive exploration by means of induced field observation became a widely used investigation technique. Although any electrical current also induces a magnetic field, it has not been possible for a long time to explore it since its magnitude is much smaller than environmental magnetic field (induced magnetic field is around five orders of magnitudes smaller than urban magnetic noise, and nine orders smaller than earth magnetic field). The introduction of Superconducting Quantum Interference Device (SQUID - for details, see [68]) made it possible to explore also magnetic fields too. In 1968, Cohen observed the alpha rhythm with a SQUID-based measurement device [44]. Since that time more and more complex architectures for detecting induced magnetic fields on the brain have been developed, and now helmet sensors, that employ up to 150 channels and can record the activity on the whole brain, are being used (in a magnetic shielded environment).

The measured magnetic activity is the result of electrical currents in the neurons within the cortex: each sensor stimulus or neural pulse moves through the nerve system by means of a depolarization wave, generating an *action potential.* If the head is assumed to be a spherical conductor, then it is evident that the only neuronal currents that will produce an observable magnetic field outside the brain are the ones that are parallel to the surface. Moreover, activity of a single neuron does not cause a strong enough signal to be measured. Therefore, signals obtained will be related to pools of neurons acting synchronously, or rather departing from the

overall synchronous activity and so reducing its effect.
To model the measured field distribution Maxwell's equations

$$\begin{aligned}
\nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon} \\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\
\nabla \cdot \mathbf{B} &= 0 \\
\nabla \times \mathbf{B} &= \mu(\mathbf{J} + \varepsilon \frac{\partial \mathbf{E}}{\partial t})
\end{aligned} \tag{3.31}$$

and current continuity equation

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} \tag{3.32}$$

are employed.
However some assumptions are usually made. In fact, since frequencies involved are below 100 Hz, the quasi-static approximation can be made; moreover magnetic permeability μ in biological tissues can be assumed to be equal to the one of empty space (i.e. $\mu = \mu_0$). The quasi-static assumption is fulfilled if the time-varying (i.e. time derivatives) are smaller with respect to ohmic currents, and for the typical values of MEG applications ($\rho = 0.3\Omega^{-1}m^{-1}$, $\varepsilon = 10^5\varepsilon_0$ and $f = 100 \quad Hz$), this assumption holds. This does not mean that time-varying phenomena are neglected, but $\frac{\partial \mathbf{B}}{\partial t}$ is not considered in $\nabla \times \mathbf{E}$ evaluation, and so is $\frac{\partial \mathbf{E}}{\partial t}$ for $\nabla \times \mathbf{B}$ evaluation.
The electric field is such that, in the quasi-static approximation:

$$\nabla \times \mathbf{E} = 0 \tag{3.33}$$

For what concerns current density $\mathbf{J}(\mathbf{r})$ produced by neuronal activity, it can be modeled as the sum of two currents

$$\mathbf{J} = \mathbf{J}^p(\mathbf{r}) + \mathbf{J}^V(\mathbf{r}) \tag{3.34}$$

where:

- $\mathbf{J}^p(\mathbf{r})$ is the *primary* current, caused by the ions movement within the membrane, therefore it is connected to chemical activity of the cells.
- $\mathbf{J}^V(\mathbf{r})$ is the *volume* current generated by passive movement of free charges under the influence of primary current.

MEG analysis aims at individuating primary currents, but induced volume currents have to be accounted for in field computations.
Several techniques have been developed to find the primary currents in the brain related to an experiment, but no unique solution can be provided. In section 8.2.1 the localization procedure that has been employed will be described.

3.4.1 ICA analysis

In the last decade ICA has been employed successfully to extract consistent information from MEG recordings [167]. Several studies have proved its effectiveness in removing artifacts and extracting relevant activations from MEG and EEG signals [112, 168, 82, 17, 86].
A first challenging issue in Blind Source Separation neurophysiologic applications is the choice of the contrast function used to extract sources: the non-Gaussianity assumption in the ICA model and the imposition of an orthogonality constraint between extracted components (ICs) produce source estimates which are active during short time intervals with minimal overlap. Therefore, ICA seems to be effective for separating neuronal signals corresponding to sources that exhibit burst behavior, coming from spatially distinct compact sources. The magnetic field patterns of these ICs are close to those produced by isolated current dipoles [113, 125]. In this way, ICA achieves both temporal and spatial separation of source activity and can significantly enhance imaging accuracy [181, 125]. On the other hand, ICA is insensitive to the time ordering of the data points; instead other BSS algorithms have been recently claimed to be more suitable for cerebral sources separation, by exploiting second-order statistics of the source signals to decompose the recorded mixture, as for example minimizing a set of time-lagged cross-correlations [162]. At present, many different BSS packages are available, implementing both high-order ICA algorithms and second-order BSS techniques; validation of obtained results has to be investigated case by case. A second key point in applications is how to assign the neurophysiological and neuroanatomical meaning and interpretation to the extracted sources, since often “interesting” characteristics are not effectively separated in a single component but they can remain partially mixed, or split into more than one component. Usually, a post-extraction analysis of spectral and spatial IC properties is applied to select the relevant ones, leading to the definition of clusters of “similar” components with respect to some criteria [113, 73, 18]. The necessity of

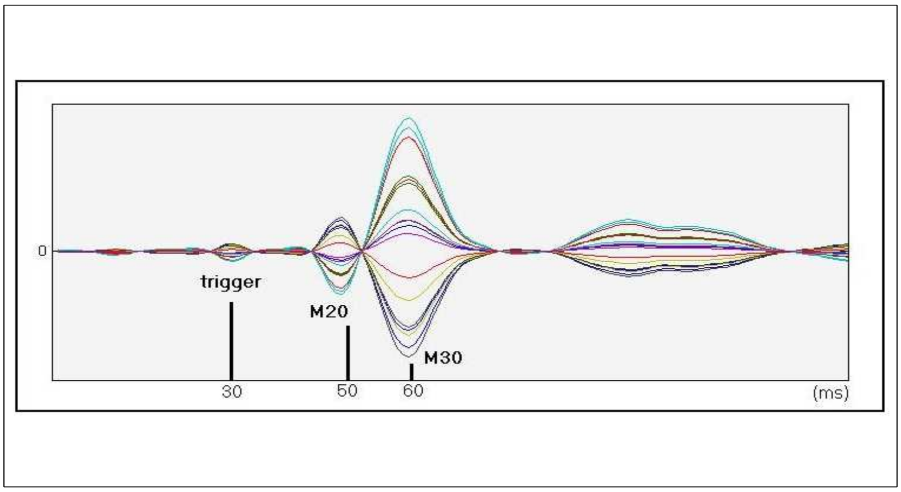


Figure 3.8: MEG channels signals reconstructed by means of a *single* independent component

this post-processing is the consequence of the blindness of the approach, since ICA does not take other information into account than the statistics of the data; the advantage is the generality of the assumptions, that make these techniques powerful and flexible tools with respect to hypothesis-driven procedures, which are highly dependent on the accuracy of a predefined model/template.
The ICA model for MEG data analysis is quite straightforward: each source of neural activity is measured by different sensors with different weights, and all sources are mixed up linearly, together with environmental noise. It is usual to consider noise as an additional component, therefore the model can be described as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{3.35}$$

where $\mathbf{x}$ are the measured magnetic fields, $\mathbf{s}$ are as usual the independent components and $\mathbf{A}$, the mixing matrix, describes the mixing process. Although ICA does not solve the inverse problem explicitly, it is to be noted that information contained in the mixing process can be used to localize the sources. In fact localization for independent sources is performed by spatial retro-projection, i.e. in order to localize component $\mathbf{s}_i$, its contribution to reconstructed data is found as follows:

$$X_{rp_X} = \mathbf{a}_i \times \mathbf{s}_i \tag{3.36}$$

where $\mathbf{a}_i$ is the estimated mixing vector (i.e. the ith column of $\mathbf{A}$) for the source

$\mathbf{s}_i$, and X_{rp_X} is the retro-projection on the sensor channels of the estimated source. Therefore each reconstructed channel has the same waveform but a different amplitude (see Figure 3.8, where an independent component contribution to signal reconstruction is depicted for the different channels) related to mixing coefficients estimates. As a consequence of this, the field distribution obtained by retro-projecting only one component is time-invariant up to a scale factor; consequently, the subtending current distribution is time-independent.

Chapter 4

Dedicated Architectures for fast ICA Implementation

4.1 Introduction

The main principles and application fields of Independent Component Analysis have been showed in Chapter 2. Since this technique is quite general and can be used for different applications, there is an increased need for performance in both speed and accuracy in independent components extraction. In fact, for what concerns biomedical signal processing, accuracy of the separation, robustness to noise and to outliers is a crucial aspect, while for other problems, like telecommunication signal processing, the use of ICA algorithm is mainly focused on speed of extraction.
Another interesting aspect of the problem is the possibility of *embedding* the signal extraction into a processing chain or a single device, so that this device may be portable. The best solution to this task in terms of performances, power consumption and space occupation would be to design a custom integrated circuit for the application. However, this strategy could not be optimal in terms of design time and cost, since the production of dedicated ASIC usually requires a high number of produced pieces to justify the efforts to produce it.
The use of dedicated programmable architectures, like Digital Signal Processor (**DSP**) and Field Programmable Gate Array (**FPGA**) units is in between the two strategies of employing a custom ASIC and of employing a general purpose microprocessor. DSP architecture, in particular, is especially suited for scientific calculations that involve high numbers of “simple”and repetitive tasks with the re-

quirement of real-time processing.

In this chapter some aspects of performances of ICA algorithms have been investigated, and the study of the feasibility of an embedded ICA separating algorithm has been explored on a Digital Signal Processor (**DSP**) architecture.

The ICA algorithm employed is FastICA ([79],[80]), one of the most effective separation algorithms both for its accuracy and for its speed. As seen in Section 2.8, FastICA is based on the maximization of an approximation of negentropy by means of a fixed-point iteration; its speed comes from the approximation of the Hessian matrix of the associated Lagrange optimization problem, with an identity matrix multiplied by a scalar, making the extraction fast and precise. The optimization can be carried out in two alternative ways: extracting one component at a time (deflation approach) or extracting all the sources together (symmetric approach). For an implementation on a dedicated architecture the first seems to be the most suitable. In fact, consider the extraction of n independent components with T samples each. A symmetric approach would handle $n \cdot T \cdot B$ bits of data, where B is the number of bits assigned for the representation of a value, at each iteration, while in the deflation scheme data dimension reduces to $T \cdot B$ (for each component). On the other hand, however, the deflation approach can encounter some problems in the last extracted components, since errors do propagate from one component to all the others.

The core of the FastICA algorithm is the update of the unmixing coefficient. Recalling eq. 2.148, at each iteration the algorithm must perform the following steps for the p-th component:

$$\begin{aligned} \mathbf{w}_p &\leftarrow \mathbf{w}_p \mathrm{E}\{g'(\mathbf{w}_p^T \mathbf{z})\} - \mathrm{E}\{\mathbf{z} g(\mathbf{w}_p^T \mathbf{z})\} \\ \mathbf{w}_p &\leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1} (\mathbf{w}_p^T \mathbf{w}_j)\mathbf{w}_j \\ \mathbf{w}_p &= \frac{\mathbf{w}_p}{\|\mathbf{w}_p\|} \end{aligned} \tag{4.1}$$

where the most time consuming operations in the computation are the evaluation of non-linearities and the expected values. Moreover, it will be showed in the following that also decorrelation and normalization steps may be troublesome for a fixed-point architecture and they have to be implemented carefully to avoid overflows or precision loss.

To test the performances of a DSP implementation of FastICA algorithm, we decide

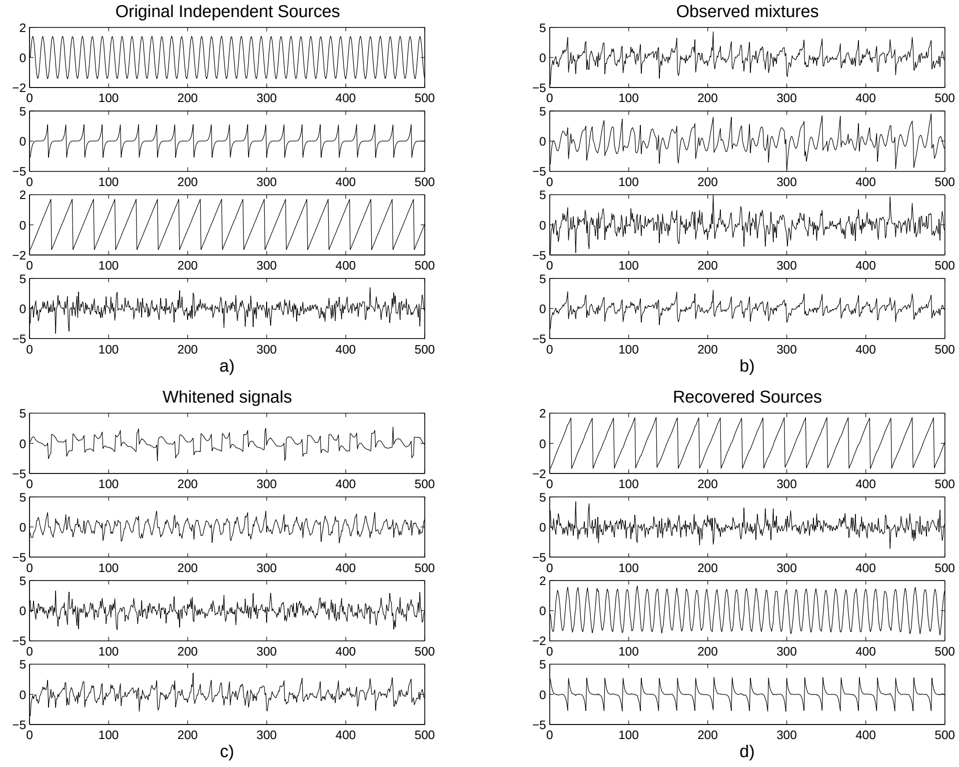


Figure 4.1: Test data set for independent component separation. (a) Original sources, (b) observed signals (linear mixture of (a)), (c) whitened observed signals, (d) recovered signals with FastICA

to extract some “test”independent sources mixed by means of a random matrix. Original independent signals **s** were generated with FastICA package (CITARE), and mixed linearly by means of a matrix **A** whose elements are i.i.d. uniformly in $[-0.5, 0.5]$. The observed signals, together with the real and estimated independent components and whitened data are depicted in Figure 4.1.

We started our analysis by whitened data rather than from original one, to evaluate the time employed only by the independent component extraction. To evaluate the results some proper measures of effectiveness of the separation were employed, and the algorithm was tested on a PC, on a floating point and on a fixed point DSP architecture. In Section 4.2 the details of the implementation and the results obtained will be discussed in detail.

4.2 DSP Implementation of FastICA

If compared with general purpose microprocessors, Digital Signal Processors (DSP) are optimized for scientific computation, and are often employed as a stand-alone processing unit or as co-processors associated to a microprocessor. To obtain real-time performances, together with low power consumption and with the possibility of being embedded in some environments, their architecture is particularly optimized with respect to memory usage and transfer. Moreover, a high level of parallelization can be achieved, together with the possibility of accessing to more than one memory location within one cycle.

In the classical Von Neumann architecture the Arithmetic Logic Unit (ALU) and the control unit are connected to a single memory that stores both the data values and the program instructions. During execution, an instruction is read from the memory and decoded, appropriate operands are fetched from the memory, and, finally, the instruction is executed. The main disadvantage is that memory bandwidth becomes the bottleneck in such an architecture. To overcome this limitation DSPs implement a Harvard architecture, where instructions and data are stored in separate memories and can therefore be trasferred in parallel. The difference between Harvard and Von Neumann architectures is described in Figure 4.2.

The most common operation a standard DSP processor must be able to perform efficiently is multiply-and-accumulate. This operation should ideally be performed in a single instruction cycle. This means that two values must be read from memory (one of them might reside in a register) and (depending on organization) one

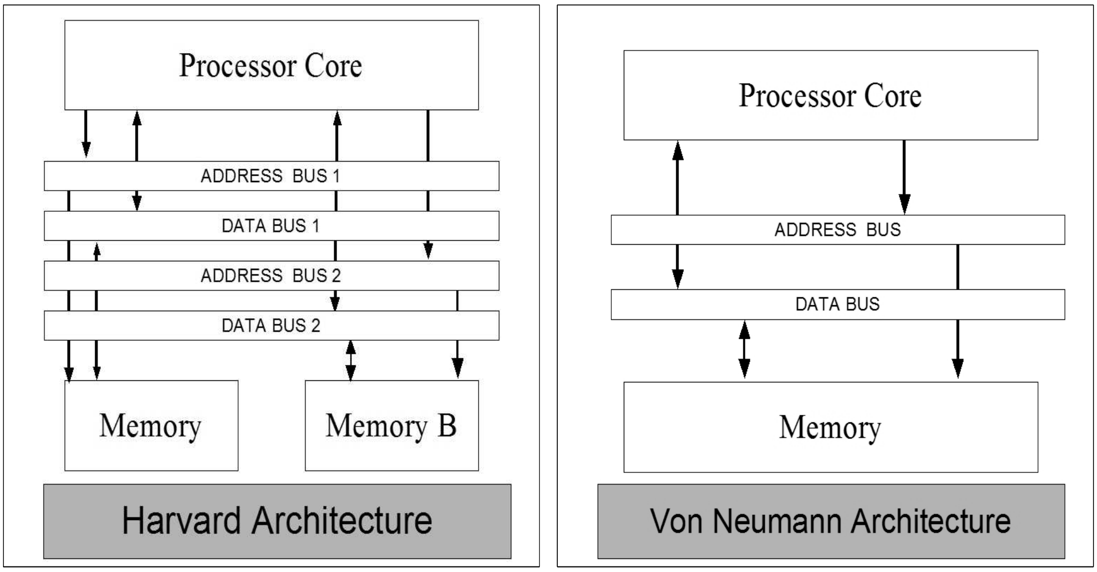


Figure 4.2: Harvard (left panel) and Von Neumann (right panel) architectures

value must be written, or two or more address registers must be updated, in that cycle. Hence, a high memory bandwidth is just as important as a fast multiply-and-accumulate operation. Several memory buses and on-chip memories are therefore used so that reads and writes to different memory units can take place concurrently. Furthermore, pipelining is used extensively to increase the throughput. Current DSP architectures use multiple buses and execution units to achieve even higher degrees of concurrency. Chips with multiple DSPs processors and a RISC microprocessor are also available.

One of the major strengths of DSP computation is the possibility of performing in one cycle the multiply-and-accumulate operation (Multiply And Accumulate, MAC), that is particularly desirable while dealing with data processing in real-time. For further details on DSP architecture and functionalities, see [111], [95] and [171] FastICA algorithm implementation was explored on two different architecture, floating and fixed point, on two DSPs from Texas Instruments^{TM} of the family C6000. Floating point architectures are more sophisticated than fixed point ones, and their choice should be motivated by reasons of portability and embedding ability, since they can be easily outperformed by general purpose microprocessor in terms of speed. On the contrary, the major strength of fixed point architecture is their relatively higher speed, since higher clock frequencies (if compared with floating point ones) can be achieved.

DSPs algorithms can be easily and reliably implemented on Texas Instruments^{TM} devices by means of Code Composer Studio (CCS), a software that allows C code implementation on DSP platforms. Since Assembler is the native language in DSP architecture, it is still possible to implement the algorithm in that language, but this could not be the best solution in terms of developing time. In fact, CCS gives the programmer a fully accessible environment allowing to compile, build and run a C code on DSP, with the possibility of configuring memory allocation, interrupts and libraries. Moreover, the efficiency of a C code implementation can be around $80-100\%$ of an assembler one, making C programming the best trade-off in terms of time effort and performances. A C code was implemented therefore for FastICA, and it was focused only on the independent components extraction. As said in 4.1, some aspects of the algorithm implementation must be studied carefully to achieve speed performances. In particular, the most expensive computation is the contrast function evaluation, since an average on all the dataset has to be made for each iteration. Another computationally expensive operation is decorrelation. For what concerns fixed point implementation, another issue was the normalization step, that had to be modified in an adaptive fashion to avoid overflow in the elaboration.

4.2.1 Texas Instruments C6000 Family

Before explaining in detail the floating and fixed point implementation of FastICA algorithm, a brief description of C6000 family will be given. Among all of the DSP families of Texas Instruments, we choose C6000 class since it is the most suitable for complex calculus, even if it is the most expensive. There are some features common to all the DSPs in C6000 families, mainly regarding the processing unit scheme (Figure 4.3). As seen in Figure 4.3, right panel, the C6000 architecture has 8 independent functional units, and four of them can be employed to perform two simultaneous MAC operation, allowing therefore a high level of parallelization that can be exploited even at C level by means of CCS compiler; due to parallelization C6000 family can perform in general from 200 to 2400 MMAC (Mega Multiply-and-Accumulate) operations per second. Further real-time oriented design strategies have been implemented on the internal bus, as seen in Figure 4.4. Five simultaneous bus operations can be performed at a time:

- one program read
- two data read/write

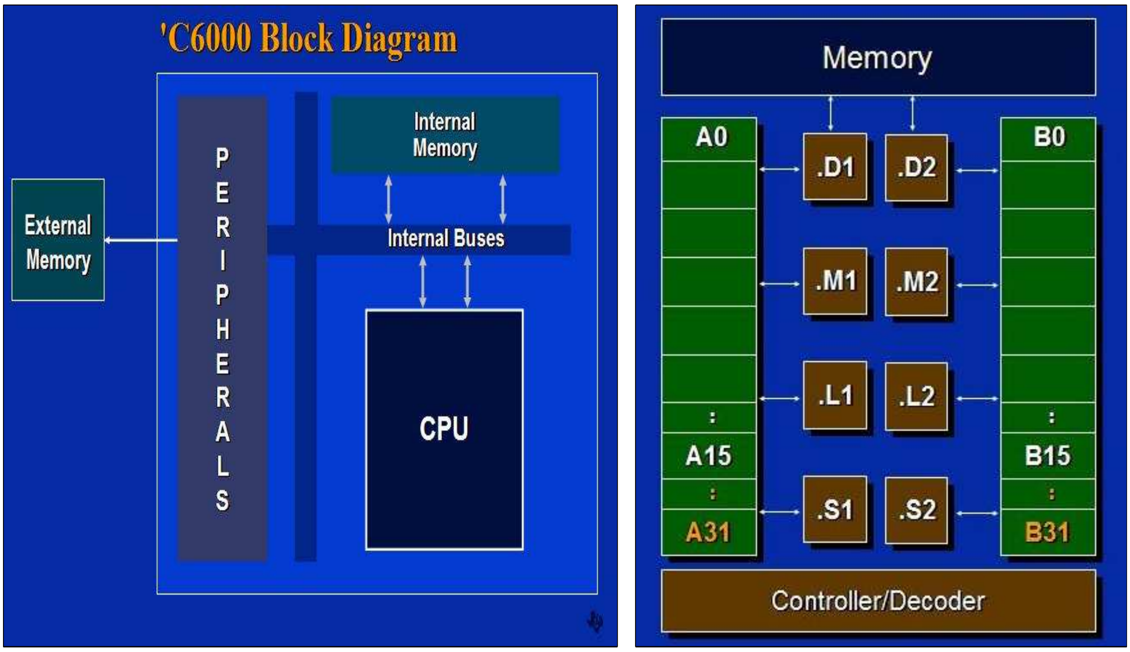


Figure 4.3: C6000 Family Block Diagram (left) and Architecture (right)

- one DMA write
- one DMA read

Direct Memory Access (DMA) allows to optimize memory usage, since it can move data from and to memory, without use of CPU resources, by means of proper interrupts.

The DSP has on-chip two-level cache. The first cache level is split into program and data memory, consistently with the Harvard architecture implemented, while second-level cache and main memory are shared between program and data, allowing the flexibility in usage typical of von Neumann architectures. Each DSP used was embedded in a Developer Starting Kit (DSK), which provided the DSP with interfaces, additional memory banks, A/D converters and ports for expansions. Usually a DSK is connected to a PC, and the communication can hold via JTAG emulator (by means of a parallel or USB port) or via Host Port Interface (HPI).

4.2.2 Floating Point Implementation

Floating Point realization of FastICA algorithm was implemented on a TMS320 C6711 DSP with a clock frequency of 150 MHz. This DSP was provided with a

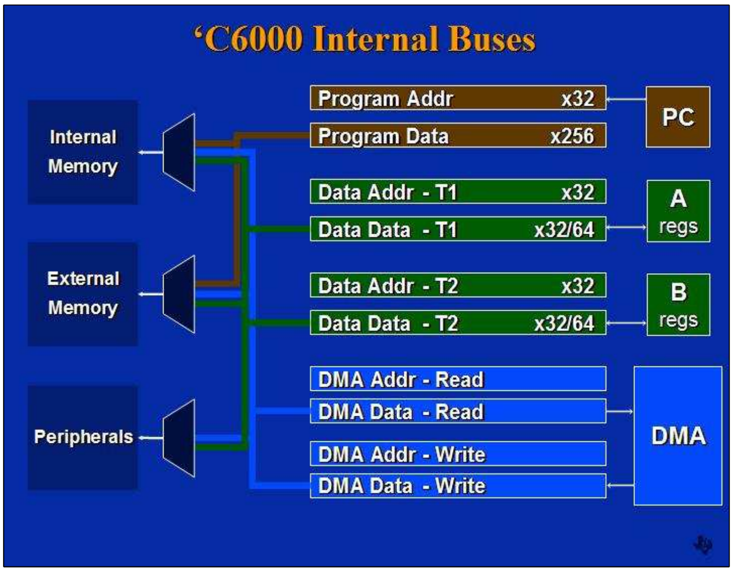
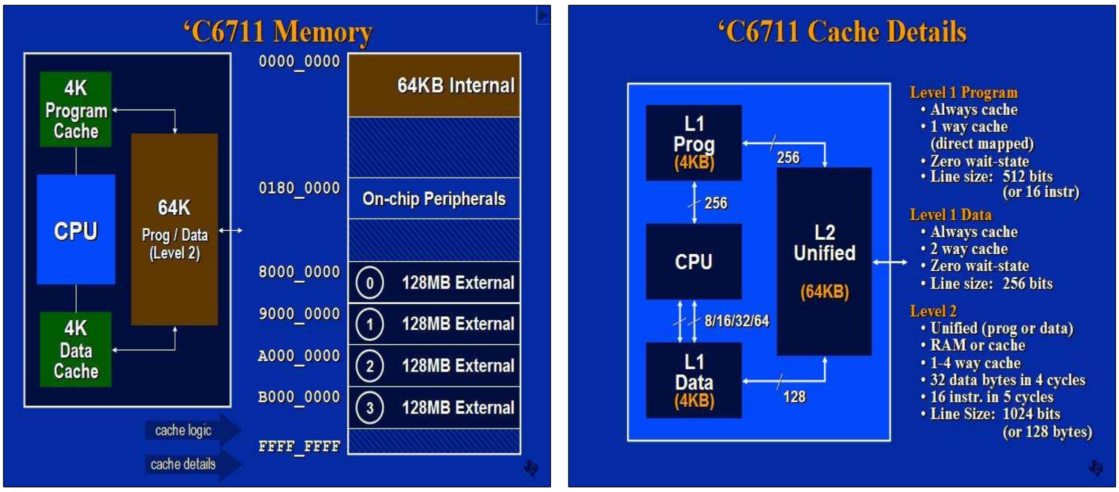


Figure 4.4: Texas Instruments C6000 internal buses

DSK whose interface is the parallel port. The DSP used has a floating point unit, together with other units as shown in Figure 4.3, that allows a more precise number representation, if compared with a fixed point one. Moreover, C6711 is provided with additional features to speed-up the execution: memory mapping, depicted in figure 4.5, has three different levels:

- First level cache (L1), separated for data and programs (4KB are provided for each of them). Program cache can store up to 1000 instructions (each line of 512 bytes can store 16 instructions)
- Second level cache (L2), unified for program and data, with a dimension of 64KB. This memory module is divided in four blocks can be configured independently as cache or RAM memory .
- External memory, that can account for additional memory modules.

The implementation of a FastICA algorithm for a floating-point architecture is straightforward, therefore a preliminary version of the algorithm based on C code has been realized and tested. By means of debug tools provided with Code Composer Studio it is possible to quantify the amount of cycles needed for each part of the code, therefore the average time elapsed. The next step has been to optimize the


'C6711 Memory

'C6711 Cache Details

Level 1 Program
- Always cache
- 1 way cache (direct mapped)
- Zero wait-state
- Line size: 512 bits (or 16 instr)

Level 1 Data
- Always cache
- 2 way cache
- Zero wait-state
- Line size: 256 bits

Level 2
- Unified (prog or data)
- RAM or cache
- 1-4 way cache
- 32 data bytes in 4 cycles
- 16 instr. in 5 cycles
- Line Size: 1024 bits (or 128 bytes)

Figure 4.5: TMS320 C6000 memory (left) and cache (right) schemes

code structure and applying compilation optimization to the debugged code. Compiler optimization is based on massive parallelization of instructions that do not have dependencies (two instructions are independent if one is not based on the output of the other), but usually "manual"code changes that are oriented to parallelization enhance compiler ability to recognize independent instructions dramatically.
To help pipelining of the instructions, some further modifications can be made to the code. In particular, by means of compiler directive MUST_ITERATE it is possible to "force"the program to run a cycle for at least a fixed number of iterations, having a more efficient pipelining.
Further compiler directives to speed up execution are DATA_ALIGN (for global vectors) and DATA_MEM_BANC (for vectors defined locally), that force data vectors to be aligned in 8 bytes groups in memory, so that cache transfers are optimized.
With these modifications, performances were evaluated again, and results will be discussed in section 4.2.4.

4.2.3 Fixed Point Implementation

Fixed point version of FastICA algorithm was implemented on a TMS320 C6416 DSP from Texas Instruments. The DSK board consisted basically of:

- a DSP C6416 with fixed-point architecture, 600 MHz, 1 MB RAM
- 16 MB external SDRAM memory and a 512 KB flash memory

- an A/D converter
- a JTAG emulator

The architecture of the DSK is depicted in Figure 4.6.
Memory map is similar to the one of C6711 depicted in Fig. 4.5, but it has greater storing capabilities:

- The first level is still separate for data and instructions, and stores 16KB for each of them. Each line of program cache contains 8 instructions.
- Second level cache is common for programs and data, and stores 1 MB. The size of the line is 64 bytes.

The major issues of the fixed point representation if FastICA are precision and overflow. In fact

- Data must be represented with the available bits, therefore particular care must be taken in assigning bits to the significant digits of the numbers.
- The precision chosen for data may affect the convergence properties of the algorithm
- Particular attention must be paid to internal computations of the algorithm, that may cause overflow.

Denote now with WL the number of bits available for an integer representation and let the number of integer part bits be I_{WL} and the number of fractional part bits be F_{WL}: the number of bits needed for a value representation is $I_{WL}+F_{WL}+1$, where the last term accounts for the sign. The range R of numbers representable with I_{WL} in fixed point notation is $-2^{I_{WL}} \leq R \leq 2^{I_{WL}}$. Of course, relative precision of numbers is in any case 2^{-WL}, while absolute precision is $2^{-F_{WL}}$.
Taking a larger value F_{WL} allows a more precise representation of the number, but on the other hand it reduces the bits dedicated to the integer part, reducing the range of admissible values: a trade-off between precision and range, taking into account convergence properties of the algorithm, must be found.
The major issues in fixed point conversion iteration are:

1. Evaluation of nonlinearity: the chosen nonlinearity was $G(y) = 1/4y^4$, therefore in (4.1) $g = y^3$ and $g' = 3y^2$. Before taking the expected values of

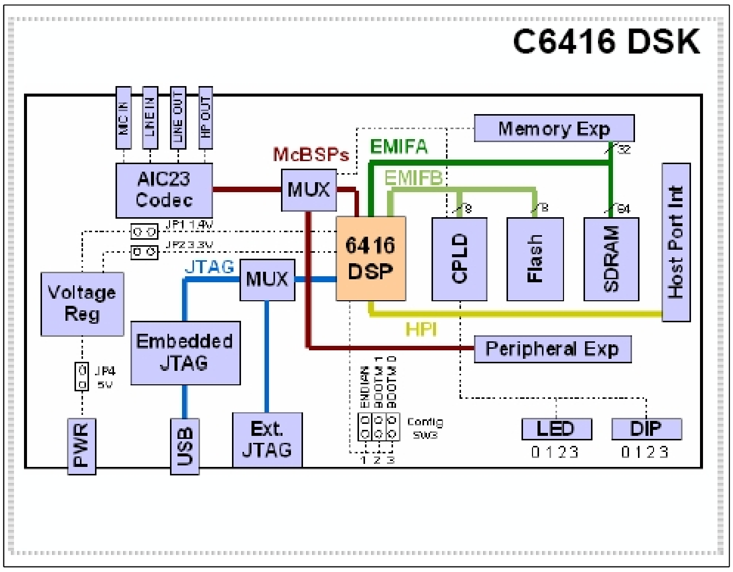


Figure 4.6: TMS320 C6416 Scheme

the nonlinearities on the whole dataset, values may go out of range, causing overflow.

2. Products: any product in the algorithm may cause overflow

3. Normalization: the evaluation of the norm may be troublesome in terms of overflow.

The following solutions were taken to avoid overflow:

1. Consider the computation

$$\mathrm{E}\{g'(\mathbf{w}_p^T\mathbf{z})\} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{w}_p^T\mathbf{z})\cdot(\mathbf{w}_p^T\mathbf{z}) : \tag{4.2}$$

incorporating the term $\frac{1}{m}$ into the summation operation in the following way

$$\sum_{i=1}^{m}\frac{1}{\sqrt{m}}(\mathbf{w}_p^T\mathbf{z})\cdot\frac{1}{\sqrt{m}}(\mathbf{w}_p^T\mathbf{z}) \tag{4.3}$$

reduces the range of values needed to perform the computation. The same procedure can be done for the term $\mathrm{E}\{\mathbf{z}g(\mathbf{w}_p^T\mathbf{z})\}$.

2. According to the bits allocated for the integer part, the maximum value R_{MAX} for a number is evaluated. Even if the representation chosen is suitable for input and output, it may happen that intermediate values of some variables overflow. If the product of two terms a and b is grater than R_{MAX}, the values of a and b are scaled dividing by powers of 2 until their product becomes lower that R_{MAX}. The entity of the truncation is stored and then the result is scaled back to the result once it is possible without causing overflow.

3. The norm evaluation is the most crucial part in the algorithm. In fact, at each iteration unmixing coefficients vector $\mathbf{w}$ must be divided by its norm

$$\|\mathbf{w}\| = \sqrt{\sum_{i=1}^{n} w_i^2} \tag{4.4}$$

   and errors in norm evaluation *do propagate* from an iteration to the other. We decided to adopt an adaptive truncation, choosing the minimum truncation that avoided overflow, avoiding thus to lose too much precision in the norm evaluation.

The maximum dimension allowed for an integer representation in DSP C6416 is the *long* format, that occupies 40 bits, i.e. signed integers from -2^{39} $(= -549755813888)$ to 2^{39} (=549755813888) can be represented with such format.

In a preliminary analysis, ranges of all the variables of the algorithm have been evaluated by tracing on the example data described above. For each variable, its mean, maximum and minimum value together with the standard deviation have been estimated. The best trade-off between fractional part representation and overflow has been the following: 20 bits were allocated for the fractional part, 19 for the integer part and 1 for the sign. Fixed point representation of input data x_{FP} is obtained by multiplying the number by 2^{20} $(\approx 10^6)$ and discarding the fractional part of $x_{FP} \cdot 2^{20}$.

By means of this choices, the algorithm was able to run without overflow and to reach an estimate of the indepdent components. The results are presented in next section, together with those of the floating point implementation.

4.2.4 Discussion

Evaluation of performances for both implementations has been conducted both on accuracy and speed, and compared with a C code running on Personal computer with a AMD Sempron 1.8 GHz processor, with a a 512 KB cache and 512 MB RAM.

Evaluation of accuracy of the implementations

For what concerns accuracy, a typical performance index is the following:

$$E = \sum_{i=1}^{m}\left(\sum_{j=1}^{m}\frac{\mid p_{ij}\mid}{\max_k \mid p_{ik}\mid} - 1\right) + \sum_{j=1}^{m}\left(\sum_{i=1}^{m}\frac{\mid p_{ij}\mid}{\max_k \mid p_{kj}\mid} - 1\right) \tag{4.5}$$

where p_{ij} is the ijth element of matrix $\mathbf{P} = \mathbf{BA}$, with $\mathbf{B}$ being the estimate of the unmixing matrix $\mathbf{W}$. If the separation is "perfect", $\mathbf{P}$ reduces to a permutation of the identity matrix, and E is 0. Higher values of E_1 denote some errors in independent components extraction. The acceptable value of E depends on the number of independent sources and on the sample size for the observations.
A Matlab implementation of the algorithm was considered first, and E_1 was evaluated for it. Then the same index was evaluated for the floating point implementation and the fixed point one (for the fixed point implementation, data were converted to floating point and compared on a PC, therefore on a floating point). Mean values for the extractions are the following:

- **MATLAB:** $E = 0.95$
- **Floating Point:** $E = 0.95$
- **Fixed Point:** $E = 0.6$

The equivalence between MATLAB on a PC and C6711 in terms of performances is rather intuitive, since they are based on the same architecture and perform basically the same operations. On the contrary, it could seem quite surprising that a fixed point implementation, that makes truncations and approximations, may perform better than a floating point version. The explanation maybe is in the fact that truncation in the values of $\mathbf{w}$ may reduce the effect of non optimal representation of $\mathbf{P}$ as a permutation of identity matrix. Moreover it is to be noted that if the fixed point implementation reaches a solution, this solution is the same as in the floating point implementation, even if its representation is limited by the used precision.

Evaluation of time performances

Time execution performances were evaluated for the implementations on a PC with MATLAB Code and C code, and for the two implementations in floating and fixed point. Time evaluation procedures were different for the two DSPs, since C6711 was provided with debug options that enabled getting information on the cycles needed for *each* code selection, while for C6416 this was not possible. MATLAB and C codes speed were evaluated by averaging the execution time on several repetitions. The results of this evaluation are the following:

- **C6711 DSP:** the average of the amount of cycles needed for an extraction was considered. In particular, the code referring only to FastICA main iteration needed an average amount of 1.064.535 cycles. Since the clock frequency of C6711 is 150 MHz, the average time for the extraction is around 7.09 *msec*.

- **C6416 DSP:** in this case it was not possible to evaluate in debug mode the amount of cycles needed for each code selection. Therefore the internal timer of the DSP was employed, and the average amount of time needed for the extraction was around 23 μsec.

- **MATLAB:** the MATLAB version of the algorithm was modified such that the extraction procedure was repeated for 10^6 times, and the total elapsed time was evaluated. Each independent component extraction lasted 6,26 *msec* on average.

- **C CODE on PC:** the same procedure as the MATLAB case was implemented, and the average execution time was 10.95 μsec.

4.3 Conclusions and future work

Two implementations on DSP architecture have been explored on a test dataset, and a comparison in terms of speed and accuracy has been carried out for both the fixed point and the floating point implementations. For what concerns the speed, the fixed point solution has been proved as the most suitable, while the floating point one seems to be highly inefficient. The reasons for this behavior may be several: first of all, TMS C6711 is a *first generation* DSP, while TMS C6416 is a *second generation* one, and therefore is better designed both in terms of memory usage and in terms of architecture optimization. In fact the considerably larger amount of cache memory and the higher level of parallelization, may have contributed to a difference of elapsed time of roughly two orders of magnitude. Moreover, fixed point DSPs show a promising trend for what concerns clock frequencies: in fact, the TMS

C6416 employed in this work has a clock frequency of 600 MHz, but recently developed C64XX DSPs from Texas Instruments exceed 1 GHz frequency, this meaning that the performances explored on this test case can be further enhanced (almost doubled already on currently available HW) by means of more powerful architectures.

However, a fixed point implementation has not been yet *fully* explored. In fact, data precision may be a crucial issue while considering different and considerably larger datasets. In this case *whitened* data were provided directly to the DSP. In a complete independent component extraction procedure, this step should be performed directly by the DSP, and this may cause some problems. In fact whitening procedure (see section 2.6) transforms by means of a linear transformation the original dataset into a new one that has a unity covariance matrix: this means that data variance is reduced to 1, allowing a more precise representation in a fixed point architecture. Therefore a suitable procedure for whitening the data within the necessary precision should be implemented, to allow a complete processing from raw data to independent components. Moreover, some troubles may be encountered in moving to larger datasets (even if whitening procedure ensures that data variability will be somehow bounded). Range evaluation for variables was performed on a rather limited dataset. While whitening ensures a substantial limiting of range, more extensive evaluation should be performed on other datasets from different application environments.

To sum up, the present exploration has proved that a fixed point implementation of FastICA algorithm is feasible, and a trade-off between precision and overflow may be found (also in adaptive fashion, like in the norm evaluation case). Some important issues have still to be addressed, but the speed and the continuously increasing performances of fixed point DSPs make them a suitable choice for embedding the FastICA algorithm into a real-time processing chain. In any case, the study reported shows that we cannot expect a single DSP to outperform a standard PC significantly enough to justify the cost of designing such a solution. To further increase the performances in terms of speed, a multi-threading architecture is at present under study, since FastICA core iterations can be divided into concurrent tasks and parallelized. A distributed computing architecture, made by several DSP units embedded in a multiprocessor board with one DSP distributing computational load to the others is currently under study. This architecture is scalable and may significantly outperform a PC-based solution with contained cost, good portability and ease of upgrading.

Chapter 5

Statistical Non-Linear Model of HEMT MMIC's with BSS

5.1 Introduction

In the present work a rigorous validation procedure of a previously implemented model [42] has been performed. In particular, a population of 196 HEMT devices has been built by means of a Quasi-two-Dimensional (Q2D) simulation. Proper statistical variation of the physical parameters, and distance-dependent device-to-device correlation have been imposed. The generation of the HEMT population is described in detail in section 5.2. The statistical non-linear model has been extracted from DC I_{DS} current and S-parameters of the devices. Further details of the model and on the extraction procedure will be given in section 5.3. Finally, capability of the model to account for physical parameters variation, and to reproduce statistical distribution of the S-parameters has been checked by hypothesis testing, PCA and, for the first time, ICA (section 5.4).

5.2 Database Generation

The nominal model of conventional HEMT transistor has been generated by means of HELENA ([72]) software, that is designed for study of electric and noise properties of HEMT transistors. HELENA employs a Q2D (Quasi 2D) model to solve the physical equations, that incorporates the most important bi-dimensional effects into a mono-dimensional model with relatively small computational effort. The de-

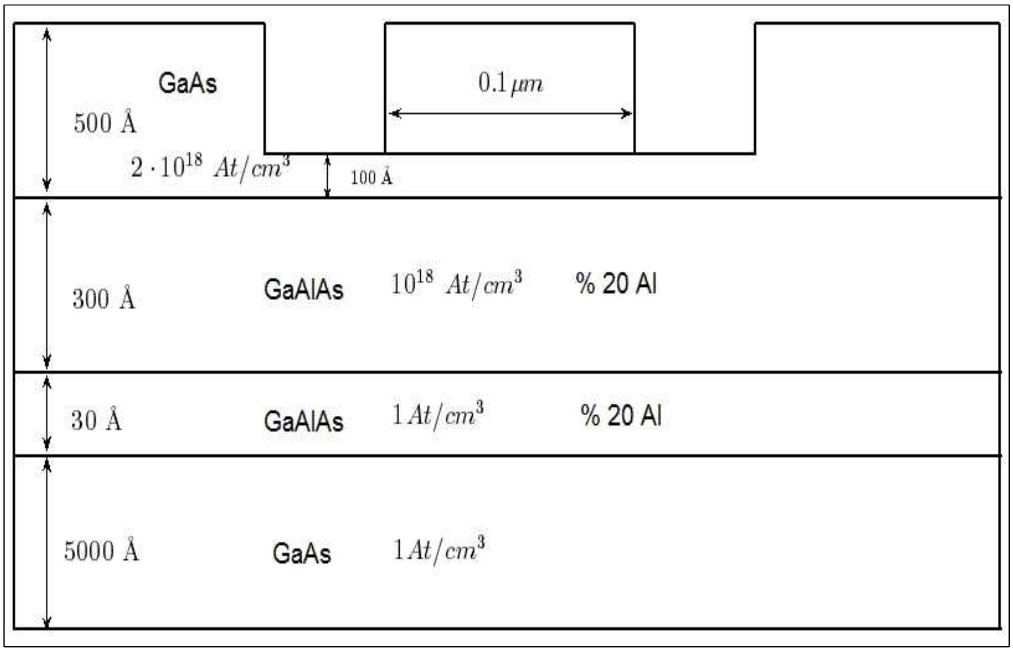


Figure 5.1: Geometrical and Physical parameters of the simulated HEMT device

vice considered for this simulation was the GaAs/AlGaAs HEMT proposed in [115]. The structure of the device is depicted in Figure 5.1. The device has a highly doped ($2 \cdot 10^{18}$ atoms/cm^3) cover layer, with a thickness of 500 Å, a 0.1 μm gate length, while the 300Å AlGaAs layer is doped with a concentration of 10^{18} atoms/cm^3, and has a 20% mole fraction of Al. The spacer layer has a relative reduced thickness (30 Å), while the buffer layer is not doped and has a thickness of 5000Å. The small signal equivalent circuit of the HEMT employed is depicted in Figure 5.2 It consists of an RLC network in input, a voltage driven current generator, an RLC network in output and a coupling capacity between input and output. C$_{gs}$ capacity is related to the charge in accumulation layer and is determined by the voltage variation between gate and source, while C$_{gd}$ is related to the charge in accumulation layer between gate and drain. The current generator driven by voltage at C$_{gs}$ is related to the fact that current of the channel is modulated by gate voltage by means of thickness of the accumulation layer, while time delay τ is due to the carriers and to the signal propagation through gate, since its dimensions cannot be neglected at the frequencies where the HEMT is employed. For a more detailed description of the model, see [115].

The database has been generated by means of statistical variations of physical pa-

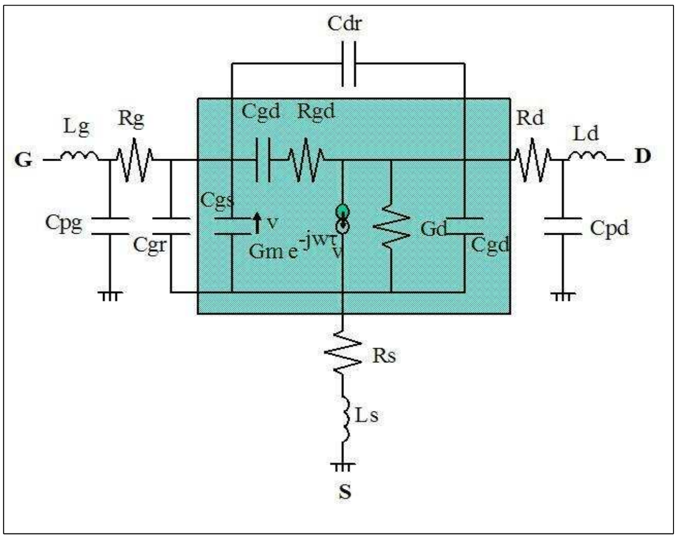


Figure 5.2: Small signal equivalent circuit

rameters, taking also small scale variations into account. The procedure employed is the following:

- A nominal model has been implemented for a single transistor, starting from physical parameters.
- A sensitivity analysis has been carried out in order to select the statistical variables of the model. Large scale and small scale variability have been taken into account.
- A population of of devices has been generated by perturbation of physical parameters of the nominal transistor.
- Statistical tests have been performed on the MMIC realization in order to test if the specification on mean, variance and covariance matrix is fulfilled.
- A database of measures for S parameters and for I_{DS} has been generated by simulation in HELENA.

The small-signal model extracted by HELENA has been employed to simulate the MMIC (Monolithic Microwave Integrated Circuit) consisting of 196 transistors, on a square grid of 14 × 14.
A preliminary sensitivity study has been conducted in order to decide which physical

Physical Parameter	**Mean Value**	**Standard Deviation**
Thickness of the layer under gate	100 Å	3
Thickness of the 300 Å layer	300 Å	9
Thickness of the separation layer	30 Å	0.9
Gate length	0.1 μm	0.015
Carrier density in 300 Å layer	10^{18} atoms/cm^3	3^{16}

Table 5.1: Statistical physical parameters of the model, together with their mean value and their standard deviation

parameters influence the transistor circuital parameters the most. This sensitivity study, together with some results already known from literature ([161, 10, 96]), has led to the choice of the parameters described in Table 5.1. Each of the five statistical parameters was assumed to have a *Gaussian* distribution with mean and standard deviation of Table 5.1

To account for lack of manufacturing uniformity, statistical variation has been imposed to some physical parameters. In particular, two kinds of variations have been considered:

- **Large Scale:** the *same* devices belonging to two different MMICs may be different
- **Small Scale:** the devices belonging to the *same* MMIC have some variability that is related to their distance.

Both these variabilities were considered to generate the simulated measure database. Each of the five statistical parameters was considered independent from the others (this is consistent with MMIC production techniques), having a Gaussian distribution, while correlation matrix relative to homologous parameters at different positions in the same chip depended on distance (since neighbor devices should be more "similar" than distant ones).

The correlation coefficient $Y(d)$ between two transistors at distance d was computed in the following way:

$$Y(d) = Y_1(d) + Y_2(d) \tag{5.1}$$

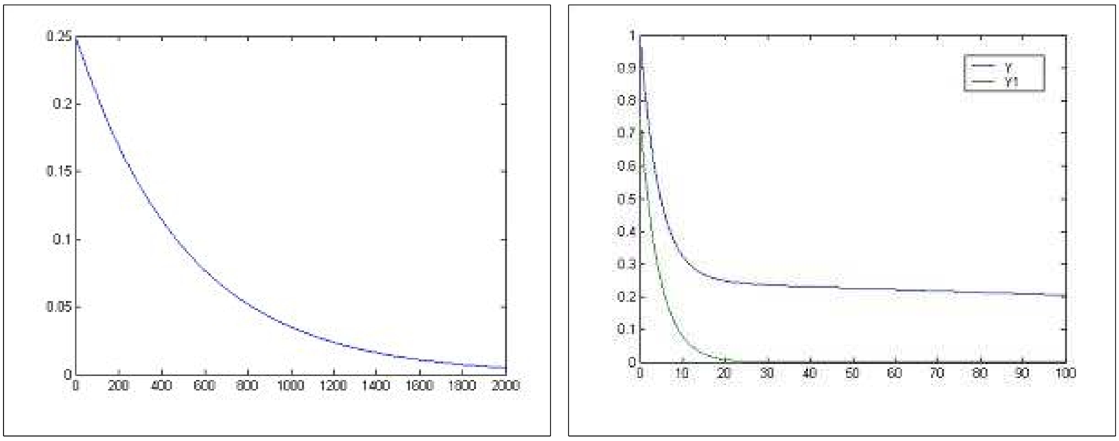


Figure 5.3: Distance-dependent correlation coefficient for statistical parameters. Left panel: Y_2. Right panel: Y and Y_1

where

$$\begin{aligned} Y_1(d) &= Ae^{-k_1 d} \\ k_1 &= -\frac{1}{d_1}\ln(\frac{10^{-4}}{A}) \end{aligned} \tag{5.2}$$

$$\begin{aligned} Y_2(d) &= (A-1)e^{-k_2 d} \\ k_2 &= -\frac{1}{d_1}\ln(\frac{10^{-4}}{1-A}) \end{aligned} \tag{5.3}$$

with $A = 0.75$, $d_1 = 40\mu m$ and $d_2 = 4000\mu m$. Values of Y, Y_1 and Y_2 are depicted in Figure 5.3.

By means of this correlation matrix, it is possible to simulate a more realistic database, since it is plausible that non-uniformities in manufacturing process affect neighboring transistors in a very similar way.

To account for large scale variability, 50 realization of a chip consisting of 30 × 30 transistors have been generated for *each* of the five statistical parameters, having thus 50^5 different MMICs. The next step of database generation has been to choose one of those 50^5 realizations, and to test if it has been generated correctly. For each of the five physical parameters statistical tests have been conducted satisfactorily on mean, variance, on Gaussianity of the marginal distribution and on correlation matrix.

The final step of measurement database has been to simulate each transistor to obtain S parameters and I$_{ds}$ currents. By means of HELENA, it has been possible to consider S parameters for different operation points: V$_{ds}$ ranged from 0.25 V

to 3 V with a step of 0.25 V. V_{gs}, instead, ranged from a value α and 0 V, with α depending on the transistor selected (usually from -0.9 V to -0.5 V). For each operation point, parameters S were evaluated for frequency F, with F ranging from 1 to 25 GHz, with a step of 1 GHz. Each S parameter had an average of 14000 different measurements.
To implement I_{DS} measurement, due to the problems encountered with HELENA in performing this simulation, another simulation tool, Agilent ADS, has been employed. Starting from the non-linear models extracted with HELENA, it has been possible to evaluate I_{DS} with ADS considering a V_{ds} and V_{gs} varying like in the S-parameters simulation.
The final database has been restricted to a 14 × 14 chip for numerical problems encountered with HELENA and for the reason that usually a larger measurement database is rarely available.

5.3 Non-Linear Model

Starting from the measurement database discussed in 5.2, a non-linear model has been extracted and subsequently validated.
The non linear model, described in [42] has been extracted from the current database. An MMIC linear model consists of a non linear nominal model for a device of the chip and of a covariance matrix that accounts for the relationship between different parameters of the model for a single device and for different devices. the model extraction procedure is depicted in Figure 5.4. An empirical non linear model is chosen to describe the behavior of the devices, therefore only a limited number M of the overall parameters set is chosen as "statistical" (only the parameters whose variation influences in a relevant way the model are considered statistical). From the device in the center of the chip the M parameters of the non linear model and the non statistical ones are estimated. Subsequently statistical parameters are estimated on the whole chip, and their mean and variance is evaluated. Once the model of single device has been described by means of its mean and variance, the covariance matrix of the MMIC is estimated. Considering all the couples at a given distance d, the correlation among all feasible distances in the chip is evaluated, and therefore the MMIC covariance matrix is evaluated. The nominal non linear model employed is depicted in Figure 5.5. The model consists of intrinsic and extrinsic elements. Moreover, it is possible to divide the parameter set into a class of linear elements,

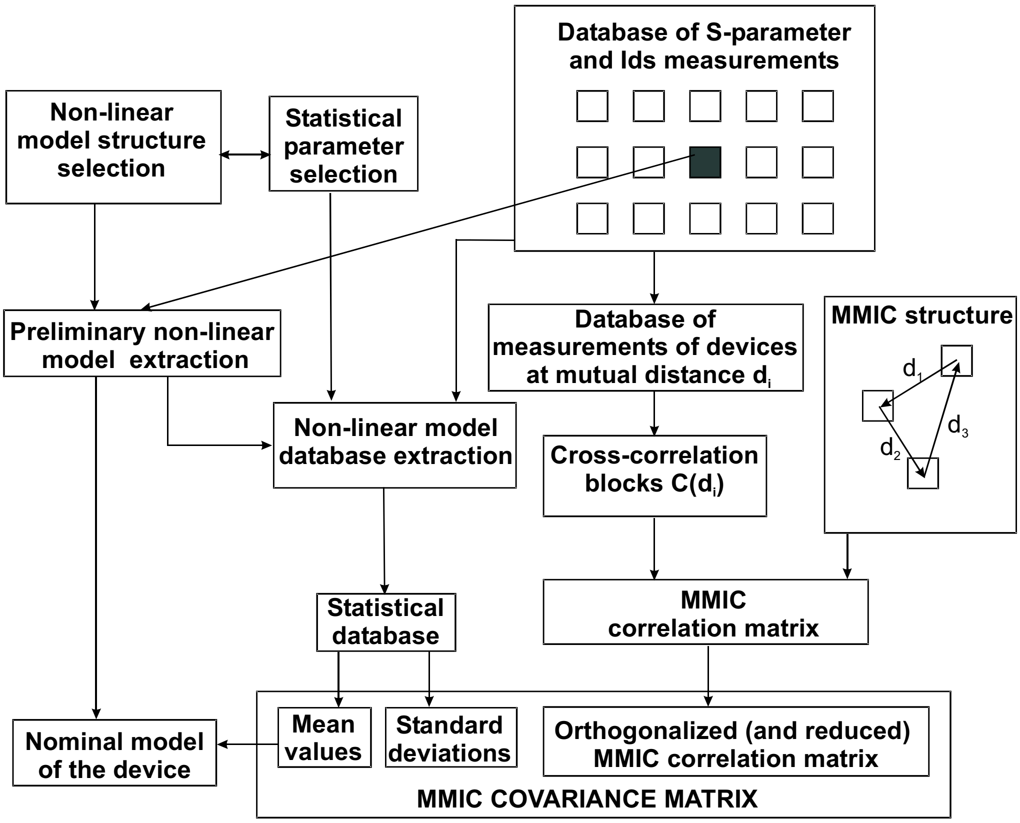

Database of S-parameter and Ids measurements
Non-linear model structure selection
Statistical parameter selection
Preliminary non-linear model extraction
Database of measurements of devices at mutual distance d_i
MMIC structure
d_1
d_2
d_3
Non-linear model database extraction
Cross-correlation blocks $C(d_i)$
Statistical database
MMIC correlation matrix
Nominal model of the device
Mean values
Standard deviations
Orthogonalized (and reduced) MMIC correlation matrix
MMIC COVARIANCE MATRIX

Figure 5.4: Statistical model extraction procedure

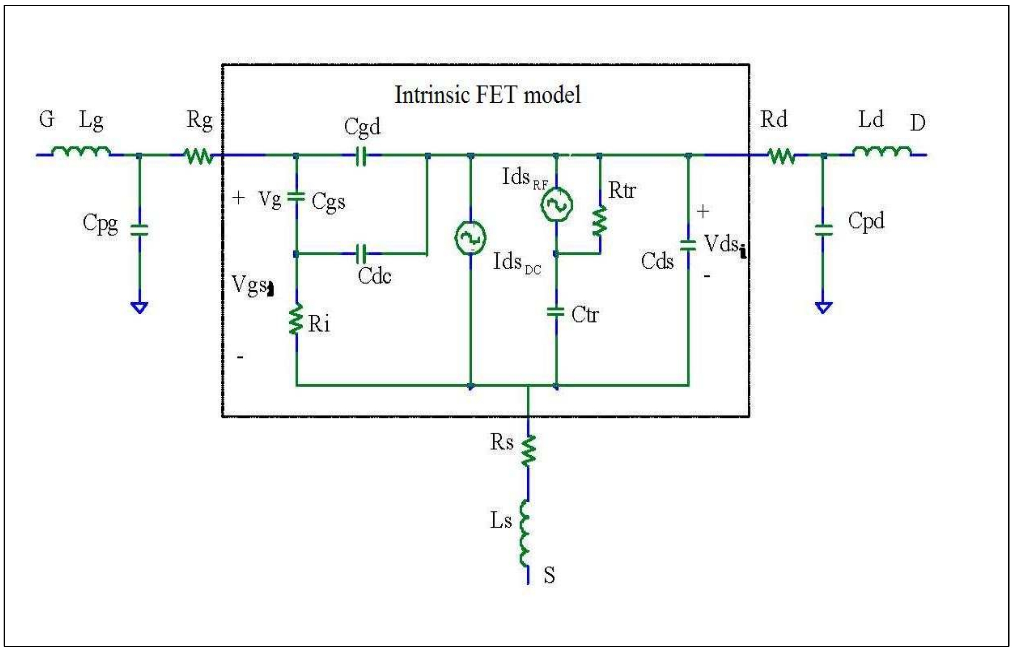


Figure 5.5: FET Nominal model

whose value is independent from the polarization of the circuit, and a class of non linear elements, containing C$_{gs}$, C$_{gd}$, I$_{ds}$, DC I$_{dsRF}$ and R$_i$, whose value depends on the polarization. Capacitors C$_{gs}$, C$_{gd}$ and C$_{ds}$ account for charge on gate contact, while I$_{dsDC}$ and I_{dsRF} model frequency dispersion of transconductance and output conductance. All the non-linear elements in the model are described by means of empirical equations [6, 47, 42]. Statistical parameters are chosen considering:

- Empirical parameter p variance.
- S parameters sensitivity to p.
- Noise introduced by optimization techniques employed to estimate p.

According to these studies, a set of 11 parameters has been described as statistical [42]:

- **I$_{pk}$:** DC current amplitude factor.
- **v$_{pk0}$:** V$_{gs}$ that gives the maximum DC transconductance.
- **d:** Linear dependence coefficient of v$_{pk0}$ from V$_{ds}$.

- **u:** output DC conductance.
- $\mathbf{I}_{pk1}$, $\mathbf{I}_{pk2}$: amplitude factor of RF transconductance model.
- $\mathbf{I}_{pk3}$, $\mathbf{I}_{pk4}$: amplitude factor of RF output conductance model.
- $\mathbf{Q}_{ogs}$: amplitude factor of Q_{gs}
- $\mathbf{Q}_{ods}$: amplitude factor of Q_{ds}
- $\mathbf{Q}_{ogd}$: amplitude factor of Q_{gd}

Non-linear model extraction can be summarized by the following steps:

1. Extraction of *linear* elements from parameters measures on different operation points.
2. Extraction of multi-bias linear model by means of the solution of a set of equations to obtain extrinsic parameters in different operation points.
3. I_{dsDC} paramters extraction by means of Angelov empirical equation ([6]) for the static current values measured.
4. Extraction of the parameters of the empirical equations that describe non-linear components by means of fitting of the values obtained at points 1, 2 and 3. Usually this step is done by means of a Simulated Annealing-Gradient mixed approach

MMIC covariance matrix is then estimate by considering different sets of transistor couples at distance d (ranging from a minimum value d_{min} to a maximum value depending on chip dimensions). For each distance, the correlation between all the parameters of the two transistors of the couples is evaluated. It is evident that, as the distance d increases, the number of couples at distance d decreases, therefore suitable hypothesis tests to assess the reliability of such correlation for a given distance are performed to set the maximum distance d_{max} that allows a significant value of correlation.

The correlation matrix obtained by this procedure is a block matrix, where block (i, j) is the correlation matrix of the parameters of transistor i and of transistor j.

5.4 Model Validation

Once the model has been extracted, a validation procedure has to be implemented in order to evaluate the correspondence between the modeled circuits and the real ones. Usually a new database is generated starting from the model parameters by means of a Monte Carlo simulation, and it is confronted to the measured one by means of hypothesis testing. In particular, mean, standard deviation, autocorrelation and cross correlation blocks of the MMIC covariance matrix are tested.

5.4.1 Statistical Testing

S parameters population in the measured database and in the simulated one are confronted with statistical tests. The significance level is related to the probability of false negatives (i.e. the probability of having an error between the two populations while there is no error), and usually a significance level α ranging from 0.05 to 0.1 is used. If the population examined consists of m variables, the probability of a false negative is:

$$\alpha_{cumul} = 1 - (1 - \alpha)^m \tag{5.4}$$

A different significance level α has to be employed to obtain $\alpha_{cumul} = 0.1$, according to the hypothesis tested. In particular:

- **Mean values and standard deviations:** since the 4 scattering parameters have both real and imaginary parts, there are 8 dimensions, therefore $\alpha = 0.013$.
- **Autocorrelation coefficients:** the number of dimensions in this case is 64, but the autocorrelation matrix is symmetric and its main diagonal elements are equal to 1, therefore there are only 28 independent variables and $\alpha = 0.00263$.
- **Crosscorrelation coefficients:** as before, the number of dimension is 64 and the matrix is symmetric, but the main diagonal elements are not forced to be 1, therefore the number of independent dimensions is 36, and $\alpha = 0.00164$.

For mean values testing, the following variable is employed:

$$Z = \frac{\mid \mu_{meas} - \mu_{mod} \mid - \Delta}{\sqrt{\frac{\sigma_{Meas} + \sigma_{Mod}}{Q}}} \tag{5.5}$$

where the null hypothesis is

$$H_0 :\mid \mu_{meas} - \mu_{mod} \mid = \Delta \tag{5.6}$$

against the alternative hypothesis:

$$H_1 :\mid \mu_{meas} - \mu_{mod} \mid < \Delta \tag{5.7}$$

and μ and σ are the mean values and standard deviations of the modeled and measured distributions, z is a Gaussian variable and Q is the sample size. The region where the hypothesis H_1 is accepted is the one where:

$$Z < -Z_a \tag{5.8}$$

where z_a is the cumulative distribution function of z for a suitable value of a.
For what concerns standard deviation, hypothesis testing is passed with a maximum error Δ_σ with a significance level α with the following test:

$$Z_{\sigma^2} = \frac{(\mid \sigma^2_{Meas} - \sigma^2_{Mod} \mid - \Delta_{\sigma^2})}{\sqrt{\frac{\sigma_{\sigma^2_{Meas}} + \sigma_{\sigma^2_{Mod}}}{Q}}} \tag{5.9}$$

where the null hypothesis:

$$H_0 :\mid \sigma^2_{Meas} - \sigma^2_{Mod} \mid = \Delta_{\sigma^2} \tag{5.10}$$

is tested again the alternative hypothesis:

$$H_1 :\mid \sigma^2_{Meas} - \sigma^2_{Mod} \mid < \Delta_{\sigma^2} \tag{5.11}$$

, where Z_{σ^2} is a Gaussian variable, σ and σ_{σ^2} are the variance and the variance of the variance of the modeled distribution and Q is the sample size. The error on variances is related to the error on standard deviations by means of the following:

$$\Delta_{\sigma^2} = \begin{cases} 2\Delta_\sigma - (\Delta_\sigma)^2 & \text{if} \sigma_{Meas} > \sigma_{Mod} \\ 2\Delta_\sigma + (\Delta_\sigma)^2 & \text{if} \sigma_{Meas} < \sigma_{Mod} \end{cases} \tag{5.12}$$

The acceptability region for H_2 is defined by:

$$Z_{\sigma^2} < -Z_{\sigma^2 \alpha} \tag{5.13}$$

where $Z_{\sigma^2\alpha}$ is the cumulative distribution function of Z_{σ^2} for a given significance level α. Test on auto- and cross-correlation is done by means of Fisher's Z, with the following test:

$$Z_c = \frac{\mid c_{Meas} - c_{Mod} \mid - \Delta_c}{\sqrt{\dfrac{2}{Q-3}}} \tag{5.14}$$

where the region of acceptance of the alternative hypothesis $H_1 :\mid c_{Meas} - c_{Mod} \mid < \Delta_c$ is defined as in (5.8)

5.4.2 Principal Component Analysis

Since the parameters of the model are correlated, before performing a Monte Carlo simulation PCA (see section 2.3.4) is usually performed. In fact, by means of this technique it is possible to:

- Transform the original dataset into a new one whose parameters are uncorrelated
- Reduce the dimensions to a specific value (defined, for instance, by the amount of explained variance)

Moreover, since many CAD simulation tools are not able to perform a Monte Carlo simulation on a set of correlated variables, it is possible to overcome this limitation by performing orthogonalization by means of PCA. The main drawback of such a technique is that the new parameters do not have a straightforward relationship with physical parameters, making it difficult to interpret the results.
After performing PCA, each parameter of the model can be expressed in terms of principal components in the following way:

$$p = \mu_p + \sigma_p(\sum_{i=1}^{N_{PC}} F_i PC_i) \tag{5.15}$$

where N_{PC} principal components has been considered, and F and P are the principal factors and principal components respectively.

5.4.3 Independent Component Analysis

Independent Component Analysis has been applied to the parameters of the model, as a further step with respect to PCA. In fact, since uncorrelatedness is a looser

requirement than independence, and they are the same only if the variables are Gaussian, it is plausible to look for an independent representation rather than for an uncorrelated one, once it is known that parameters do not have a Gaussian distribution. While it is plausible to assume that technological parameters of the MMIC exhibit Gaussian distribution, the extracted parameters of the model, since they are related by empirical non-linear equations to the physical parameters, are in general not Gaussian.
Therefore the following procedure has been implemented to estimate the independent components:

- Data were preprocessed to remove the mean and to set the standard deviation to 1
- Independent components were extracted by means of FastICA algorithm, with a symmetric approach (to have better precision)
- Each independent component probability distribution has been estimated.

However the Monte Carlo simulation tool could not implement probability densities different from normal, uniform and log-normal, and therefore each independent component density was approximated with one of these available functions. In particular, since the majority of components were super-Gaussian, independent component IC_i was first transformed into IC_{i1} by adding a suitable mean μ_i (to make the probability density defined only for positive values), the corresponding log-normal density was estimated with maximum likelihood, and then μ_i mean was subtracted. This procedure, however, could be somehow troublesome since there were some independent components whose pdf could be fitted unsatisfactorily by a log-normal density. The results of the ICA implementation for the parameters database will be discussed in the next section. After performing ICA, each parameter of the model can be expressed in terms of independent components in the following way:

$$p = \mu_p + \sigma_p(\sum_{i=1}^{N_{IC}} A_i IC_i) \tag{5.16}$$

where N_{IC} is the number of independent components, and A and IC are the mixing coefficients and the independent components respectively.

5.4.4 Results and Discussion

Statistical validation on the model has been performed, together with PCA and ICA preprocessing. Results of the statistical tests will be shown first, then an analysis of Principal components relationship with physical parameters will be discussed. Finally, a comparison between PCA and ICA simulation will be done in order to evaluate the effectiveness of the two models.

Different part of the MMIC have been considered. In particular, tests have been conducted on an MMIC consisting of 2 active devices at distance d_{min} and on an MMIC consisting of two devices at distance $2d_{min}$. The statistical test have been performed on the real and imaginary part of S-parameters in the 1-25 GHz frequency range, with a cumulative level of significance $\alpha = 0.1$. Tests have been conducted employing PCA analysis to reduce dimension such that the current representation explains the 95 % of the overall variance. The results of the validation procedures for the S-parameters in terms of mean, standard deviation, auto- and cross-correlation are shown in Tabb. 5.2-5.4 for the case without PCA and in Tables 5.5-5.7 for the case where PCA prerocessing has been performed. Statistical test on means and standard deviations gives the percentage difference that has to be accepted in order to have a given percentage of success (in this case 50 and 75 %). Statistical test on auto- and cross-correlation instead gives the success percentage according to the error accepted (in this case 0.3, 0.4, 0.5, 0.6)

Statistical test show that:

- The difference of the means of simulated and measured dataset is almost identical for all S-parameters.
- The standard deviation difference is reduced in the case of PCA pre-processing. In fact, S_{21} parameter fails both in its real and imaginary part without PCA, while with PCA the imaginary part of it passes the test with probability 75% with an accepted error of 94%.
- The auto- and cross-correlation show a smaller difference in the case of PCA preprocessing.

To sum up, these results confirm the fact that PCA leads to a more efficient representation, since its components are uncorrelated.

S-parameters	Mean Values		Standard Deviations	
	50 %	75 %	50 %	75 %
Re[S_{11}]	0.2	0.6	25	25.5
Im[S_{11}]	0.7	0.8	25	25.5
Re[S_{12}]	1.7	1.75	23	23.5
Im[S_{12}]	1.6	1.7	31	34
Re[S_{21}]	1.2	1.4	> 100	> 100
Im[S_{21}]	2.5	2.6	> 100	> 100
Re[S_{22}]	0.9	1.1	> 100	> 100
Im[S_{22}]	1.9	2	91	98

Table 5.2: Statistical test on means and standard deviations of S-parameters: difference accepted between simulated and measured vales to obtain 50 and 75 % success for hypothesis tests on mean and standard deviations in 1-25 GHz frequency range

Accepted Error	0.3	0.4	0.5	0.6
Autocorrelations	28.66	41.44	47.11	62.11
Autocorrelation signs	78.22	78.22	78.22	78.22

Table 5.3: Statistical test on S-parameters autocorrelation coefficients: percentage of success of hypothesis test between simulated and measured parameters. The accepted error is the accepted percentage difference

Accepted Error	0.3	0.4	0.5	0.6
Cross-correlations	39.37	50.12	69.25	82
Cross-correlation signs	87.87	87.87	87.87	87.87

Table 5.4: Statistical test on S-parameters cross-correlation coefficients: percentage of success of hypothesis test between simulated and measured parameters. The accepted error is the accepted percentage difference

S-parameters	Mean Values		Standard Deviations	
	50 %	75 %	50 %	75 %
Re[S_{11}]	0.1	0.4	18	18.5
Im[S_{11}]	0.6	0.7	19	19.5
Re[S_{12}]	2.3	2.4	16	16.5
Im[S_{12}]	2.1	2.15	19	27
Re[S_{21}]	1.5	1.6	> 100	> 100
Im[S_{21}]	2.5	2.6	94	> 100
Re[S_{22}]	1	1.1	17	17.5
Im[S_{22}]	2.5	2.6	25	26

Table 5.5: Statistical test on means and standard deviations of S-parameters with PCA: difference accepted between simulated and measured vales to obtain 50 and 75 % success for hypothesis tests on mean and standard deviations in 1-25 GHz frequency range

Accepted Error	0.3	0.4	0.5	0.6
Autocorrelations	44.22	50.11	54.6	60.6
Autocorrelation signs	80.66	80.66	80.66	80.66

Table 5.6: Statistical test on S-parameters autocorrelation coefficients with PCA: percentage of success of hypothesis test between simulated and measured parameters. The accepted error is the accepted percentage difference

Accepted Error	0.3	0.4	0.5	0.6
Cross-correlations	56.62	70.68	88.75	87.75
Cross-correlation signs	88.50	88.50	88.50	88.50

Table 5.7: Statistical test on S-parameters cross-correlation coefficients with PCA: percentage of success of hypothesis test between simulated and measured parameters. The accepted error is the accepted percentage difference

A further investigation has been conducted on the principal components of the parameters of a single device, in order to check if it is possible to relate the extracted component to the original physical parameters. In [94] such investigation had been conducted in order to diagnose the physical mechanism limiting manufacturing uniformity of InP HEMT devices. Ten DC figures of merit were measured for 50 devices and the principal components of the correlation matrix were extracted and related to the physical parameters. Here, the principal components are evaluated for the statistical non-linear model of the device, in order to check if the proposed model is able to preserve the relations between extracted model parameters and original physical parameters. The analysis of the relative importance of each extracted principal component PC_i in accounting for the total variance shows that the first 3 components account for $93,2\%$ of the total variance: in particular particular, PC_1 accounts for 48.2%, PC_2 for 30.9%, PC_3 for 14.1%. The contribution of each component PC_i to the variance of the 11 statistical parameters of the model is depicted in Figure 5.6. The association of the first three principal components to the model parameters I$_pk$ (DC current amplitude factor), V$_{pk0}$ (Vgs for maximal DC transconductance), u (DC output conductance), allows to point out their relation to variations of the physical parameters n_s, the carrier concentration in the channel of the device mainly depending in turn on the thickness and carrier concentration in the highly doped layer, and of L_g. PC_1 is associated with the variations of L_g: in fact, PC_1 shows negative correlation both with I$_{pk}$ and V$_{pk0}$, and this suggests a relation between L_g increase and transconductance drop due to velocity saturation. PC_2 is associated with n_s, as it is strongly correlated both to V_{pk0} (with the minus sign) and to I$_{pk0}$ (with the plus sign). Finally, PC_3 is strongly correlated to the output conductance u, but no association with any physical parameter has been found [94]. It has to be noted that PC_1 and PC_2 explain also almost the totality of the dynamic model parameters variance.

Further tests have been conducted on an MMIC to make a comparison between ICA and PCA performances. In particular the database has been validated once again by means of Monte Carlo analysis using the independent components of the model rather than the principal components, this time for a *single* device. For what concerns S-parameters auto-correlations, ICA-based model seems more accurate. In fact, statistical validation yields a success percentage of 57.1 % for an accepted error of 0.4, while for the PCA model it is 50.1%. The result of the validation procedure in terms of mean and standard deviation of the S-parameters with an

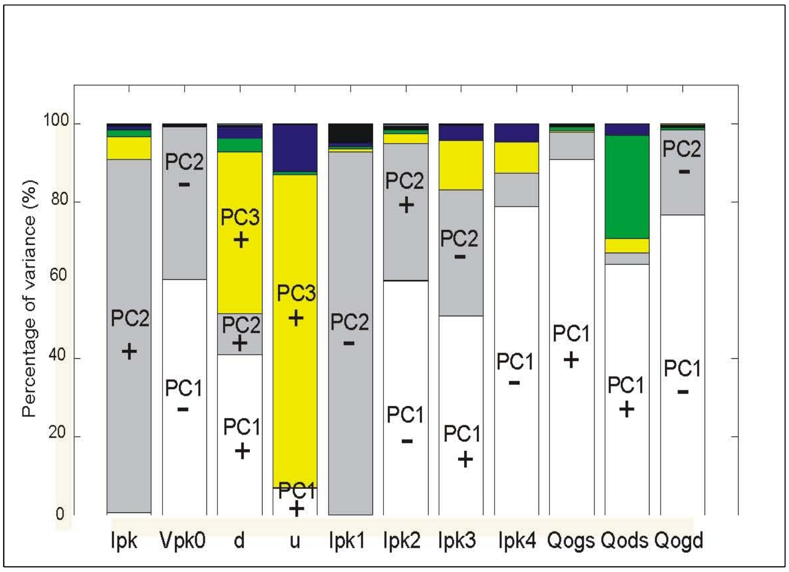


Figure 5.6: Contribution of each component PC_i to the variance of the 11 statistical parameters

Independent-Component-based single transistor model are described in Table 5.8. A comparison between the measured database and the simulated one for both PCA and ICA in terms of S-parameters population, mean and variance is depicted in Figures 5.7-5.18, where the PCA-based model analysis is on the left side of the figures, while the ICA-based one is on the right. For what concerns S-parameters statistical population, the comparison is carried out for both real and imaginary part. For mean and standard deviation analysis the measured values are in red and violet respectively, while the simulated ones are in blue and green respectively. From the analysis shown, two considerations can be made. It is evident that ICA is more effective in representing S_{21} parameters. In fact, in the correlated model and in the PCA based model, almost all of the S_{21} parameters failed the statistical test, while in the ICA based model an acceptable difference between simulated and measured parameters was obtained for both real and imaginary parts of S_{21}.

On the other hand, the ICA based model suffers from lack of precision on the other parameters, if compared with PCA. One of the main reasons for this behavior may be the fact that log-normal density approximation was based on heuristic parameter choice that may reduce accuracy of the fitting. Moreover, the log-normal function may lack the ability to represent the variability of the independent components, and somehow "reduces" the set of values spanned by each component.

S-parameters	Mean Values		Standard Deviations	
	50 %	75 %	50 %	75 %
Re[S_{11}]	2	4	48	48
Im[S_{11}]	5	5	52	53
Re[S_{12}]	2.3	2.4	16	16.5
Im[S_{12}]	5	6	45	45
Re[S_{21}]	2	2	30	45
Im[S_{21}]	6	7	18	20
Re[S_{22}]	3	3	58	60
Im[S_{22}]	6	6	67	70

Table 5.8: Statistical test on means and standard deviations of S-parameters with ICA: difference accepted between simulated and measured vales to obtain 50 and 75 % success for hypothesis tests on mean and standard deviations in 1-25 GHz frequency range

Based on these considerations, more precise density fitting by means of Gaussian mixtures is currently under study. In fact, Gaussian mixtures have the additional advantage that they can be easily implemented in any CAD tools capable of performing Monte Carlo analysis, since each parameter can be treated as a linear sum of Normal distributed parameters, whose mean, variance and weight is fitted by means of maximum likelihood estimation. Moreover, Gaussian Mixture Models can be extracted automatically, without resorting to heuristic parameter choices as in the case of the log-normal distribution.

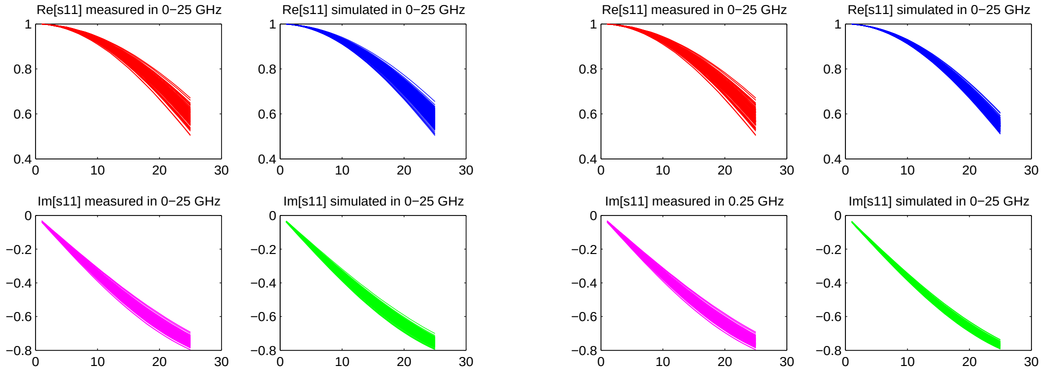


Figure 5.7: Simulated and measured S11 parameter. Left panel: PCA results. Right panel: ICA results

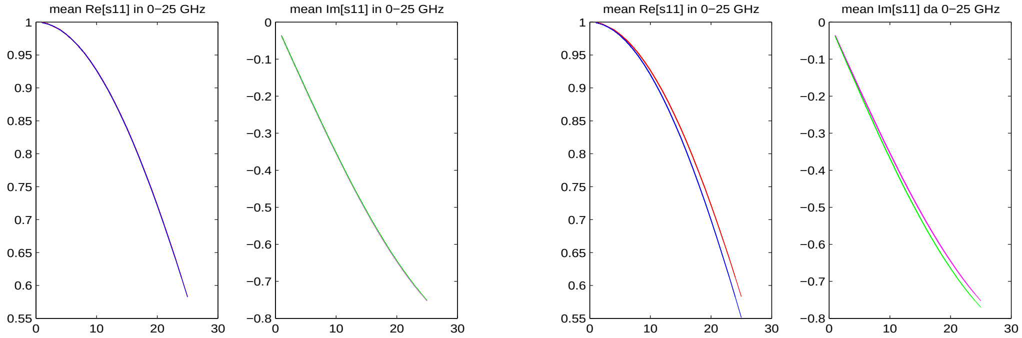


Figure 5.8: Simulated and measured S11 parameter: mean values. Left panel: PCA results. Right panel: ICA results

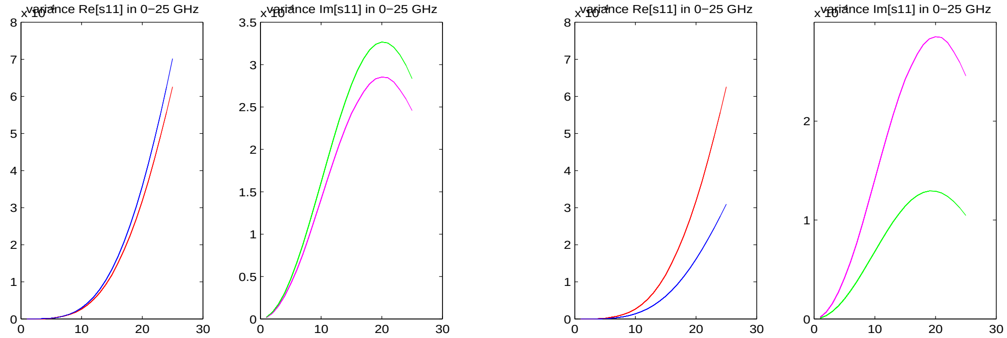


Figure 5.9: Simulated and measured S11 parameter: standard deviations. Left panel: PCA results. Right panel: ICA results

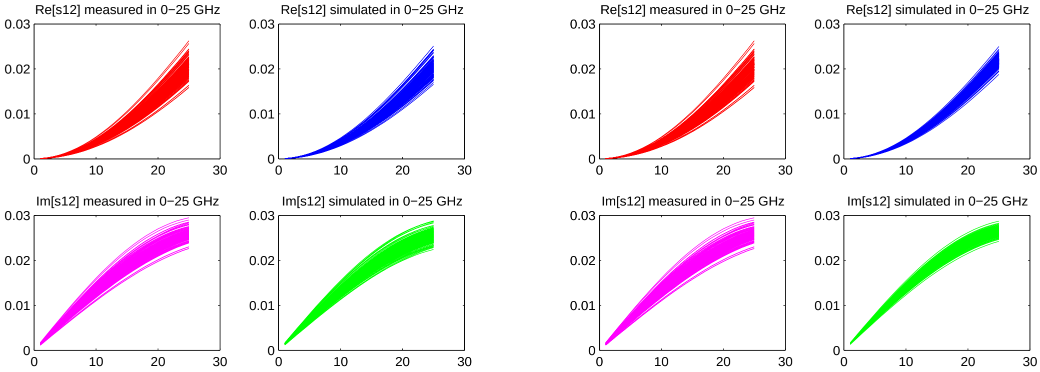


Figure 5.10: Simulated and measured S12 parameter. Left panel: PCA results. Right panel: ICA results

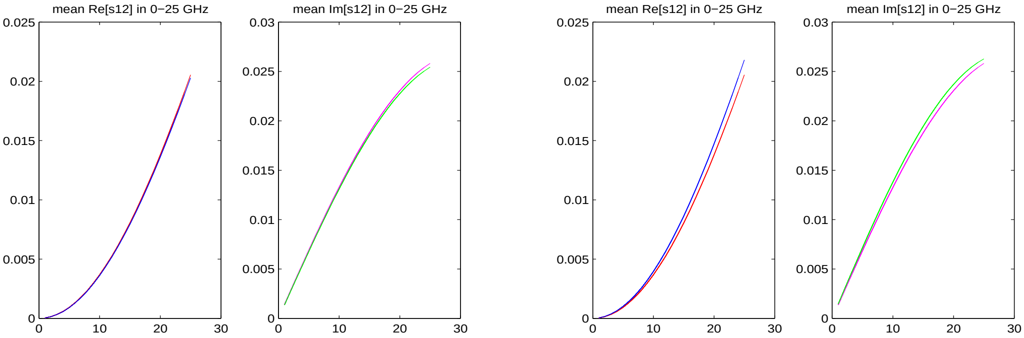


Figure 5.11: Simulated and measured S12 parameter: mean values. Left panel: PCA results. Right panel: ICA results

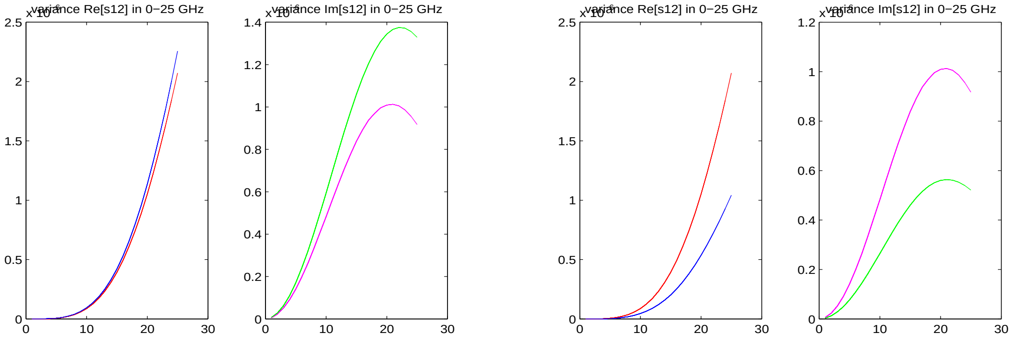


Figure 5.12: Simulated and measured S12 parameter: standard deviations. Left panel: PCA results. Right panel: ICA results

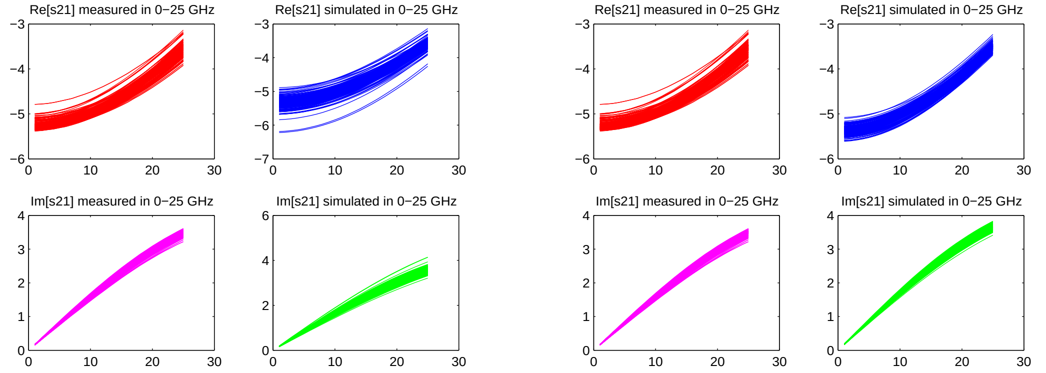


Figure 5.13: Simulated and measured S21 parameter. Left panel: PCA results. Right panel: ICA results

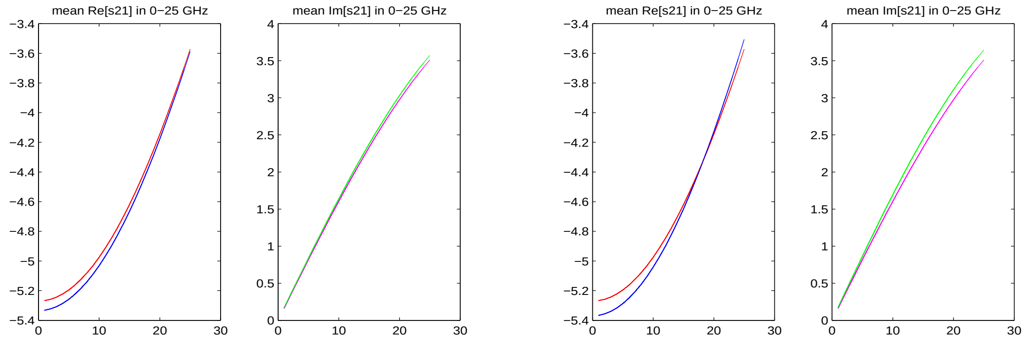


Figure 5.14: Simulated and measured S21 parameter: mean values. Left panel: PCA results. Right panel: ICA results

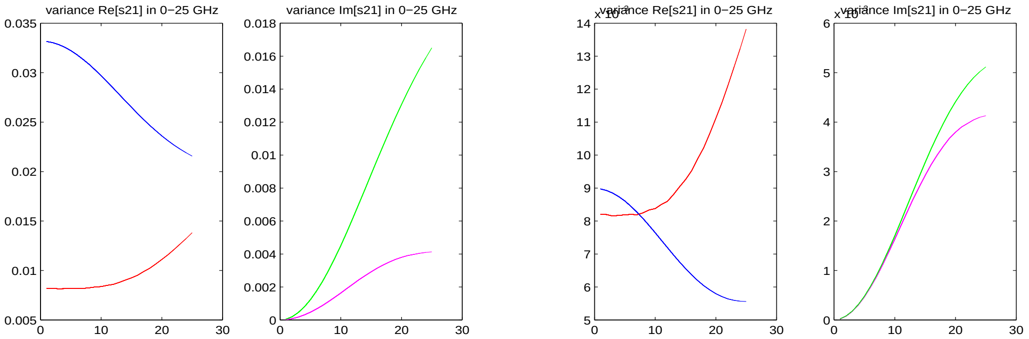


Figure 5.15: Simulated and measured S21 parameter: standard deviations. Left panel: PCA results. Right panel: ICA results

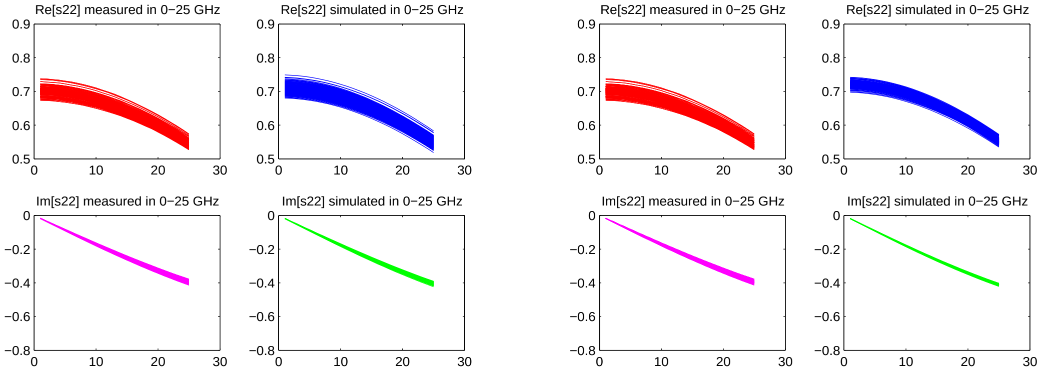


Figure 5.16: Simulated and measured S22 parameter. Left panel: PCA results. Right panel: ICA results

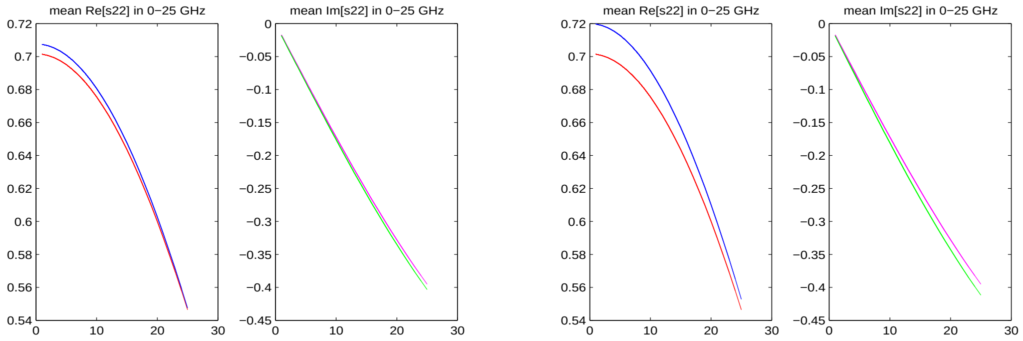


Figure 5.17: Simulated and measured S22 parameter: mean values. Left panel: PCA results. Right panel: ICA results

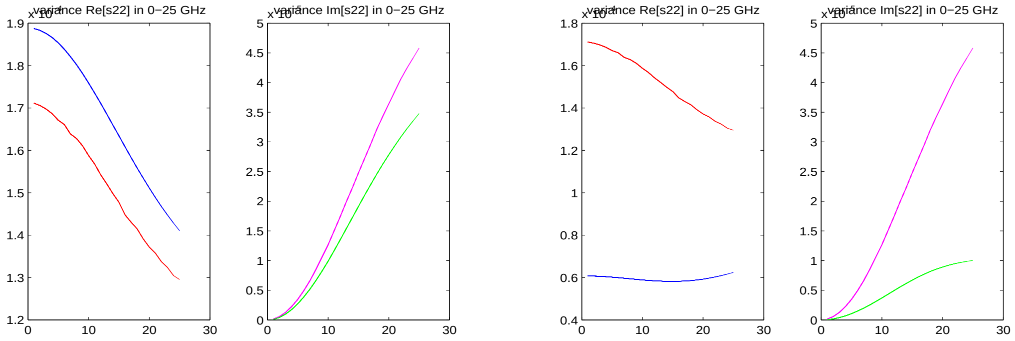


Figure 5.18: Simulated and measured S22 parameter: standard deviations. Left panel: PCA results. Right panel: ICA results

Chapter 6

Incorporation of prior knowledge in ICA

6.1 Introduction

In Chapter 2 the main principles underlying Independent Component Analysis (ICA), together with the most used approaches to extract sources, have been shown while in Chapter 3 some applications related to biomedical signal processing have been discussed. The results from previous discussion lead to the consideration that ICA is an effective technique as long as its assumptions are met, and it is possible to recover the sources in most of the cases. However, as pointed out in section 2.7.5, this technique is not *explicitly* meant for "structured" data, meaning that no ordering of the points of the observed signals or no regularities in original sources are exploited while performing extraction, and one may want not to lose this information while recovering sources.
Consider Fig. 6.1: the mono-dimensional signals in panels (a)-(c) are considerably different if one looks at them as "signals". However, if they are considered as *realizations* of a *random variable*, their histograms (an thus all statistics) are *identical*, as depicted in panel (d). In fact, signals in panels (b)-(c) have been generated by means of permutations of the points of signal in (a). Therefore, the use of an algorithm based merely on statistics neglects information content that is present in the signal structure. If some spectral characteristic of a particular source is known *a priori*, it is not possible to point this out using classical algorithms. A solution could be to use an approach based on second order statistics, presented in section

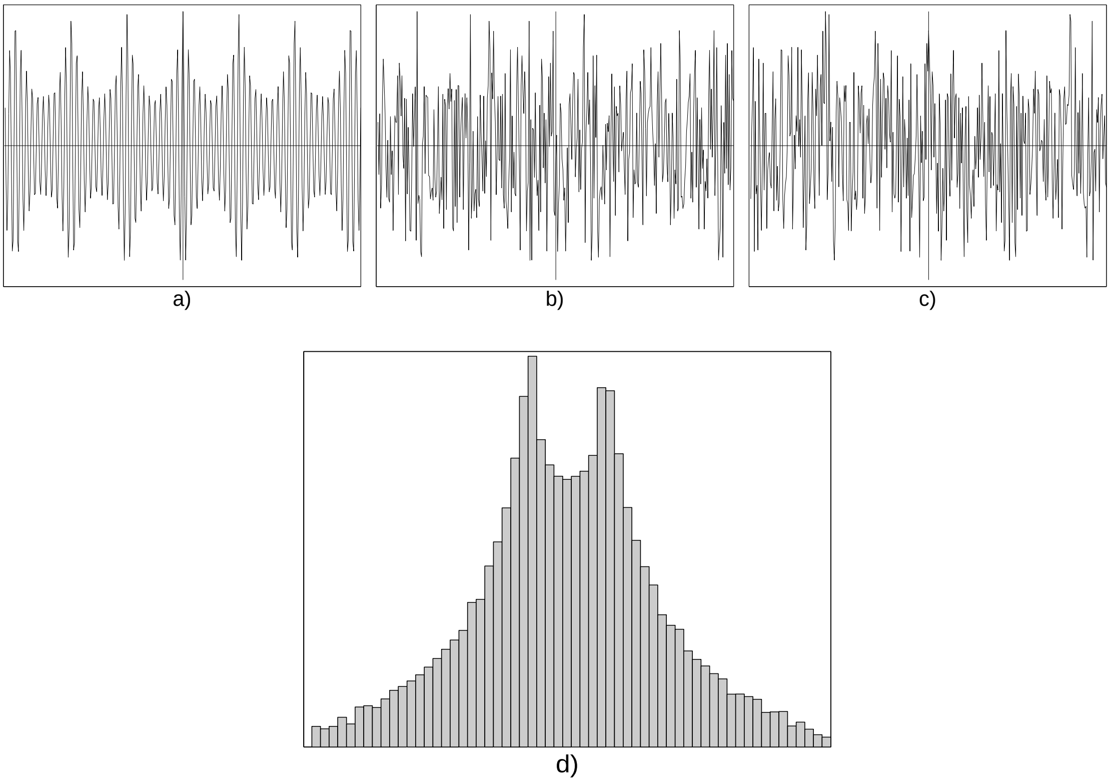


Figure 6.1: Example signals that have different spectral properties (panels (a)-(c)) while having the same histogram (panel (d))

2.7.5; however this solution is not flexible enough to take into account several kinds of prior information, as it points out only temporal regularities looking for lagged decorrelation, and its results are strongly affected by the choice of the lag.
This considerations are also evident in the fMRI case: consider the spatial map of an independent source (for instance, a single slice). It is known *a priori* that physiological activities that are observed by means of ICA do not (usually) involve regions of the brain described by a single voxel. Therefore an isolated active voxel in a spatial map is usually considered a noise artifact. However, while performing ICA on fMRI data, three-dimensional spatial data are considered neglecting the spatial ordering of points: one could scramble all the points in the same way for all the volumes, but the solution would be the same, in terms of mixing coefficients. This is at the same time a strength and a weakness of the ICA approach: while, on the one hand, the technique is completely blind and it is driven *only* by statistical independence, on the other hand it neglects additional information *available* on data, that therefore

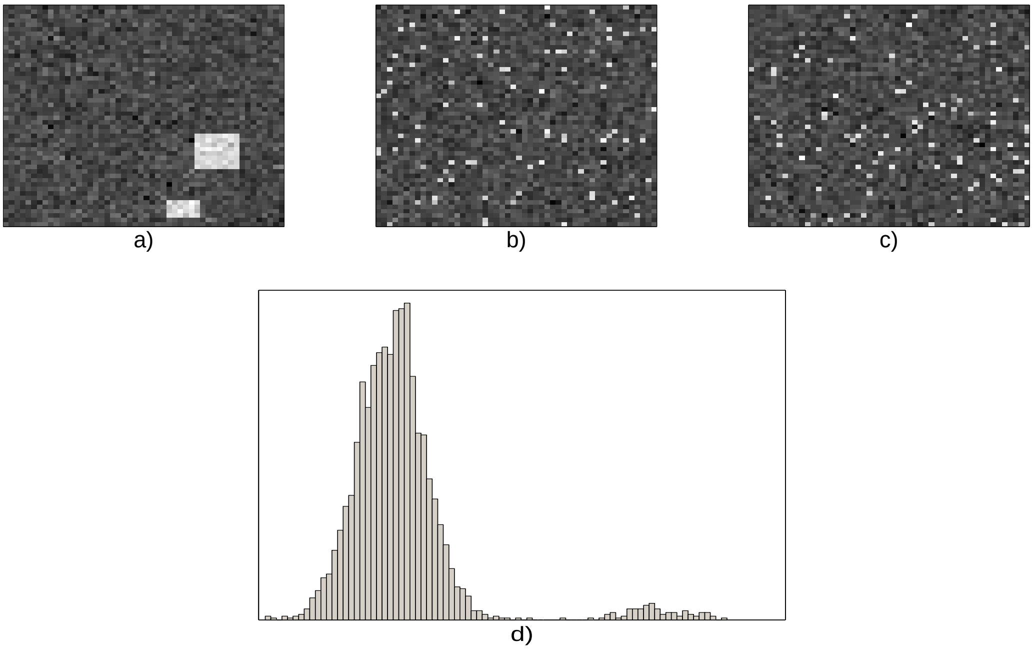


Figure 6.2: Example images that have different spectral properties (panels (a)-(c)) while having the same histogram (panel (d))

cannot be considered in the extraction. Consider now figure 6.2: in panels (a)-(c) three spatial maps with the *same* histograms are depicted. The first one (panel (a)) is the best candidate for a "realistic" activation, while the other two do not carry any information that can be used by a human expert to evaluate the results of an experiment. However, all the three maps have the same histogram, like the case presented in figure 6.1, and thus they will have the same statistics.

In general this situation may not be a problem. Consider, in fact, an independent component analysis using FastICA algorithm (Section 2.8) extracting one component at a time: for a given linear combination of original data a cost function is evaluated and it is maximized by means of a fixed point iteration. The fact that all the three maps have the same histogram *does not imply* that they will be extracted as sources: in fact, an independent source is pointed out by a local maximum in the cost function, not by the value of the cost function itself. Therefore, in the noiseless case, the fact that also a "meaningless" map has the same histogram as an "interesting" one is not a problem, since it has been proved that ICA is able to recover sources effectively. However, in presence of noise, that could be Gaussian, multidimensional, uncorrelated or correlated with some sources, the situation is not the same: the non-gaussianity measure itself becomes less and less robust with respect to that noise, and it may happen that meaningless activations can be considered as sources. One may want to use the prior information on spatial regularity to make the extraction more robust to noise, while preserving the strength of the blind approach.

Another situation where the use of prior information may help is the case when some specific property of only one or a group of sources is known in advance. In this case, if one is looking only for those independent components, it would be better to extract only "desired" sources, avoiding performing a full ICA decomposition and selecting *a posteriori* the components. This is particularly useful in problems where the number of the sources is considerably high, and there are some requirements on the extraction time.

In the next section the works presented in literature will be revised, while in section 6.3 a new framework to account for prior information in an ICA problem will be presented.

6.2 Previous work

In the previous section it has been pointed out that there are some cases where prior information on one or more sources is available and that it could be useful not to discard this knowledge. This kind of prior information can be divided in two classes: the case where the prior knowledge is "precise", and the one where this knowledge is rather "loose", i.e. when there are some clues about some sources, or about some properties of these sources. To give an example, knowing the spectrum of one source, or knowing the time course of an experiment is a rather specific knowledge, while for the signals presented in Figg. 6.1 and 6.2 the knowledge to discriminate between the signals is not as precise as in the previous case: knowing that the sources are "structured" may be enough to extract them effectively, even without looking for a particular power spectrum, but moving through a wider class of signals. It is possible to classify the work presented in literature on this topic according to this distinction, and in sections 6.2.1 and 6.2.2 current techniques will be revised.

6.2.1 Precise Prior information on some sources

The situation where some specific knowledge on one or more sources is available has been addressed in literature. To make the extraction more effective, i.e. to recover those sources first, several techniques have been proposed.

The situation described in [136] and in [19] is the following: suppose that the autocorrelation function of one source is known *a priori*, or at least an approximation of it is available. In this case, the aim of the proposed techniques is to recover that source first, since classical ICA algorithms do not order extracted components. It may be useful, in fact, to extract only that source, without extracting all the sources and identifying the desired one *a posteriori*. In [136] this is achieved by means of a modified cost function that considers also an error ϵ defined as follows:

$$\epsilon(\mathbf{w}) = \sum_{\tau=0}^{K-1} [r_{ss}(\tau) - r_{model}(\tau)]^2 \tag{6.1}$$

where K is the number of time points, $r_{ss}(\tau)$ is an estimate of the autocorrelation function of the source s and $r_{model}(\tau)$ is the known autocorrelation function. The new cost function becomes:

$$J(\mathbf{w}) = J_G(\mathbf{w}) - \lambda\epsilon(\mathbf{w}) \tag{6.2}$$

where $J_G(\mathbf{w})$ is a one-unit non-Gaussianity measure (i.e. negentropy, see Section 2.7.6). The optimization is carried out by means of a gradient method with different starting points (to avoid being always trapped in the same local maximum), and convergence is evaluated according to the weighting factor λ. The set of starting points that lead to the desired optimum are observed while increasing the value of λ, and it has been observed that this set expands as λ increases until an optimum value λ_o, and then it decreases.

Starting from the same assumption (i.e. some knowledge of the autocorrelation function), in [19] the authors propose a different solution, using second order methods (see Section 2.7.5). In this case sources are assumed to be uncorrelated at any time lag, such that:

$$\begin{aligned} \mathrm{E}\{s_i(k)s_i(k-\tau_r)\} &\neq 0 \\ \mathrm{E}\{s_i(k)s_j(k-\tau_r)\} &= 0 \quad \forall i \neq j \end{aligned} \tag{6.3}$$

and sources *must* have different autocorrelation functions to guarantee the identifiability of the model. Information on the autocorrelation of one source is exploited by means of an error function ϵ defined as follows:

$$\epsilon(t) = y(t) - by(t-p) = y(t) - by_p(t) \tag{6.4}$$

where b is a coefficient of a FIR filter with delay z^{-pT_s}, with the sampling period T_s is usually assumed to be one. Since sources are constrained to have different autocorrelation functions, a suitable choice of the delay p allows to identify sources whose autocorrelation has a peak for that delay. The algorithm proposed is based therefore on the minimization of the expected value of eq. (6.4) with respect to parameters $\mathbf{w}$ and b. The cost function is therefore:

$$\xi(\mathbf{w}, b) = \mathrm{E}\{\epsilon\} = \mathbf{w}^T\mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} - 2b\mathrm{E}\{y_p\mathbf{w}^T\mathbf{x}\} + b^2\mathrm{E}\{y_p^2\} \tag{6.5}$$

With the orthogonality constraint and some simplifications, the update rule becomes:

$$\mathbf{w} \leftarrow \mathrm{E}\{y_p\mathbf{x}\} \tag{6.6}$$

The ability of the algorithm to extract the desired source relies on the choice of the delay p that allows to discriminate between sources. The authors propose a methodology based on the autocorrelation $r_\mathbf{x}$ of observed signals $\mathbf{x}$, individuating the optimal delays by means of the peaks of $r_\mathbf{x}$. Once the delay has been chosen, they proved analytically that the optimization process extracts the desired source first.

Another algorithm to extract a source on which there is some prior knowledge has been proposed in [84]. The algorithm is called Principal Independent Component Analysis (PICA), and it is based on a cumulant approach where the learning is driven by a reference generator based on prior information on the source. It has been proved analytically that such approach allows extracting the desired source first. Another algorithm, presented in [105], that is based on Joint Approximate Diagonalization of Eigen-matrices (JADE), allows accounting for prior information that can be expressed in the form of a quadratic constraint.

The Bayesian approach, presented in section 2.7.2, makes it quite easy to express the additional prior information in terms of a prior to the problem. In [93] it has been shown how to incorporate some prior knowledge of the sources. The additional information considered was about the decorrelation of the observed signals (that can be obtained by whitening. section 2.6, and that is not required in a maximum likelihood approach), about the mixing coefficients and about the shape of the sources probability densities.

In [81] it has been shown how to account for prior information on the mixing matrix, considered in the form of a probability, and how to include this information in some of the most known ICA algorithms (like INFOMAX, [24], the Juttén-Herault algorithm, [88]). It has been shown that this prior information helps the sources extraction in terms of convergence and speed of the algorithms.

For what concerns biomedical data, the use of prior knowledge is particularly appealing, since often a considerably limited number of independent components is considered after the extraction, and some information is available on them. Starting from this consideration some approaches have been proposed to point out this characteristic in biomedical signal processing. In [108] and [109] the problem is addressed by means of constrained ICA (c-ICA), a technique derived from Lagrange multipliers optimization. In this way it is considered a constraint of correlation between the time course of a *temporal* independent component of an fMRI dataset and a "rough" template of the time-course of the experiment. Results presented in [109] show that is possible to extract first the task-related components, with no need of post-selection, by having some clue on their time course. The cICA framework is particularly flexible in including constraints in a theoretical rigorous way. The contrast function that points out independence is, as usual, negative entropy, and the additional information is considered as a constraint. Therefore the problems becomes:

$$\begin{aligned} &\text{maximize} \quad J_G(\mathbf{W}) \\ &\text{subject to} \quad \mathbf{g}(\mathbf{W}) \leq \mathbf{0} \quad \text{and/or} \quad \mathbf{h}(\mathbf{W}) = \mathbf{0} \end{aligned} \tag{6.7}$$

where $\mathbf{g} \in \Re^{\mu}$ accounts for the *inequality* constraints and $\mathbf{h} \in \Re^{\nu}$ for the equality ones.

The problem as posed can be solved by means of Lagrange multipliers, introducing the Lagrangian function $\mathcal{L}$:

$$\mathcal{L}(\mathbf{W}, \mu, \lambda, \mathbf{z}) = J_G(\mathbf{y}) + \mu^T \widehat{\mathbf{g}}(\mathbf{W}) + \frac{1}{2}\gamma \left\|\widehat{\mathbf{g}}(\mathbf{W})\right\|^2 + \lambda^T \mathbf{h}(\mathbf{W}) + \frac{1}{2}\gamma \left\|\mathbf{h}(\mathbf{W})\right\|^2 \tag{6.8}$$

where $\widehat{\mathbf{g}} = (\widehat{g}_1, \widehat{g}_2, \ldots, \widehat{g}_u)$, with $\widehat{g_p} = g_p + z_p^2$, where z_i is a set of μ slack variables that transform the inequality constraints into equality ones; the vectors $\mu \in \Re^{\mu}$ and $\nu \in \Re^{nu}$ are the vectors of *positive* Lagrange multipliers; the two quadratic penalties terms $\frac{1}{2}\gamma\| \cdot \|^2$ ensure that the in the minimization problem the condition of local convexity holds (i.e. the Hessian matrix is positive-definite). The optima of the problem in eq. (6.7) are those where the gradient of the Lagrangian function $\mathcal{L}$ with respect to $\mathbf{W}$, μ and ν is zero. The authors propose a Newton-like algorithm for the Lagrangian optimization, and a gradient ascent for multipliers update.

This algorithm was applied also in [85] to electromagnetic recordings of the brain (EEG and MEG). For EEG signals, information from ocular artifact was taken directly by thresholding a particular channel where this artifact was mostly evident generating a "reference" function for it. Therefore cICA with the aim of minimizing the correlation between estimated component and the reference function was performed, successfully extracting the artifact source first. The same technique was also applied to MEG recordings to identify the ocular and ECG artifacts.

A different approach was used by Calhoun in [35], to include information about the time course of fMRI activities in a spatial ICA extraction. The approach, called Semi-Blind ICA (sbICA), takes its moves from the INFOMAX algorithm ([24]) with the update rule based on the natural gradient (see Section 2.7.2 and [2]). Differently from [109] (where constraints are applied to time courses in a *temporal* ICA extraction), in this case *spatial* independent maps are constrained to have a *temporal* time courses that correlates up to a predefined amount with known reference signals. Each column $\mathbf{a}_i$ of matrix $\mathbf{A}$ (that is the estimate of the mixing matrix) is updated according to this criterion: first the classical INFOMAX update rule is applied, then at the next iteration, correlation ρ_i between each column and the desired time course is computed, and according to a tolerance value t_i defined by user, the column is updated in the following way:

$$\mathbf{a}_i = \begin{cases} \mathbf{a}_i & \text{if} \quad \rho_i \geq t_i \\ \mathbf{a}_i + c[f(\mathbf{a}_i)] & \text{if} \quad \rho_i < t_i \end{cases} \tag{6.9}$$

where $f(\mathbf{a}_i)$ is a function that tries to correct to estimated time course according to the prior information on it. In particular, after each iteration components are resorted such that every time each column will correspond to a specific time course, as the INFOMAX itself does not order components. The algorithm has proved to be effective on simulated and real fMRI data, and it has been confronted with GLM showing more robustness to a choice of the reference time-course (i.e. sbICA performs better than GLM when the information on the reference function is not perfect).

6.2.2 Loose Information on the sources

For what concerns a loose prior information on all the sources, or on a subset of the whole number, several works have been proposed in literature. In particular some works of Stone ([155, 156]) explore the possibility of considering other criteria together with independence. In [156], a new contrast function derived from INFOMAX by means of a linear combination of *temporal* and *spatial* independence for fMRI data analysis has been proposed. The starting hypothesis is that fMRI activations may be independent in the spatiotemporal domain, better than in the spatial or temporal domain alone. The algorithm proposed, called *spatiotemporal ICA*, has been tested on simulated sources that did not fulfill exactly the ICA assumptions ([118]) and was proved to be more effective than spatial or temporal independence. Moreover, information on the asymmetric probability density function was accounted for by means of skewed functions as approximating functions in the contrast function.

Another approach, presented in [155] is the use of *temporal predictability* as a *regularization* term in the contrast function. Temporal predictability of y is defined in this case as the log ratio of long term variance to short term variance of y, that is:

$$h_p = \frac{1}{2} \log \frac{\sum_{i=1}^{m} (\overline{y}_i - y_i)^2}{\sum_{i=1}^{m} (\tilde{y}_i - y_i)^2} \tag{6.10}$$

where the quantities $\overline{y}_i$ and $\tilde{y}_i$ are both exponentially weighted sums defined as follows:

$$\overline{y}_i = \lambda_S \overline{y}_{i-1} + (1 - \lambda_S) y_{i-1} \quad \text{with} \quad 0 \leq \lambda_S \leq 1 \tag{6.11}$$

$$\tilde{y}_i = \lambda_L \tilde{y}_{i-1} + (1 - \lambda_L) y_{i-1} \quad \text{with} \quad 0 \leq \lambda_L \leq 1 \tag{6.12}$$

with the half life h_L of $\tilde{y}$ much longer (a ratio of 100 has been proposed) than the corresponding half life of $\overline{y}$. The term in (6.10) is added to the spatial independence term with a weigh $\beta = 0.5$. Results on simulated overlapping maps showed that this regularizing factor led to an improvement in the separation.

Another approach proposed for fMRI data analysis is the so called *Probabilistic ICA*, proposed by Beckmann in [22, 23]. In this case, by means of a Bayesian framework, it is possible to give a different weight to different voxels in the brain. The aim of this approach is to improve the separation by imposing a smoothing factor in both temporal and spatial domain, selecting moreover the voxels that contribute the most to the activation maps.

In [67], a Bayesian framework with some priors on the autocorrelation of the sources have been proposed. The technique is applied to text recovery in linear mixtures, and Gibbs priors are employed to account for the edges in the images and results on simulated signals have been proposed, showing the effectiveness of the approach.

6.3 New approach to take information into account

As seen in section 6.2, several attempts have been made in order to incorporate prior information into the ICA extraction. It has been shown that this prior information can be rather "specific" or "loose"; according to this difference, several approaches have been proposed. However, all of the proposed approaches, given the fact that they are flexible within the class of prior information that the algorithm employed itself allows to add into the extraction, lack the possibility to account for information that cannot be *differentiable*, that cannot be expressed in a *closed form*, or that is heterogeneous. To overcome this limitation, and to provide a comprehensive framework for including almost "any" kind of prior information into an ICA extraction, a new approach has been developed.

As pointed out in section 2.7, to find an ICA decomposition, a contrast function has to be defined and it has to be optimized by means of some procedure. The new approach starts from FastICA algorithm (section 2.8), in the deflation approach, i.e. one component is extracted at a time; for the FastICA algorithm the contrast function is an approximation of negentropy (see eq. (2.118)), and the optimization is carried out by means of a Newton iteration.

For what concerns the new proposed algorithm, before explaining it in detail, it is possible to define which are both the contrast function and the optimization procedure.

- **Contrast Function**: To account for prior information, the new proposed contrast function is:

$$F = J_G(\mathbf{w}) + \lambda H(\mathbf{w}), \tag{6.13}$$

  with $F : \Re^m \rightarrow \Re$.

- **Optimization Procedure**: optimization of F is carried out by means of Simulated Annealing

Therefore the prior information is included into the algorithm by means of an additive term in the contrast function. The terms in eq. (6.13) will now be explained in detail:

J_G: this function is defined as in eq. (2.118), and is an approximation of negentropy by means of a non linearity G: $J_G(\mathbf{w}) = [\mathrm{E}\{G(\mathbf{w}^T\mathbf{z})\} - \mathrm{E}\{G(\nu)\}]^2$. As shown in section 2.7.6, maximization of J_G leads to independence.

H: the prior information is taken into account by means of this additional term in the contrast function. H can be related to some properties of the sources or of the mixing process and it holds: $H : \Re^m \rightarrow \Re$. According to the problem faced, H can be of any form, also a non differentiable one. Moreover, H has to be such that $H(a\mathbf{w}) = H(\mathbf{w})$, $\forall a \neq 0$ (a contrast function for ICA, as seen in Definition 3, must preserve its value for a class of equivalence). However, in this case, as $\mathbf{w}$ is constrained to have unit norm, the previous requirement reduces to: $H(-\mathbf{w}) = H(\mathbf{w})$.

λ: this parameter weighs the two functions and its importance can be crucial for the optimization.

The choice of Simulated Annealing, that will be explained in more detail is section 6.4, comes from the fact that additional contrast function H can be also non differentiable, or it cannot be expressed in closed form, but only some measures of it are available according to a specified value of the mixing coefficients $\mathbf{w}$.
According to the value of λ and to H it is possible to perform two different types of optimization; in fact, both constrained optimization, whose aim is to point out a specific feature of a single component or of a group of components, and multi-objective optimization, where a term is added to the measure of independence, can be treated using the proposed algorithm.
Suppose that the values of F and H in the proximity of a negentropy optimum (i.e. a solution of a classical ICA decomposition) are known: the optimization process can be performed in two ways:

i) H is weighed by λ such that λH is greater (i.e. at least two orders of magnitude) than J_G, and H is considered only until a threshold κ: the optimization becomes a *constrained optimization.*

ii) λH and J_G are weighed such that their values are similar. In this case, the optimization can be considered *multi-objective.*

These two cases need further investigation, and this is the topic of the next sections.

6.3.1 The Constrained Optimization case

Constrained optimization is employed when, in addition to the minimization (maximization) of a function f, some further requirements on the optimization landscape are required. A constrained optimization problem can be expressed as:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{w}) \\ \text{subject to :} \quad & h(\mathbf{w}) = \mathbf{0} \\ & g(\mathbf{w}) \leq \mathbf{0} \end{aligned} \tag{6.14}$$

where $f : \Re^m \rightarrow \Re$, $h : \Re^{m_1} \rightarrow \Re^{n_1}$ and $g : \Re^{m_2} \rightarrow \Re^{n_2}$. Function h is related to *equality* constraints, while g accounts for *inequality* ones.

It has to be noted that also FastICA can be seen as a constrained optimization, where f is an approximation of negentropy, and h is the constraint of unit norm.

If the three functions are continuous and differentiable, it is possible to give an elegant formulation of the problem by means of Lagrange multipliers [28], as seen in [108, 109, 85]. However there may be some cases where the function f or the constraints g and h are not differentiable, or cannot be defined in closed form but are defined in a procedural fashion. For ICA applications, it could be useful to account for prior information without the requirement to express this knowledge with a differentiable function or in closed form.

To overcome these limitations, the approach presented in section 6.3 is capable to deal with constrained optimization. In fact, it is possible to consider the additional contrast function as a *penalty* term and transform the constrained problem into an unconstrained one.

Penalty methods are particularly useful in looking for optimal solutions of a function f in a feasible region S. The problem then becomes:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{w}) \\ \text{with} \quad & \mathbf{w} \in S \end{aligned} \tag{6.15}$$

Suppose that is possible to define a continuous function $\psi : \Re^m \rightarrow \Re^+$ such that:

$$\psi(\mathbf{w}) \begin{cases} = 0, & \text{if } x \in S \\ > 0, & \text{if } x \notin S \end{cases} \tag{6.16}$$

then a new function to find the optima of f within the feasible region S is:

$$P(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda\psi(\mathbf{w}) \tag{6.17}$$

where λ often is called *penalty factor*. It will be shown that, as $\lambda \to +\infty$, the *unconstrained* optimization of f in $\Re^m$ leads to the optimum of it in S. To show this, consider a sequence of $\lambda_i \in \Re^+$ such that:

$$\lambda_{k+1} > \lambda_k, \qquad \lim_{k\to+\infty} \lambda_k = +\infty \tag{6.18}$$

and suppose that these two statements hold:

i) The problem of maximizing (6.17) has a solution, i.e. a point $\mathbf{w}^*$ exists such that

$$f(\mathbf{w}^*) = \min_{\mathbf{w}\in S} f(\mathbf{w}) \tag{6.19}$$

ii) $\forall\lambda_k > 0 \quad \exists\, \mathbf{w}_k \in \Re^m$ such that:

$$P(\mathbf{w}_k, \lambda_k) = \min_{\mathbf{w}\in\Re^m} P(\mathbf{w}, \lambda_k) \tag{6.20}$$

then the following theorem holds:

Theorem 10. *Suppose that both f and ψ are continuous functions, and that it holds* (6.16)*, and that conditions (i) e (ii) are satisfied. Let λ_k be a sequence that fulfills* (6.18)*; consider the sequence of $\mathbf{w}_k$ as defined in (ii). Then, $\forall k$ it holds:*

(a) $P(\mathbf{w}_k, \lambda_k) \leq f(\mathbf{w}^*)$;

(b) $\psi(\mathbf{w}_{k+1}) \leq \psi(\mathbf{w}_k)$;

(c) $f(\mathbf{w}_k + 1) \geq f(\mathbf{w}_k)$;

(d) $P(\mathbf{w}_{k+1}, \lambda_{k+1}) \geq P(\mathbf{w}_k, \lambda_k)$.

Proof. Let us consider the four points (a)–(d) separately:

(a): $\mathbf{w}_k$ is an *unconstrained* minimum for $P(\mathbf{w}_k, \lambda_k)$. Since $\psi(\mathbf{w}) = 0$ for $\mathbf{w} \in S$, it holds:

$$\begin{aligned} P(\mathbf{x}_k, \lambda_k) &= \min_{\mathbf{w}\in\Re^m} (f(\mathbf{w}) + \lambda_k\psi(\mathbf{w})) \leq \min_{\mathbf{w}\in S}(f(\mathbf{w}) + \lambda_k\psi(\mathbf{w})) = \\ &= \min_{\mathbf{w}\in S} f(\mathbf{w}) = f(\mathbf{w}^*). \end{aligned} \tag{6.21}$$

(b): Since $\mathbf{w}_k$ and $\mathbf{w}_{k+1}$ are the minima of $P(\mathbf{w}, \lambda_k)$ and $P(\mathbf{w}, \lambda_{k+1})$ respectively, it holds that:

$$\begin{aligned} f(\mathbf{w}_k) + \lambda_k \psi(\mathbf{w}_k) &\leq f(\mathbf{w}_{k+1}) + \lambda_k \psi(\mathbf{w}_{k+1}) && (6.22) \\ f(\mathbf{w}_{k+1}) + \lambda_{k+1} \psi(\mathbf{w}_{k+1}) &\leq f(\mathbf{w}_k) + \lambda_{k+1} \psi(\mathbf{w}_k), && (6.23) \end{aligned}$$

and summing equations (6.22) and (6.23), after some algebraic manipulations, we have:

$$(\lambda_{k+1} - \lambda_k)\psi(\mathbf{w}_{k+1}) \leq (\lambda_{k+1} - \lambda_k)\psi(\mathbf{w}_k), \tag{6.24}$$

and, as $\lambda_{k+1} - \lambda_k > 0$, it holds that $\psi(\mathbf{w}_{k+1}) \leq \psi(\mathbf{w}_k)$.

(c): from eq. (6.22) it holds that

$$f(\mathbf{w}_k) - f(\mathbf{w}_{k+1}) \leq \lambda_k(\psi(\mathbf{w}_{k+1}) - \psi(\mathbf{w}_k)), \tag{6.25}$$

therefore (c) comes from (b).

(d): from eq. (6.22), considering that $\lambda_{k+1} > \lambda_k$, it holds:

$$f(\mathbf{w}_k) + \lambda_k \psi(\mathbf{w}_k) \leq f(\mathbf{w}_{k+1}) + \lambda_k \psi(\mathbf{w}_{k+1}) \leq f(\mathbf{w}_{k+1}) + \lambda_{k+1} \psi(\mathbf{w}_{k+1}) \tag{6.26}$$

therefore (d) is true.

□

By means of Theorem 10 that is related to the property of monotonicity, it is possible to prove the convergence of the penalty method:

Theorem 11. *Suppose that f and ψ are continuous functions, that psi fulfills condition* (6.16)*, and that hypotheses (i), (ii). Let $\{\lambda_k\}$ be a sequence of positive numbers satisfying eq.* (6.17)*, and $\{\mathbf{w}_k\}$ be the sequence of unconstrained minimum points of function $P(\mathbf{w}, \lambda_k)$. Suppose that all points $\mathbf{w}_k$ stay in a compact set $D \subset \Re^m$. Then:*

(c_1) $\lim_{k\to\infty} \psi(\mathbf{w}_k) = 0$

(c_2) $\lim_{k\to\infty} f(\mathbf{w}_k) = f(\mathbf{w}^*)$, $\lim_{k\to\infty} P(\mathbf{w}_k, \lambda_k) = f(\mathbf{w}^*)$*;*

(c_3) Every accumulation point of $\{\mathbf{w}_k\}$ is an optimum of problem (6.15)*;*

(c_4) $\lim_{k\to\infty} \lambda_k \psi(\mathbf{w}_k) = 0$

Proof. From points of (a) and (c) of Theorem 10, it comes:

$$f(\mathbf{w}^*) \geq P(\mathbf{w}_k, \lambda_k) = f(\mathbf{w}_k) + \lambda_k \psi(\mathbf{w}_k) \geq f(\mathbf{w}_1) + \lambda_k \psi(\mathbf{w}_k), \tag{6.27}$$

therefore, considering the superior limit:

$$f(\mathbf{w}^*) - f(\mathbf{w}_1) \geq \lim_{k\to\infty} \sup \lambda_k \psi(\mathbf{w}_k). \tag{6.28}$$

Now, since $\lambda \to \infty$ and $\psi \geq 0$, it holds that $\limsup \psi(\mathbf{w}_k) = 0$ (if this were not true, then the second term of eq. (6.28) would tend to $+\infty$ for some subsequence, and this would be an absurd). As $\psi \geq 0$, it holds:

$$\lim_{k\to\infty} \psi(\mathbf{w}_k) = 0 \tag{6.29}$$

therefore (c_1) is true.
From hypotheses D is a compact set and $\mathbf{w}_k \in D\ \forall k$, therefore the sequence $\{\mathbf{w}_k\}$ has accumulation points. Suppose that $\overline{x}$ is an accumulation point, i.e. a subsequence exists such that:

$$\lim_{k\to\infty, k\in K} \mathbf{w}_k = \overline{\mathbf{w}} \tag{6.30}$$

From (c_1) and since f and ψ are continuous functions, it will be $\psi(\overline{\mathbf{w}}) = 0$, therefore $\overline{\mathbf{w}} \in S$. From Theorem 10, sequences $\{f(\mathbf{w}_k)\}$ and $\{P(\mathbf{w}_k, \lambda_k)\}$ are monotonic non decreasing and upper bounded (in fact $f(\mathbf{w}^*) \geq P(_k, \lambda_k) \geq f(\mathbf{w}_k)$), therefore they have a limit. From (a) of Theorem 10, considering the superior limit for $k \in K$, it holds:

$$\begin{aligned} f(\mathbf{w}^*) &\geq \limsup_{k\to\infty, k\in K} P(\mathbf{w}_k, \lambda_k) = \lim_{k\to\infty} P(\mathbf{w}_k, \lambda_k) = \\ &= \lim_{k\to\infty} f(\mathbf{w}_k) + \lim_{k\to\infty} \lambda_k \psi(\mathbf{w}_k) = \\ &= f(\overline{\mathbf{w}}) + \lim_{k\to\infty} \lambda_k \psi(\mathbf{w}_k) \geq f(\overline{\mathbf{w}}) \geq f(\mathbf{w}^*) \end{aligned} \tag{6.31}$$

where the last inequality comes from the fact that $\overline{\mathbf{w}}$ is a feasible point. From inequalities in (6.31), proposition (c_2) is verified. Again from (6.31) come the fact that $f(\mathbf{w}^*) = f(\overline{\mathbf{w}})$, $\overline{\mathbf{w}}$ is an optimal solution, thus (c_3) and (c_4) are verified. □

Theorems 10 and 11 guarantee that the penalty method, under some conditions, leads to a solution to the problem of optimizing f in a region S defined by means of a penalty function ψ. For what concerns the ICA approach developed, penalty methods are particularly useful, because it is not required for the constrains to be differentiable, a case where Lagrange multipliers optimization, that requires gradients, is not able to perform optimization.

If we define U_{unc} as the set of the solutions of a classical (unconstrained) ICA problem, the solution set of the constrained problem will be $(U_{unc} \cap S) \cup U_c$, where S deontes the feasible region and U_c is the set of solutions of the constrained problem that are at the edges of the feasible region.

Therefore, in the applications where there is some strong requirement on an independent component, the constrained approach for the contrast function can be useful, as long as the feasible set S is well defined by the additional function H.

6.3.2 The Multi-objective case

The second way to modify ICA is the multi-objective case. In this way, the additional contrast function is not weighted much more than independence, but rather it is considered together with it to reach the optimum configuration.

It is clear that, in this case, that the new optima of the problem will not be a subset of the classical ICA solution: consider a solution of an ICA problem, where $\mathbf{W}$ is the estimate of the unmixing matrix, estimated by maximizing the negentropy of the estimated components $\mathbf{y} = \mathbf{Wx}$. If we consider a new contrast function F defined as in (6.13), the solution of the optimization process will be a new unmixing matrix $\mathbf{W}_1$. It is evident that the negentropy of $\mathbf{y}$ will be greater or equal to the one of $\mathbf{y}_1 = \mathbf{W}_1\mathbf{x}$, by definition, meaning that the solution in the multi-objective approach is less “independent” than the classical one. The question why such an approach could be an improvement to the extraction is not evident in most cases. However, especially in real world applications, most of the times the measure of independence may be “deteriorated” by the presence of noise, or the “interesting” signals themselves may not be “completely” independent. Consider the example in section 6.1, and in particular Figure 6.2: in that case, the three spatial maps related to a simulated fMRI experiment have the same histogram. If no noise is present in the environment, the sources will be retrieved correctly. However, if a source has a poor Signal to Noise Ratio (SNR), its contrast value cannot be fully reliable, and it may happen that the configuration that maximizes negentropy is not the optimal one in terms of interpretation of the results. The starting point of this approach is that in real world problems, there is much more than independence that one can consider to extract the sources. Consider, for instance, the analysis of fMRI time series (Chapter 3): it is known that independent spatial maps, to be considered an estimate of spatiotemporal activity in the brain, must entail some kind of regularities in space and time. This is absolutely not pointed out by classical ICA, and there is no reason to discard this information if it may make the source recovery easier and allow extraction of more physically plausible independent components. In Chapter 7 an example on a simulated fMRI dataset at different levels of SNR will be shown to support such arguments.

6.4 Simulated Annealing

6.4.1 Introduction

Simulated Annealing (SA) is an optimization procedure inspired by statistical mechanics, particularly suitable in cases where a deterministic solution or a gradient optimization cannot be used (like in combinatorial optimization problems, or when no closed form for the contrast function is available). The basic idea behind the use of simulated annealing is to simulate the behavior of condensed matter at low temperature: consider the problem of growing silicon in the form of highly ordered, defect free crystals for use in semiconductor manufacturing. This means coercing a solid in into a low energy state, that (usually) means a highly ordered state. To accomplish this, the material is *annealed*: it is heated to a temperature that allows many atomic rearrangements, and then it is cooled *slowly*, until the material freezes into a good crystal. If the material is not heated to a suitable temperature, or it is not cooled slowly, the crystal formation process may present a great number of imperfections. The aim of the slow cooling is to allow statistical equilibrium within a single state associated with the temperature.

Simulated Annealing is inspired to this considerations, and performs a stochastic global optimization of a (cost, or contrast) function, being quite general, in the sense that different kind of functions can be optimized without changing the optimization procedure. To perform *global* optimization by means of local search Ω, several solutions have been proposed. As the optimization may be trapped in local maxima (or minima) while performing local search, a so called *multistart* procedure is usually implemented, where the algorithm starts from different points in order to find the global optimum. Of course such a procedure could be extremely demanding, since one has to explore many starting points, and the algorithm will be trapped several times in the same optimum. Theoretically, a local search procedure Ω should not be called more than once for each region of attraction (the region of attraction of a solution $\overline{x}_k$ w.r.t. a local search procedure Ω is defined as the set of points from which the local search Ω will converge to $\overline{x}_k$).

Simulated Annealing, as an alternative to multistart algorithms, performs a random search technique, where the algorithm avoids getting trapped in local optima by accepting not only all movements that lead to an increase in the contrast function value but also, with a finite probability, movements leading to a deterioration . This probabilistic acceptance criterion is designed to avoid being trapped in a local optimum, and leads to a global solution as the control parameter (related to the acceptance probability) reaches values such that this probability tends to zero.

6.4.2 Simulated Annealing optimization

While dealing with SA, it is common to illustrate the search that is performed at a *fixed* temperature, called *Metropolis Algorithm*, which consists of a number of simple principles. As mentioned before, the search is random (in the sense that there is no favorite direction of perturbation from the original state), but with the probabilistic acceptance of “wrong” solutions.

Consider a state s from which the algorithm starts, then consider a perturbation s'. If the contrast function value $f(s')$ is greater than $f(s)$ it is accepted, otherwise it is accepted with a probability $p_T(\Delta) = \min(1, e^{-\Delta/T})$, where $\Delta = f(s') - f(s)$ and T, often called temperature, is the control parameter that determines the probability of transition. Consider the Metropolis procedure applied to the case of maximizing function f over a state space X. The procedure can be summarized by this pseudocode:

```
Generate some random initial configuration $s$
REPEAT
    Determine a neighbor state $s'$
    Evaluate $\Delta = f(s) - f(s')$
    Evaluate $p_T(\Delta) = \min(1, e^{-\Delta/T})$
    IF $random[0,1] \leq p_T(\Delta)$ THEN move to state $s'$
```

```
7. UNTIL FALSE
8. END
```

If the new state s' is such that leads to an improvement in contrast function value ($\Delta < 0$), then the new $p_T(\Delta) = 1$, therefore the new state is accepted. On the contrary, if $\Delta \geq 0$, meaning that the new state is lower in terms of contrast function, the new state is accepted with probability related to the temperature. Considering the set of states $\Phi_s \in X$ such that $\phi \in \Phi$ can be reached from s in one move (perturbation), then each move must be reversible ($\phi \in \Phi_s \Rightarrow s \in \Phi_\phi$). Consider a discrete problem and suppose that the probability of each move is $1/\omega$, following the Metropolis procedure means moving through states according to a Markov process with the following transition probability:

$$\Phi_T(s'|s) = \left[\begin{array}{ll} \frac{1}{\omega} p_t(f(s') - f(s)) & \text{if} s' \in \Phi_s \\ 1 - \sum\limits_{\phi \in \Phi_s} \frac{1}{\omega} p_t(f(\phi) - f(s)) & \text{if} s' = s \\ 0 & \text{otherwise} \end{array}\right. \tag{6.32}$$

If $T > 0$ the process is *irreducible*, meaning that there is a nonzero probability that state ϕ will be reached from state s, with $\phi, s \in X$. The stationary distribution of the process π_T is:

$$\pi_T(s) = \frac{e^{(-f(s)/T)}}{\sum\limits_{\phi \in \Phi} e^{(-f(\phi)/T)}}, \quad \forall s \in \Phi \tag{6.33}$$

Simulated annealing starts from the Metropolis algorithm and introduces the concept of temperature decrease to find the global optimum. In fact, as the temperature decreases, fewer and fewer solutions will be accepted, and the optimization will be “trapped” in the global optimum. With the same notation as before, it is possible to express the optimization in terms of a pseudo-code:

```
1. Generate a random initial state $s = s_0$
2. Consider a starting temperature $T = T_0$
3. WHILE (stopping criterion is not satisfied) DO
4.    WHILE (required number of states has not been generated) DO
```

```
    Generate a new state $s'$ by perturbing present state $s$
    Evaluate $\Delta = f(s) - f(s')$
    IF $(\Delta \leq 0)$ THEN
      Move to state $s'$
    ELSE
      Generate a random variable $\alpha \in [0,1]$
      IF $\alpha \leq e^{(-\Delta/T)}$ THEN move to state $s'$
    END
  END
  Update $T$ (Decrement)
END
```

Even if the procedure, as seen in the pseudo-code, is rather simple, nonetheless many issues are still open: both the generation of states and all the criteria in the code are intentionally not defined, as they may change from application to application. Fortunately, some guidelines can be given in order to setup a SA optimization. Before showing the most used criteria, a convergence analysis will be given.

It has been shown that the Metropolis algorithm moves through states according to a Markov *homogeneous* process. The homogeneity of a Markov process means that, considering the conditional probability at step k

$$P_{S_i S_j}(k-1,k) = P\left(S(k) = S_j | S(k-1) = S_i\right), \tag{6.34}$$

this probability does not depend on k; if is there any dependence, the process is described by an *inhomogeneous* Markov process.

The convergence of Simulated Annealing can be formulated in terms of:

- **a homogeneous algorithm**: the algorithm can be described as a sequence of homogeneous Markov chains, where each chain is generated at a fixed value of Temperature, that is decreased in between subsequent chains.

- **an inhomogeneous algorithm**: the algorithm in this case is described by a *single* inhomogeneous Markov chain, where the value of Temperature is decreased in between subsequent transitions.

Convergence issues in the homogeneous algorithm are related to the probabilistic acceptance rule, while in the inhomogeneous case, the main convergence condition is on the temperature decrease rate.
For what concerns the homogeneous algorithm, the convergence is proved when there exists a stationary distribution, defined as the limit of the probability of being in a state as the number of iterations tends to infinity, and that the limit of this stationary distribution is a uniform distribution on the set of globally optimal configurations. In [1] the main theorems regarding convergence of homogeneous Markov chains are shown. It is possible to prove, by means of those theorems, that the probabilistic acceptance criterion employed in Simulated Annealing allows to define stationary distributions whose limits is the set of global optima of the problem. Moving to the inhomogeneous case, assuming that temperature is changed at each transition, it is possible to prove ([1]) that asymptotic converge holds if the temperature decreases as:

$$T_n = \frac{a}{\log(n + k_0 + 1)}, \quad \text{with} \quad k_0 \geq 1. \tag{6.35}$$

From the theory of Markov chains it is possible to prove asymptotic convergence of the Simulated Annealing optimization. However, for an efficient implementation, the bounds imposed by Markov chains theorems are far too time consuming, therefore some suitable choices have to made in order to guarantee limited amount of time and at the same time the ability to reach a satisfactory solution.

Choice of initial Temperature

The choice of initial Temperature is a crucial step in a SA algorithm. In fact, if the starting temperature is too low, the system will be trapped in local optima, since the probability acceptance criterion will tend to reject all sub-optimal moves. On the contrary, a starting temperature too high makes the algorithm excessively slow, consuming much more time than necessary. Moreover, since the temperature is present in the optimization scheme only in combination with the difference of cost function values in the acceptance criterion, its values must be somehow related to the contrast function: in general, starting temperature should be at least six times higher than the mean of Δf among all moves. A suitable and more general criterion to set up the temperature is the following:

I. Begin with a random T.

II. Consider a number of iterations of the Metropolis procedure at temperature T, recording the number of accepted moves A and the number of rejected ones R.

III. If $A/(A+R) < 0.8$ then raise the temperature (*e.g.* $T_1 = 2T$), and repeat the process until the system is warm enough.

To improve the previous procedure, one may decide to avoid the ratio $A/(A+R)$ being too high (for instance, higher that 0.95), by decreasing the starting temperature for those ratios, thus avoiding the system being too "warm".

Choice of the Temperature decrease schedule

The temperature decrease schedule is another crucial aspect of a SA optimization. If the cooling scheme is too fast, it may lead to sub-optimal solutions, while if it is too slow it may take too much time to perform optimization. It has been proved that ([1]) a temperature schedule of the form:

$$T_n = \frac{c}{\log(n + k_0 + 1)} \tag{6.36}$$

for large values of n and for sufficiently large values of c, guarantees convergence. However, the scheme is far too slow to be applied in most of the problems. A more efficient procedure is the exponential decrease rate, where

$$T_n = T_{n-1}\alpha, \quad \text{with } \alpha < 1 \tag{6.37}$$

where the cooling rate is controlled by α, that is usually set between 0.5 and 0.99. This scheme has the advantage of moving relatively fast at high temperature, while focusing much more on low temperatures, near the halting of the optimization.

Conditions for algorithm termination

Various conditions for algorithm termination can be found in literature, and it was found that the value of temperature for which no further improvements can be made is:

$$T_f = \frac{E_m - E_{m'}}{\ln \nu} \tag{6.38}$$

where $E_(m)$ is the absolute maximum value for the contrast function (often is set to some predetermined value that depends on the application) and $E_{m'}$ is the next smallest value of the contrast function (according to E_m), and ν is the number of moves that takes to get from $E_{m'}$ to E_m. The temperature determined is such a way represents the worst case scenario, and other criteria can be used to increase the speed of the algorithm. Another employed criterion is to stop the algorithm when no new solution has been accepted after four consecutive temperature decreases.

Choice of number of states at each Temperature

A criterion to evaluate the number of iterations needed at each temperature is to set it at a fixed temperature as:

$$Y(T) = e^{(f_{max} - f_{min})/T} \tag{6.39}$$

where f_{max} and f_{min} are the maximum and the minimum values found so far for the contrast function f. However, $Y(T)$ gets too large as T decreases, thus requiring to upper bound this number. Another criterion is to consider the number of states at each iteration such that the ratio between accepted and rejected moves is 1:10 (but also in this case, as the temperature decreases, the number of iterations may grow too much).

6.5 Simulated Annealing for Independent Component Analysis

A Simulated Annealing optimization has been implemented for Independent Component Analysis. The contrast function, whose global optimum is the solution of the problem, is F, as defined in (6.13). The extraction can be done in two ways (like FastICA): extracting one component at a time or extracting all the sources together. Therefore, according to the extraction procedure used, the contrast function is:

- **Deflation case**: One component is extracted at a time. Therefore, a single mixing vector $\mathbf{w}$ is considered during optimization, with the following contrast function:
$$F = J_G(\mathbf{w}) + \lambda H(\mathbf{w}) \tag{6.40}$$

- **Symmetric case**: In this case, all the components are considered together, and a matrix $\mathbf{W}$ is optimized to produce the maximum of the following contrast function:
$$F(\mathbf{W}) = \sum_i J_G(\mathbf{w}_i) + \sum_i \lambda_i H_i(\mathbf{w}_i) \tag{6.41}$$
where $\mathbf{w}_i$ denotes a column of $\mathbf{W}$ and the functions H_i are computed for each estimated component.

The differences for the two extraction types are in the perturbation process, in the decorrelation procedure and in the contrast function evaluation, while the main iterations and the optimization scheme remains the same.

The main part of the optimization is the Metropolis algorithm (see 6.4.2), with the acceptance criterion related to Temperature. To show how the optimization process is carried out, it is useful to point out the main aspects of Simulated Annealing setup, that will be discussed in detail:

- Perturbation of the state

- Annealing schedule

- Stopping criterion

Perturbation, i.e. the generation of a new state from an existing one, will be examined according to the strategy employed:

- **Deflation** The new state $\mathbf{w}_{k+1}$ is generated from the previous state $\mathbf{w}_k$ according to the following rule:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + (k_0 + k_1/l)\delta_m \tag{6.42}$$

  where k_0 and k_1 are some suitable constants (decided according to the dimension of the problem and that may be changed if the optimization landscape is too much or too little rugged), l is related to the temperature and to the number of accepted and rejected states (its purpose is to reduce the perturbation step if the system is reaching the optimum) and δ_m is an m-dimensional vector whose components are uniformly distributed in $[-1, 1]$. Usually, $\mathbf{w}_{k+1}$ is not in the feasible region, since it must be $\|\mathbf{w}\|^2 = 1$; moreover, if other components have been already extracted, decorrelation from the subspace they generate should be performed. Therefore, normalization and decorrelation steps are performed before the evaluation of the contrast function in the new state.

- **Symmetric** The new state $\mathbf{W}_{k+1}$ is generated in a way similar to (6.42):

$$\mathbf{W}_{k+1} = \mathbf{W}_k + (k_0 + k_1/l)\delta_{m\times m} \tag{6.43}$$

  where k_0, k_1 and l are defined as before, and $\delta_{m\times m}$ is an $m \times m$ matrix whose components are uniformly distributed over $[-1, 1]$. Also in the symmetric case, before evaluating the value of the contrast function in the new state, symmetric decorrelation must be performed.

The annealing schedule is related to the number of states generated at a fixed temperature and to the temperature decrease rate. The following choices were made for the present implementation:

- **Temperature Decrease Rate**: A geometric decrease rate was implemented as follows:

$$T_{k+1} = T_k\alpha = T_0\alpha^{k+1} \tag{6.44}$$

  where α was usually between 0.7 and 0.99.

- **States explored**: An adaptive rule has been implemented to run across the states at a fixed temperature. The required number of states was generated if the ratio between the number of accepted states transitions and the number of rejected transitions was at least 1 : 10. For low temperatures, close to the optimum, this requirement would tend to make the process extremely time-consuming, therefore an upper bound related to the dimension of the problem was introduced.

The choice of the initial temperature has been implemented as in 6.4.2, by evaluating the ratio between the number of accepted moves and rejected ones, and setting the temperature such that this ratio is greater than 0.8.

For what concerns the stopping criterion, the states at the end of iterations at two consecutive temperatures were compared: if those two states were similar within a certain range, the algorithm stopped. For the two types of algorithm, the stopping criterion was therefore:

- **Deflation**: consider the two states $\mathbf{w}_{T_{k+1}}$ and $\mathbf{w}_{T_k}$. The algorithm is terminated if
$$\left\|\mathbf{w}_{T_{k+1}} - \mathbf{w}_{T_k}\right\| \leq \varepsilon \quad \text{or} \quad \left\|\mathbf{w}_{T_{k+1}} + \mathbf{w}_{T_k}\right\| \leq \varepsilon \tag{6.45}$$
since both $\mathbf{w}$ and $-\mathbf{w}$ are related to the same component (even if it is unlikely that the algorithm goes from $\mathbf{w}$ to $-\mathbf{w}$ during an iteration of the optimization process). Usually $\varepsilon = 10^{-4}$.

- **Symmetric**: Consider the two states $\mathbf{W}_{T_{k+1}}$ and $\mathbf{W}_{T_k}$. The sum q of the absolute value of the elements of the main diagonal of $\mathbf{W}_{T_{k+1}}\mathbf{W}_{T_k}^T$ is evaluated, and the algorithm terminates if:
$$(1-q) \leq \varepsilon \tag{6.46}$$
where ε is defined as before and usually is 10^{-4}. This criterion is the same implemented in the FastICA algorithm.

The implementation of the Simulated Annealing procedure has been shown. Now, according to the problem faced, and therefore to the approach employed, further details will be given on the Independent Components extractions.

6.5.1 Multi-objective Case

Since the multi-objective approach takes its moves from the need of improving the quality and the physical plausibility of the recovered sources, an exploratory extraction with a classical ICA algorithm (or with the new algorithm with $\lambda = 0$, as well) is performed, in order to evaluate the values of the maxima of negentropy, and let IC_i, with $i = 1, 2, \dots, m$ be the estimated components. Then the following procedure is employed:

1. Components are ordered according to Negentropy (if the algorithm employed is not global, components are not ordered).
2. For each component IC_k, the following steps are performed:
   (a) the values of negentropy J_G and of H are evaluated for estimated component.
   (b) if $H(\mathrm{IC}_k)$ is not satisfactory (i.e., the estimated component do not fulfill the requirements that are known *a priori*), then λ_k is set to a suitable value, related to the values of Negentropy $J_G(\mathrm{IC}_k)$ and of $H(\mathrm{IC}_k)$.
3. Optimization is run with Simulated Annealing with the contrast function F, with the deflation or symmetric approaches:
   - **Deflation**: for the i-th component, the contrast function becomes

$$F(\mathbf{w}) = J_G(\mathbf{w}) + \lambda_i H(\mathbf{w}) \tag{6.47}$$

   - **Symmetric**: in this case, the contrast function evaluated is:

$$F(\mathbf{W}) = \sum_i J_G(\mathbf{w}_i) + \sum_i \lambda_i H(\mathbf{w}_i) \tag{6.48}$$

For the problem dealt with in Chapter 7, λ_i was set such that $H(\mathrm{IC}_i)$ weighed approximately 10 times less than $J_G(\mathrm{IC}_i)$ in the proximity of a maximum of negentropy.

6.5.2 Constrained Case

The constrained case is different from the previous one, since a preliminary extraction with classical ICA algorithms is not needed. In fact, it has been proved in 6.3.1 that, if the weight of the penalty term tends to infinity, then the optimization of the function $J_G + \lambda H$ leads to the constrained optimum defined by H.

Additional contrast function H accounts for equality or inequality constraints, and the formulation presented in 6.3.1 need to be slightly modified since the problem is a maximization and not a minimization (however, the same results would have been achieved by inverting the sign of negentropy and looking for the minima). In this case the problem becomes, for the deflation approach:

$$\begin{aligned} \text{Maximize} \quad & J_G(\mathbf{w}_i) \\ \text{subject to} \quad & H_i(\mathbf{w}_i) \geq 0 \end{aligned} \tag{6.49}$$

while for the symmetric approach the problem is:

$$\begin{aligned} \text{Maximize} \quad & \sum_i J_G(\mathbf{w}_i) \\ \text{subject to} \quad & \mathbf{H}(\mathbf{w}_i) \geq \mathbf{0} \end{aligned} \tag{6.50}$$

where $\mathbf{H}(\mathbf{w}_i) = (H_1(\mathbf{w}_1), H_2(\mathbf{w}_2), \ldots, H_m(\mathbf{w}_m))$ and $\mathbf{0}$ is an m-dimensional vector of zeros.

To perform optimization, an initial guess on λ_0 is made. Usually, since the range of values for the additional contrast function H is known in advance, a suitable choice for the initial λ_0 is such that $\lambda_0 H_m ax$ is 10^2-10^3 times the maximum value of J. The initial choice does not affect the final results of the independent component extraction, but a non optimal choice may lead to an unnecessary amount of computations. Once the initial value of λ_0 is decided, the optimization is performed following these steps:

1. λ is set to the initial value λ_0.
2. Optimization with Simulated Annealing is performed with the contrast function $F = J_G + \lambda H$.
3. Results are evaluated by checking if the constraints are satisfied (i.e. if the value of H is positive for the solutions).
4. If the constraints are satisfied, then the solution has been reached, otherwise λ is increased (for instance, $\lambda_{new} = 10\lambda$), and the procedure goes back to step 2

Theorems 10 and 11 guarantee that, if λ is large enough, and in the worst case tends to infinity, the solution of the unconstrained problem with a penalty function is equivalent to the constrained solution. Therefore, the procedure shown converges to the global constrained optimum. It has to be noted that it may be troublesome to set too high the starting value of λ (with respect to the suggestion of three orders of magnitude), since the optimization process may encounter problems due to computer precision, because the values of the penalty function and of the negentropy are too different. Moreover, to avoid large variations for negentropy estimate J_G, it may be better to choose the nonlinear function G as in (2.119) and (2.120), i.e. $G_1(y) = \log\cosh(y)$ or $G_2(y) = -\exp(-y^2/2)$, instead of kurtosis.

Chapter 7

Use of prior information in ICA for fMRI data

7.1 Introduction

In the following chapter the methodology presented in Chapter 6 will be applied in both versions to fMRI data. In section 7.2 some examples of constrained optimization will be provided, while in section 7.3 a study on how to modify ICA with some "loose" information on the spatio-temporal nature of the sources will be shown in detail.

The two approaches reflect basically the kind of prior information one has on one or more sources. If this prior information is precise (i.e. it defines univocally an independent source), then the constrained optimization case (section 6.5.2) is the most suitable . In fact, by means of the inequality (or equality) constraints defined in the penalty function $H(\mathbf{w})$, it is possible to:

- Extract the "target" source first. This could be extremely useful when the number of independent components is high, i.e. when the use of a simulated annealing optimization on only one component is more efficient than the extraction of *all* the component with a faster algorithm *and* subsequently the selection *a posteriori* of the interesting component. This is, for instance, the case of temporal ICA, where the number of independent components (without dimension reduction) equals the number of the spatial points, that may be even a hundred thousand for a whole volume.

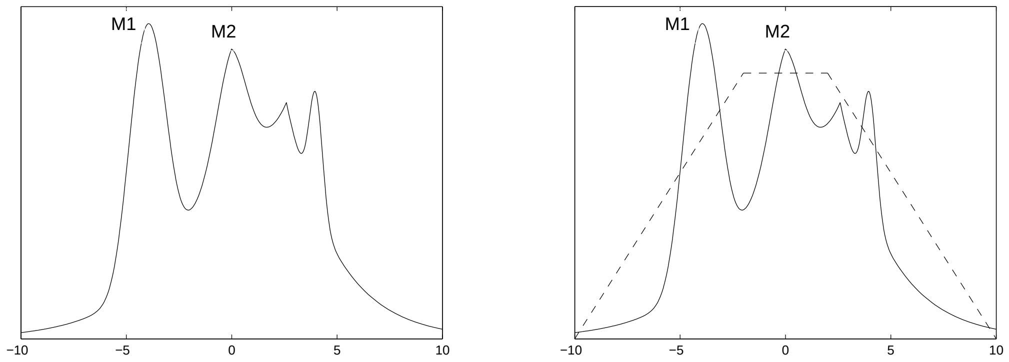


Figure 7.1: Example on constrained optimization in a mono-dimensional case. Left panel: values of contrast function. Right panel: contrast function (solid line) and penalty function (dashed line).

- If the presence of noise or if the "meaningful" signals are not completely independent, it is possible to recover them by defining the feasible region S that is known *a priori* to include the interesting source.

To clarify this aspect, consider the example in Figure 7.1: in the left panel a simple mono-dimensional optimization landscape is depicted. Since there are several local maxima, a local search optimization technique could reach one of them according to the starting point. A global optimization algorithm (like Simulated Annealing), instead, will reach $M1$ value for any starting point. Therefore local maximum $M2$ can be reached for some starting points with a local seach and will *never* be a solution for a global search technique.

However, due to noise presence or due to not full independence of the sources, it may happen that the "meaningful" maximum is $M2$ rather than $M1$, and at the same time that prior information on some of its features (which correspond somehow to position in the optimization landscape) is available. Suppose that this prior information can be expressed as a function H, i.e. it is known in advance that the "target" maximum lies in the flat region of H depicted in the left panel of Figure 7.1 (dashed line). Therefore it is possible to reach local optimum $M2$ by performing a constrained optimization on the contrast function with the inequality constraint defined by values of function H. In the following section this considerations will be employed to perform ICA analysis of fMRI datasets, to enhance performances in accuracy and to avoid post-selection of the components.

There are some situations, however, where selecting the proper maxima could not be feasible. In fact, if the prior knowledge is general enough (like in the case of spatio-temporal regularities), performing constrained optimization could be of no help in recovering the meaningful sources, if this information is on all the sources. In this case is preferable to "perturb" the local optima rather than selecting them, to have more plausible solutions. Section 7.3 will focus on this aspect.
Another important issue is the extraction order using a deflation approach. In fact, local search algorithms (like FastICA or Infomax) *do not* extract components in a precise order, since the output of the optimization procedures depends on the starting point. Consider now FastICA algorithm in deflation approach. Due to the fact that optimization is performed on a component at a time, this solution is particularly appealing for implementation issues, since only one vector of the unmixing matrix is involved for each iteration and therefore data size reduces considerably and so does memory occupation. The ordering of the components is not fixed (although it can be observed that some components have wider attraction regions than others, and therefore are more likely to be extracted first), so it is not possible to predict in advance, after the extraction of a subset of the components, if the "target" component has been recovered. Global optimization techniques do not suffer from this drawback: the ordering of the components is *always* the same, regardless of the starting point. This is a particularly appealing aspect while dealing with fMRI components, since independent sources of brain activity are in most cases super-Gaussian ([118]). It has been shown in [56] that it is not possible to detect "interesting" components according to their ranking in a mono-dimensional space (i.e. based only on non-Gaussianity of the spatial maps or to root mean square contribution of the IC to the whole dataset). However, since a task-related independent source is expected to have a high value of non-Gaussianity (even if it is not among the first ranked, that usually may be related to artifacts or to non-task-related baseline activity), it is not necessary to extract all the sources, but a reduced subset can be considered. In the constrained case, moreover, only the components pointed out by prior knowledge are extracted in a deflaction fashion, since the penalty term ensures that their contrast function values will be the highest.

7.2 Constrained optimization in fMRI data analysis

In section 3.3.6 it has been shown how to perform ICA analysis of an fMRI dataset, together with the advantages and drawback of the employment of such a technique. Consider an observed fMRI dataset $\mathbf{X} \in \Re^{m \times n}$ where each *row* of $\mathbf{X}$ contains a whole volume (or the part of volume where the analysis is performed), and each *column* contains the time course of a given voxel (i.e. volume element, in analogy to pixel). Spatial ICA looks for the independent components $\mathbf{S}$ of the dataset such that

$$\mathbf{X} = \mathbf{AS} \tag{7.1}$$

where $\mathbf{S} \in \Re^{m_1 \times n}$ is the matrix whose rows are the independent spatial maps and $\mathbf{A} \in \Re^{m \times m_1}$ is the mixing matrix. When an independent spatial map is interpreted, it is possible to consider both its spatial and temporal pattern. Consider Figure 7.2: on the left panel an independent component of a single slice analysis of an experiment is depicted, together with the original data (in gray scale) and with the relative time course (independent components display is performed in general, contrary to this simple case, with specific software, that superimposes on the anatomical scan the activations obtained by thresholding the spatial map, to better relate the activations to the corresponding brain area).

Spatial maps and time courses are interpreted in a different way from the way Statistical Parametric Maps (obtained with an hypothesis driven analysis) are usually considered. In fact, *all* the points of an independent component have the *same* time course, even if usually extreme values of the distribution are considered, since the majority of values is zero (On the other hand, this time course is realistic, in that it is extracted from the data, while the SPM time course is imposed a priori). No meaning can be given to the absolute values of the maps or of the mixing coefficients, due to the double indeterminacy in the ICA model (see section 2.4), but they can be considered together to evaluate the Root Mean Square (RMS) contribution they give to the whole dataset (like in PCA). Information in independent components lies in both their spatial maps and their temporal time courses. In fact, while the former provides information about the spatial pattern of an activity, the latter indicates how the spatial map "evolves" in time to form the dataset. Usually expert's evaluation is based on both spatial and temporal pattern. A simple criterion, pro-

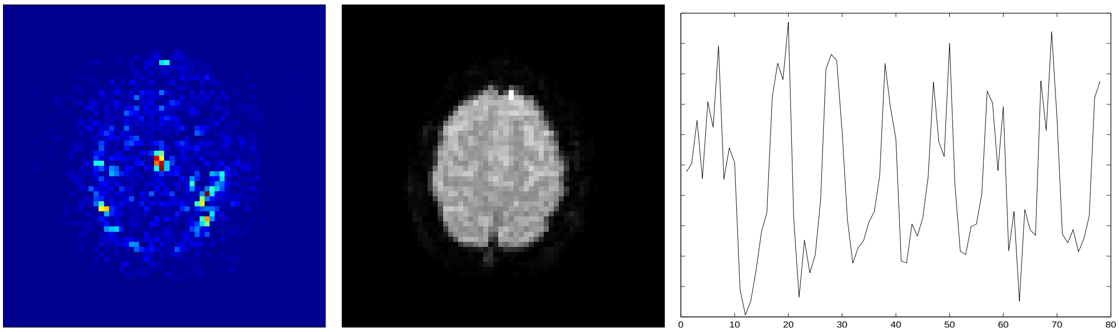


Figure 7.2: Example of spatial ICA on a fMRI slice

posed in [119], to point out task-related component(s) is to evaluate the correlation of its time-course with the paradigm of the experiment (properly convolved with an approximation of the Hemodynamic Response Function (HRF), to account for delays in BOLD signal, as discussed in section 3.3.1).

Since both spatial independent maps and temporal time courses (that are respectively the independent components and the mixing coefficients of (7.1)) carry information on the spatio-temporal activity pattern, our investigation has been focused on both aspects. In fact, additional contrast function H defined in (6.13), together with Simulated Annealing optimization, allows including very general kinds of constraints into the algorithm. Some applications of this technique with *specific* prior information on data will be discussed, and compared with existing techniques shown in 6.2. As discussed in the introduction, it is possible to recover a source first by pointing out some peculiar feature, and it is also possible to "force" a component to lie in a subspace S defined by a suitable penalty function, that is not reached by a classical extraction (like shown in Figure 7.1). The first of the two cases will be faced in section 7.2.1, while the second in section 7.2.2.
Contrast function H, that defines the constraint, is usually in the form of Figure 7.1, right panel, dashed line. In other words, if the inequality constraint $g(\mathbf{w}) > \kappa$ holds, the following function is employed in the optimization:

$$H(\mathbf{w}) = \begin{cases} g(\mathbf{w}) & \text{if} \quad g(\mathbf{w}) < \kappa \\ \kappa & \text{if} \quad g(\mathbf{w}) \geq \kappa \end{cases} \tag{7.2}$$

Such choice (rather than a function that is non-zero only when $g(\mathbf{w}) \geq \kappa$) comes from a trade-off between the need to reach the global optimum and to keep the amount of time to perform *global optimization* somehow limited. In fact, the feasible set, i.e. the set where $g(\mathbf{w}) \geq \kappa$ is flat, may be narrow, if compared with the overall dataset. Therefore, setting a sharp contrast function that takes non-zero values in region S and zero elsewhere may need more iterations to span the whole dataset, while a linear ramp tends to "guide"(at least in the last part of the optimization) the optimization algorithm to the desired maximum. Of course, a sharp function may be preferable in terms of effectiveness of the constraint, since a local maximum, lying just outside the edge, may be the global maximum of the contrast function, even if it is not in the feasible set. However, the procedure set up defined in 6.5.2, by increasing the penalty factor λ, guarantees that in such a way it is possible to select the maximum within the feasible set.

7.2.1 Recovering a source first

It is possible to extract a source first by exploiting some prior information on it. This information can be related to spatial maps on to temporal source, both in spatial and temporal ICA. Since there is no ordering of the components extracted with classical ICA algorithms, it is possible to extract a source first rather than extracting all the independent components and *post-selecting* the interesting one, if it is possible to incorporate the post-selection criterion into the extraction.

Conventional hypothesis-driven techniques define spatio-temporal activity patterns by means of correlation analysis between each voxel time course and a *reference* function, that is related to the experimental activity that is investigated. Therefore each voxel that correlates satisfactorily with the reference function is considered involved in the activity.

Conversely, ICA provides a spatial map whose voxels all share the same time course; usually one or more components exhibit a task-related time course, and it not possible to predict in advance which is the position of those components among the whole set of IC. It is therefore possible to " constrain" the first(s) component(s) to be "related" to the task by means of a suitable function H. For what concerns temporal ICA this is quite straightforward, by means of correlation between estimated independent component $\mathbf{y}$ and a reference function $\mathbf{r}_f$:

$$H(\mathbf{y}) = \begin{cases} \mid \mathrm{corr}(\mathbf{y}, \mathbf{r}_f) \mid & \text{if} \quad \mid \mathrm{corr}(\mathbf{y}, \mathbf{r}_f) \mid \leq \kappa \\ \kappa & \text{if} \quad \mid \mathrm{corr}(\mathbf{y}, \mathbf{r}_f) \mid > \kappa \end{cases} \tag{7.3}$$

The value of κ influences the effectiveness of the separation. As pointed out by [109] for the constrained ICA framework, an iterative procedure can be performed varying κ to select properly the right maximum .It has to noted, however, that the sources, i.e. the temporal time courses, are independent and therefore uncorrelated: imposing to be near to one source automatically excludes all the others.

For what concerns a *spatial* constraint on *spatial* ICA, the procedure is the same as in (7.3), since the additional function H is related to the independent component estimate. In this case the prior information exploited could be related to the position of the activation of an independent component (performing a Region Of Interest (ROI) analysis).

On the other hand, some problems may arise in constraining the temporal time course in spatial ICA and the spatial map in temporal ICA, if the threshold (κ) is not properly set. In fact, as discussed in 3.3.6, the “dual” signal in the ICA model for fMRI data analysis has no requirements of orthogonality. Temporal time courses associated to independent spatial maps may be correlated, so may be spatial maps associated to independent time courses. This considerations show that it could be troublesome to detect the task-related source first by means of a single extraction with a correlation constraint on its time course, if the threshold value is not close enough to the “real” correlation value, since more than one independent component may have a time course with a correlation at least of κ. In the next section, constraints on temporal time courses of spatial maps will be employed, but with different purposes, since in that case the threshold value κ is near or higher to the one of the target independent source.

7.2.2 Enhancing ICA capabilities: a case study on real data

To show the possibility of enhancing ICA capabilities of detecting “interesting” components, once some specific prior information is available on them, a study on 9 subjects performing an fMRI experiment has been conducted. In this case a preliminary extraction has been conducted with FastICA, and then the extraction with different constraints has been performed in order to evaluate their effectiveness.

Nine subjects performed a finger tapping experiment with the right hand: during acquisition, the participant executed a simple motor task consisting of self-paced, sequential finger-to-thumb opposition movements with the right hand, alternating equally long activity and rest periods, as ruled by acoustic signals transmitted to the subject through the scanner earphones. Time series of fMRI data were acquired from 9 healthy individuals, using a Philips Vision Gyroscan MR system operating at 1.5T and equipped for echoplanar imaging. A circular polarized volume head coil was used for radio frequency transmission and reception. Head movement was minimized by mild restraint and cushioning. Functional T2* weighted images (25 axial slices, thickness=4mm, TR=3000ms, TE=50ms, FA=90^o, FOV=240cm, 64×64 matrix, voxel dimension = $3.75mm \times 3.75mm \times 4mm$) were acquired from the participants. One data acquisition run consisted of a series of 90 such BOLD sensitive images. The 6 initial BOLD images were discarded to allow for magnet settlement. Since no anatomical scan was available, it has not been possible to point out only the voxels related to the cortex (i.e. the ones where the activation is present), and therefore the analysis has been carried out considering volume pixels. The only preprocessing performed has been removing the voxels outside the brain, to restrict the analysis only to the portion of space that may be involved with the experiment. This has been done by means of a threshold on intensity values, that allowed to discard voxels outside the brain.

A preliminary extraction has been conducted with FastICA in deflation approach. Therefore effectiveness of the separation has been evaluated in terms of correlation between the time-course of the task-related component and the experimental paradigm. Of course, being ICA an unsupervised technique, it has not been possible to evaluate in an automatic fashion the spatial map. However, the spatial maps were examined by an expert to state if they were meaningful and task related.

The next step has been testing the constrained version of the algorithm by adding two different additional functions. The first is directly related to the experimental paradigm $\mathbf{r}_f$ (derived from the task time course after convolution with an estimate of the hemodynamic response function), known *a priori*, and is:

$$H_1(\mathbf{w}) = \begin{cases} g_1(\mathbf{w}, \mathbf{r}_f) & \text{if} \quad g(\mathbf{w}, \mathbf{r}_f) \leq \kappa_1 \\ \kappa_1 & \text{if} \quad g(\mathbf{w}, \mathbf{r}_f) > \kappa_1 \end{cases} \tag{7.4}$$

where $g_1(\mathbf{w}, \mathbf{r}_f)$ is defined as the absolute value of the correlation of the map temporal time course $\mathbf{a}_i$ (reconstructed from $\mathbf{w}$ by means of dewhitening matrix) and the reference function:

$$g_1(\mathbf{w}, \mathbf{r}_f) = \mid \operatorname{corr}(\mathbf{a}_i, \mathbf{r}_f) \mid \tag{7.5}$$

The second contrast function employed was somehow similar, but more general. In fact, the frequency bandwidth of the time course has been constrained to lay in the interval $[\omega_1, \omega_2]$, defined according to the temporal paradigm of the experiment:

$$H_2(\mathbf{w}) = \begin{cases} g_2(\mathbf{w}, \omega_1, \omega_2) & \text{if} \quad g(\mathbf{w}, \omega_1, \omega_2) \leq \kappa_2 \\ \kappa_2 & \text{if} \quad g(\mathbf{w}, \omega_1, \omega_2) > \kappa_2 \end{cases} \tag{7.6}$$

where $g_2(\mathbf{w}, \omega_1, \omega_2)$ is defined as the ratio between the spectrum power in $[\omega_1, \omega_2]$ and in the all the frequency domain (computed by means of FFT):

$$g_2(\mathbf{w}, \omega_1, \omega_2) = \frac{PSD_{\omega_1,\omega_2}}{PSD_{tot}} \tag{7.7}$$

Parameters $[\omega_1, \omega_2]$ were chosen according to frequency spectrum of $\mathbf{r}_f$, in particular $\omega_1 = 0.0075\ rad/s$, while $\omega_2 = 0.35\ rad/s$, defining a quite wide frequency band centered around stimulus frequency. κ_1 was set to 0.8 (i.e. at least a correlation of 0.8 between temporal time-course and reference function was required), while κ_2 was set as 0.95 (requiring thus that at least 95 % of the frequency spectrum of the signal lay in the region defined).

Extraction comparison has been carried out in terms of correlation ρ and of non-Gaussianity value, pointed out by kurtosis $kurt$ and results are presented in Table 7.1.

Subjects 2 and 9 are considered as example of the methodology. The most physiological independent component retrieved with FastICA for the second subject has a correlation of 0.65 with the reference function, while the two proposed methodologies increase this value, losing some percentage value in kurtosis (i.e. independence). The best IC extracted by FastICA from data of subject 9, on the other hand, has a good correlation, therefore H_1 constraint just "selects" the component, while constraint H_2 reaches a higher value of correlation. Results relative to the analysis on subject 2 are depicted in Figures 7.3, 7.4 and 7.5, the ones related to subject 9 are in Figures 7.6, 7.7 and 7.8 (it has to be noted that the scans of subject 9 are in radiological convention, meaning that left and right in the transversal plane are flipped, therefore the activity in the controlater motor area appear on the other side w.r.t subject 2). Since no anatomical information was available for the subjects, the results are displayed with the activity superimposed to functional scans (in gray), and the activations have been superimposed on a volume reconstructed from the functional scans by means of MRIcro (a free display tool for fMRI data visualization, available at http://www.sph.sc.edu/comd/rorden/mricro.html).

The analysis performed suggest some important considerations. In fact, as it is possible to infer from kurtosis level of the recovered sources, this constrained

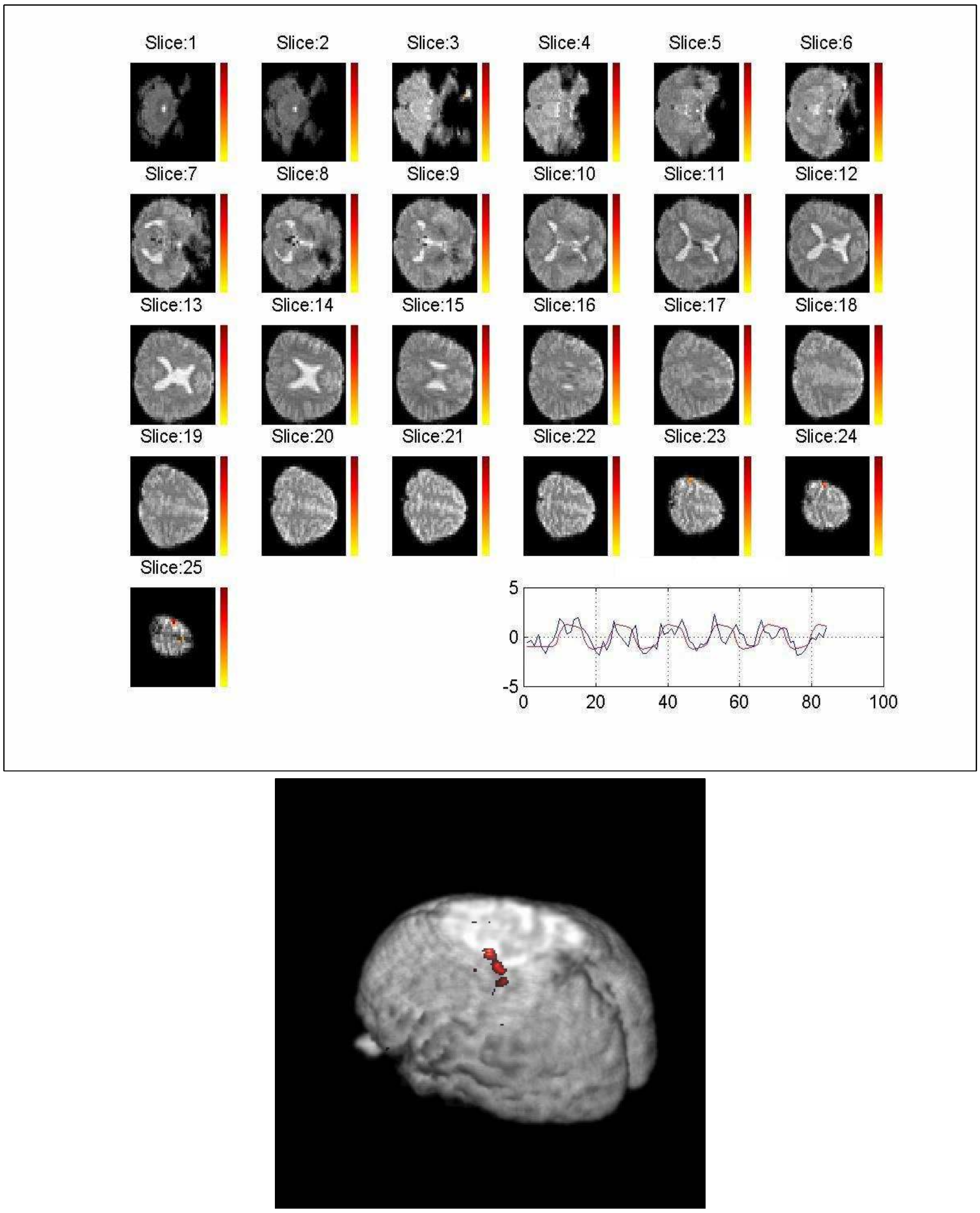


Figure 7.3: Results on 2nd subject with FastICA

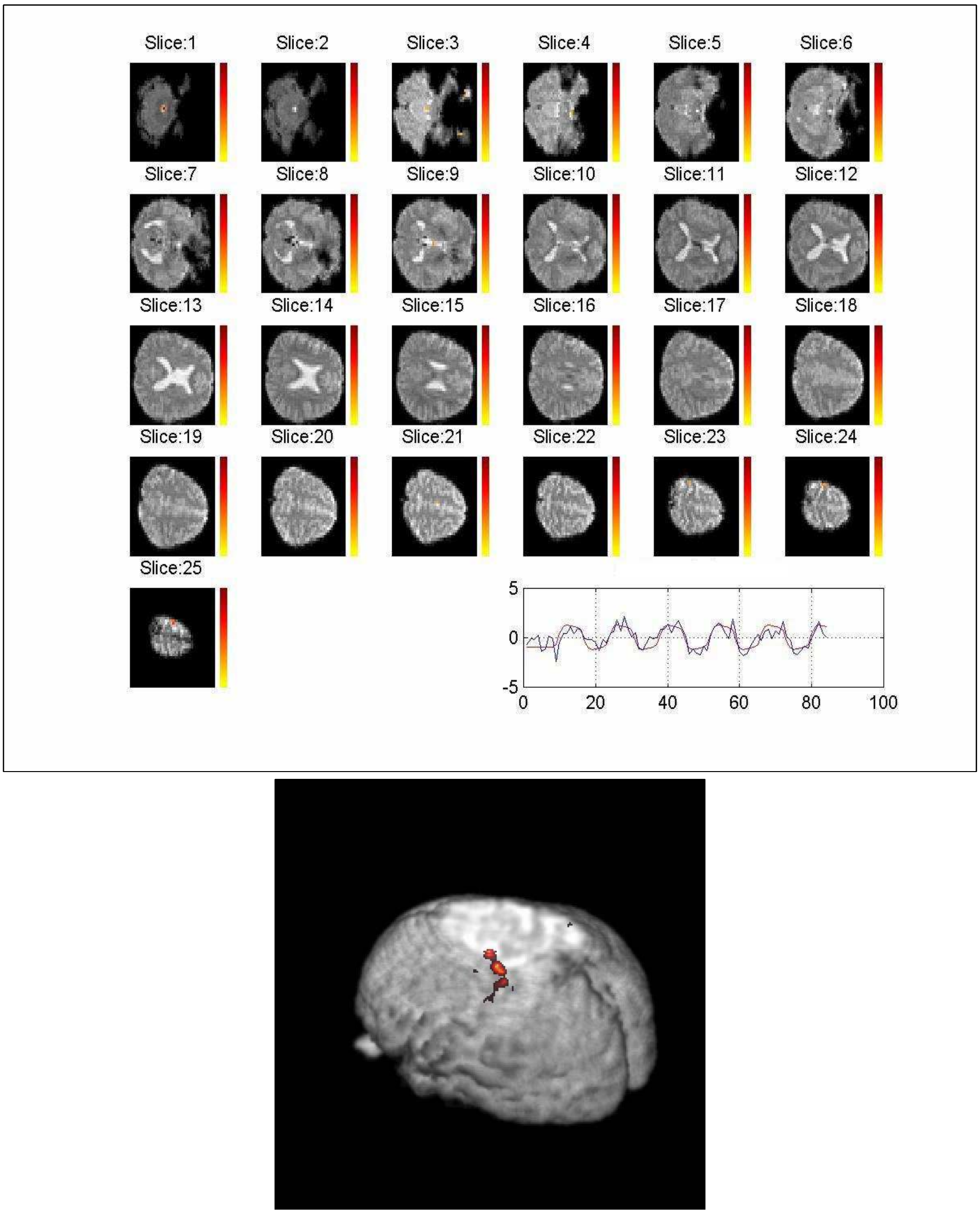


Figure 7.4: Results of the analysis on 2nd subject with constraint H_1

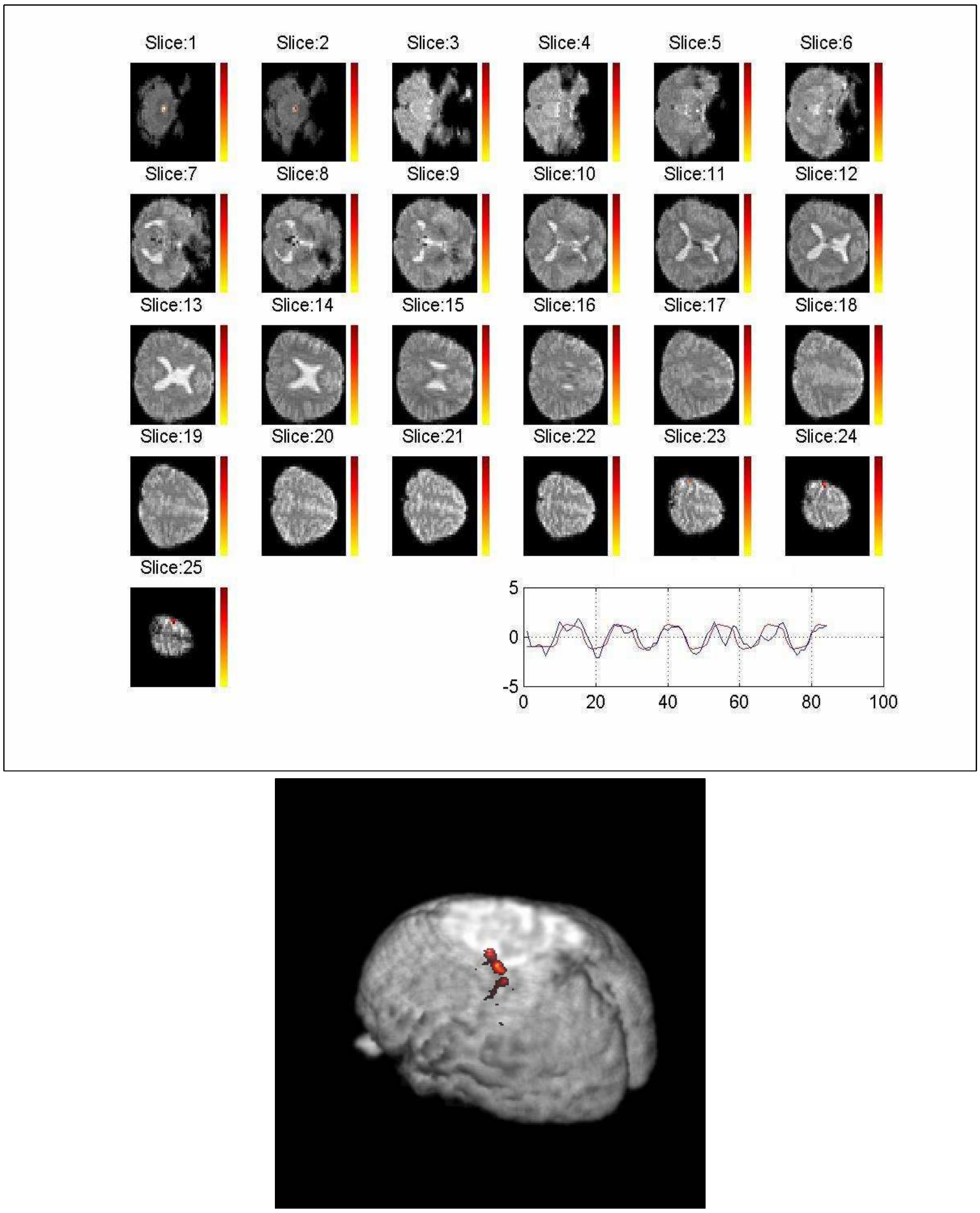


Figure 7.5: Results of the analysis on 2nd subject with constraint H_2

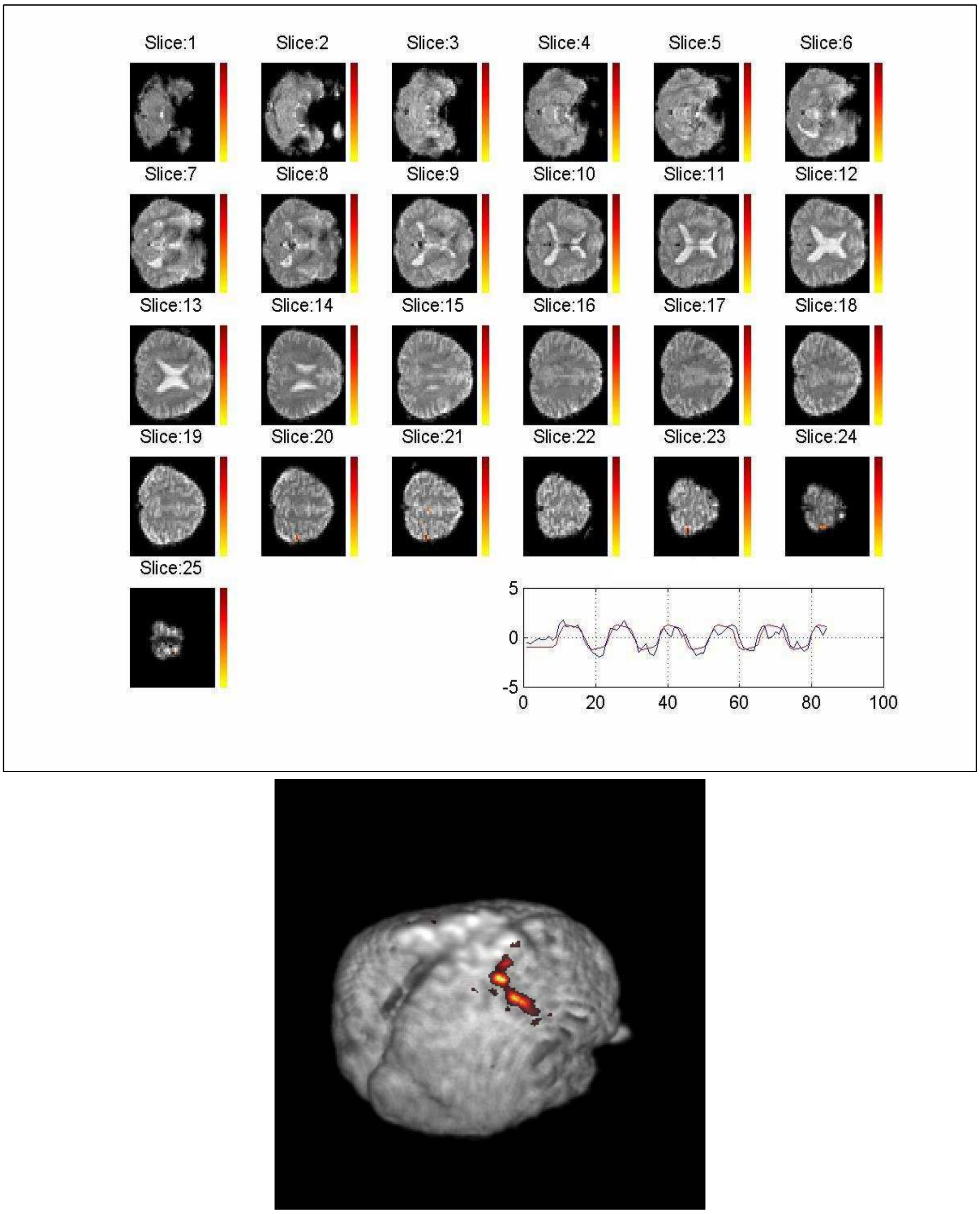


Figure 7.6: Results on 9th subject with FastICA

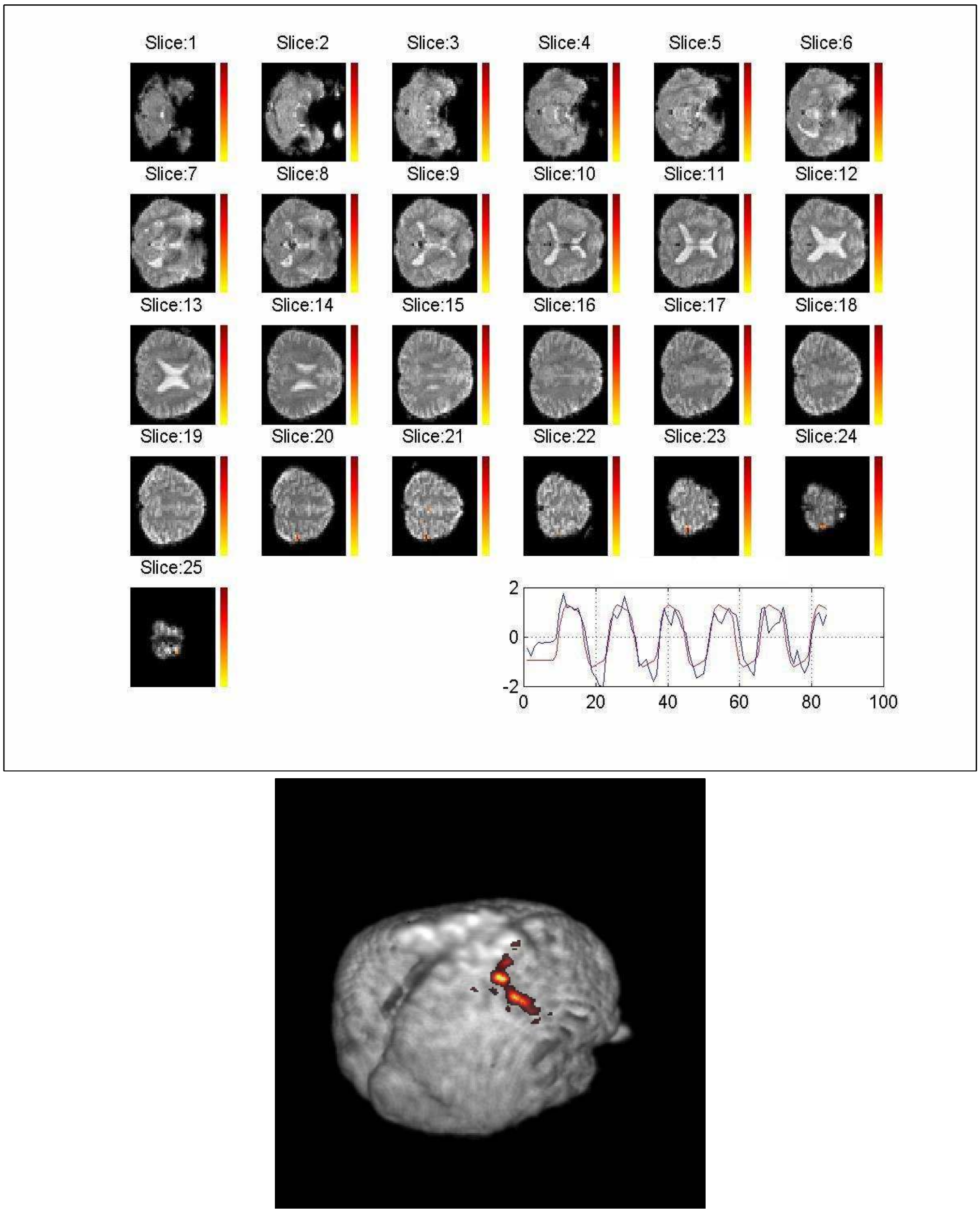


Figure 7.7: Results of the analysis on 9th subject with constraint H_1

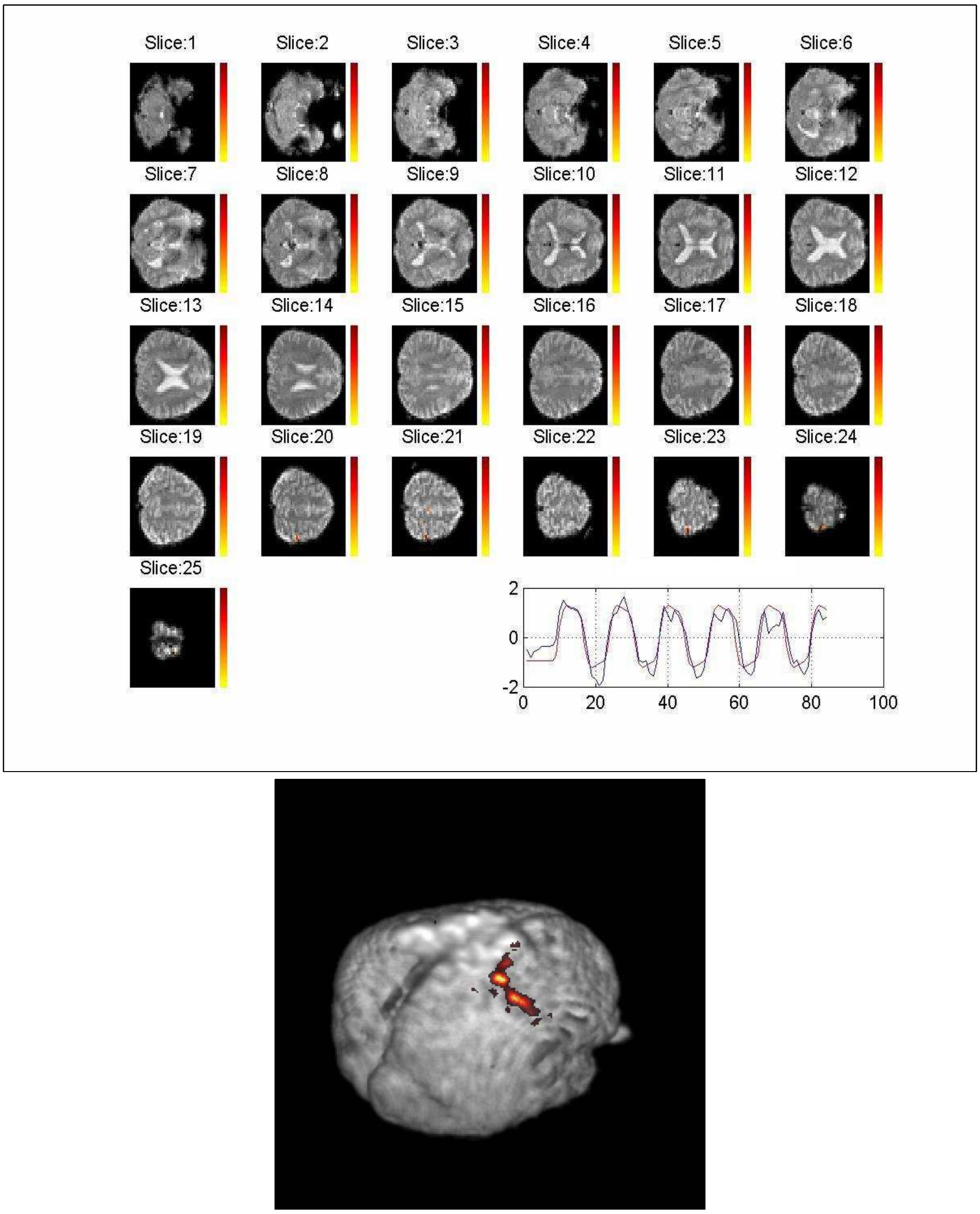


Figure 7.8: Results of the analysis on 9th subject with constraint H_2

Subject	FastICA		$J_G + H_1$		$J_G + H_1$	
	ρ	$kurt$	ρ	$kurt$	ρ	$kurt$
1	0.40	23.2	0.80	12.5	0.88	15.9
2	0.65	125.9	0.80	121.6	0.86	92.3
3	0.39	4.64	0.80	14.0	0.83	10.9
4	0.80	51.1	0.80	52.0	0.93	36.1
5	0.61	19.4	0.80	48.3	0.81	25.3
6	0.5	628	0.75	417.3	0.63	505.2
7	0.60	77	0.80	96.4	0.94	86.2
8	0.87	25.9	0.91	23.1	0.89	31.6
9	0.87	181.2	0.87	180.4	0.92	170

Table 7.1: Results of the analysis conducted with FastICA, H_1 and H_2

version of ICA finds "less" independent components, w.r.t FastICA. However, on the other hand it is possible to force some of the sources to fulfill some requirements (like correlation with a reference function or the presence in a specific band), having more plausible sources. The aim of the work presented in this section has been therefore to show how is it possible to constrain ICA to extract more "plausible" sources from a dataset by exploiting some prior information on the sources.

7.3 Multi-objective optimization

Another interesting application of the framework shown in Chapter 6 is the multi-objective optimization. As pointed out in the introduction, there are some situations where a constrained approach may not be suitable to point out some prior knowledge on the sources, since this information is rather “loose”, and therefore cannot be employed for all the components. An illustrative example for fMRI data has already been shown in 6.2: the three spatial maps have the same histogram (i.e. the same contrast function value), but they have different spectral properties. The example is theoretical, yet is indeed representative of actual situations. Isolated voxels in an ICA analysis are usually considered as *false positives*, but there is no way to prevent ICA from finding them. As it has been stated several times throughout Chapter 6, ICA is a technique *not designed for structured data*, since it looks only for statistical independence, a property related to the whole sample space of the signals, but it does not account for regularities. It has been shown that second order methods do pursue independence starting from different criteria (i.e. decorrelation at different lags), but this method is not flexible enough for the fMRI case. In fact, both temporal time courses and spatial maps, that represent physiological independent activity, show some kind of regularities that it would be useful not to ignore while extracting independent components.

The issue of how useful this information could be is not trivial, since independence still remains, in this case, the “guiding” principle for signal extraction. To prove the effectiveness of this approach, a resting state experiment has been made and some artificial fMRI-like activations have been superimposed, to show how noise level affects independence measures and how it is possible to recover sources with a modified contrast function in a way more effective than with pure negentropy (7.3.4). Before moving to the experiment, it may be useful to point out which are the principles that may be employed to recover fMRI sources and which are the best choices for the additional contrast function H.

7.3.1 Non independent strategies for fMRI

As seen in section 3.3.5, several techniques can be employed in order to retrieve information from fMRI scans. In particular, the assumption of statistical independence among different activities in the brain has proved to be realistic in most of the cases, and ICA has become a widely accepted technique for fMRI data analysis.

Generally speaking, Blind Source Separation can be performed by means of criteria other than independence. In [148] several strategies have been employed to recover task-related sources that have been mixed linearly, in particular spatial predictability and non stationarity. Spatial predictability can be seen as a measure of the complexity of a spatial map ([5]), and the sources have a spatial predictability that is higher (or equal) to those of their mixtures. It has to be noted, however, that while spatial predictability forces the extracted signals to be smooth and of low complexity, independence forces signals to be independent and, in case of positive kurtosis, to have sparse representations. By assuming that all the sources have different autocorrelation structure and are spatially correlated, it is possible to model them by means of an Auto Regressive (AR) process:

$$c_j(p) = \tilde{c}_j(p) + \sum_{q=1}^{Q} a_{il} c_j(p-q) \tag{7.8}$$

where $\tilde{c}_j$ denotes i.i.d. innovative process and a_{il} denote the linear filter coefficients. Consider a linear predictor which gives the prediction error ε_j for the estimate of the jth source:

$$\varepsilon_j(p) = y_j(p) - \sum_{q=1}^{Q} b_{jl} y_j(p-q) \tag{7.9}$$

where $\mathbf{b}_j$ gives the coefficients of the FIR filter which performs the linear prediction of the source signal y_j. In [19] it has been shown that it is possible to recover sources (by means of mixing coefficients $\mathbf{w}$) and filter coefficients, if the sources have a structure that can be modeled.

The second approach presented in the same work, non stationarity of the sources, can be employed if the noise is white and all undesirable signals are stationary.

Both these approaches were able to recover the target sources in an artificial fMRI dataset, and in an alternating checker board fMRI experiment, even if kurtosis-based methods seemed to perform better than other techniques in terms of correlation between the experimental paradigm and the time course.

A similar approach, based on temporal predictability (TP), has been proposed (for further details on temporal predictability, see section 6.2.2). TP has been employed as a regularization factor for time courses in spatial ICA for fMRI by Stone in [155], while it has been proposed as a general principle for blind source separation in [154], starting from the conjecture that the independent sources have a TP greater or equal to the TP of their mixtures. In [174] it has been demonstrated that such conjecture is wrong, and it has been proved that the TP of a mixture of independent signals is bounded by the minimum TP and the maximum TP of the sources, and a new BSS algorithm based on this principle has been proposed.

In [59], the Canonical Correlation Analysis (CCA) has been employed as an alternative approach for data-driven analysis of fMRI data. The starting principle is that "interesting" signals in fMRI are spatially and temporally autocorrelated, while their mixtures, due to uncorrelated noise, are less autocorrelated than the original sources. The algorithm proposed is able to extract task related signals, together with other activites that are usually detected with ICA (like linear trends and motion artifacts) that are highly autocorrelated.

All these works show that it is possible to employ several strategies involving other properties that are not directly related to independence of the sources, but that can be useful in retrieving task-related spatio-temporal activity patterns. The next section will focus on the choice of the additional contrast function H to add to negentropy in order to make a better extraction of sources that have remarkable spatio-temporal regularities.

7.3.2 Spatial and Temporal regularities

In the previous section some of the strategies presented in literature to perform blind source separation with a criterion different from independence have been presented. It has been shown that it possible to recover independent sources of fMRI activity also with other criteria involving some kind of knowledge of the spatio-temporal nature of sources. However, one may not want to lose the "blindness" ability of ICA to recover the source, i.e. it is not known in advance which is the linear or spatial model of each of the sources, but the only knowledge one may want to exploit is the fact that isolated active points are more likely to be false positive rather than the result of an inner cerebral activity. In other words, the prior knowledge one may want to incorporate in fMRI data analysis must be unspecific, in the sense that it is not known *a priori* where a source is localized spatially, which time course the activity has, or which is the frequency content both in space and time of each source. Therefore an ideal contrast function H should give a "measure" of how much the source is spatially and temporally regular, being insensitive to sign and amplitude of the source (due to double indeterminacy). On the other hand, since Simulated Annealing optimization requires the computation of F at each perturbation, there is also the need to avoid defining too complex contrast functions H, since their computation may take too much time and the optimization may require an overwhelming amount of time to reach the global optimum.

The functions H explored in present work are therefore spatial and temporal one-lag autocorrelation, since they are relatively easy to evaluate for any linear combination of observed signals, and they denote in a robust way the regularities over space and time, regardless of the position of the activities. Both these functions are bounded between -1 and 1, therefore it is easy to implement them as additional contrast function H and to weigh them properly by means of λ.

Temporal one-lag autocorrelation

Temporal one-lag autocorrelation can be computed easily in the following way:

$$H_t(y) = \frac{1}{\sharp(T)} \sum_{t \in T} y(t)y(t+1) \tag{7.10}$$

where $\sharp(T)$ denotes the number of time points. Since the simulated annealing optimization is performed on whitened data, it is necessary to state the relationship between mixing coefficients and time course of the spatial map. This is straightforward, considering the relationship:

$$\mathbf{A} = \mathbf{D}_{WM}\mathbf{W} \tag{7.11}$$

where $\mathbf{A}$ denotes the estimate of the mixing matrix for the original dataset, $\mathbf{D}_{WM}$ denotes the *dewhitening* matrix (i.e. the inverse of the whitening matrix that transforms original data into whitened ones), and $\mathbf{W}$ is the independent components unmixing coefficients estimate for the whitened dataset.

Spatial one-lag autocorrelation

Spatial one-lag autocorrelation can be expressed, in a very general form, as:

$$H_{sp}(\mathbf{w}) = \frac{1}{\sharp(I)} \sum_{i \in I} \sum_{q_i \in \mathcal{N}_i} \frac{F(i)F(q_i)}{\sharp(\mathcal{N}_i)} \tag{7.12}$$

where I denotes the set of all space points of a spatial maps, $\mathcal{N}(q_i)$ denotes the set of all the one-lag neighbors of point i, and $\sharp(I)$ and $\sharp(\mathcal{N}_i)$ denote the cardinality of the set of spatial points and of neighbors of spatial point i.

The procedure one has to implement to evaluate spatial one-lag autocorrelation described in eq. (7.12) is the following:

- For each point i of the spatial map:
  - Find the set $\mathcal{N}_i$ of *all* the neighbors of i.
  - Evaluate the sum of the products of the spatial map in i and in all of his neighbors
  - Average this sum by the number of neighbors of i
- Average the value of spatial autocorrelation for each spatial point over the whole spatial map.

Although the procedure may seem quite expensive, from a computational point of view, it is possible to evaluate it in a relative fast way due to the nature of the problem.

Consider a spatial map $\mathbf{y} \in \Re^{m\times 1}$, with $\mathbf{y} = \mathbf{w}^T\mathbf{X}$, where $\mathbf{w} \in \Re^{n\times 1}$ is the unmixing coefficient estimate and $\mathbf{X} \in \Re^{n\times m}$ is the whole fMRI dataset. Since a whole volume is usually considered as a single row of data matrix, it is not possible to exploit directly the neighbor relationships. To overcome this limitation, it is possible to build a *Neighbor matrix* $\mathbf{N} \in \Re^{m\times m}$ to evaluate spatial autocorrelation easily. Each element $\mathbf{N}_{ij}$ of $\mathbf{N}$ is defined as follows:

$$\mathbf{N}_{ij} = \begin{cases} \dfrac{1}{\sharp(\mathcal{N}_i)} & \text{if} j \in \mathcal{N}_i \\ 0 & \text{if} j \notin \mathcal{N}_i \end{cases} \tag{7.13}$$

Neighbor matrix therefore points out the neighbors of each voxel in a spatial map, and the quadratic form $\mathbf{y}\mathbf{N}\mathbf{y}^T$ is, up to a constant (the number of spatial points in the volume), equivalent to spatial autocorrelation, therefore:

$$H_{sp}(\mathbf{y}) = \frac{1}{\sharp(I)}\mathbf{y}\mathbf{N}\mathbf{y}^T \tag{7.14}$$

Considering that in each iteration of Simulated Annealing optimization, $\mathbf{y} = \mathbf{w}^T\mathbf{X}$, it holds:

$$H_{sp}(\mathbf{w^T}\mathbf{X}) = \frac{1}{\sharp(I)}\mathbf{w}^T\mathbf{X}\mathbf{N}\mathbf{X}^T\mathbf{w} = \mathbf{w}^T\mathbf{N}_1\mathbf{w} \tag{7.15}$$

where $\mathbf{N}_1 \in \Re^{n\times n}$ and $\mathbf{N}_1 = \frac{1}{\sharp(I)}\mathbf{X}\mathbf{N}\mathbf{X}^T$, it is possible, if spatial autocorrelation has to be considered for all the spatial points, to evaluate it in a fast and compact way. In fact, matrix $\mathbf{N}_1$ has to be computed only once before the analysis (since it depends only on problem topology and original data), and the spatial autocorrelation evaluation consists only of a matrix multiplication.

There are situations, however, when the spatial autocorrelation has to be computed only considering a limited subset of the all dataset, i.e. only on the tails of the distribution of a spatial map. In this case, since the thresholding procedure is performed for each linear combination, it is not possible to incorporate data matrix $\mathbf{X}$ into $\mathbf{N}_1$, and therefore $\mathbf{N}$ has to be employed at each iteration.
To test the proposed methodology, the algorithm has been employed on the artificial dataset and on a resting state experiment. There are some differences in the two methodologies: in fact, while the artificial dataset is such that *all* the three sources are highly autocorrelated, while the additive noise is i.i.d. with Gaussian distribution, the resting state experiment allows to account for environmental fMRI noise and for all the activities that are present in the brain regardless of the experiment performed. In the first case, therefore, a constrained optimization has been performed in order to point the task related sources, by imposing that they have autocorrelations of at least a given value. For the resting state experiment, where artificial activations have been superimposed to real brain activity, this has not been possible, since in the high noise case it was not possible to detect the injected activations. This is related to the fact that autocorrelated noise and artifacts tend to "cover" the artificial activity, and an unspecific constraint is not able to reveal the low signal to noise ratio, as it tends to privilege "strong" sources. Therefore the multi-objective approach has been implemented by adding additional terms to the contrast function.

7.3.3 Test on Simulated Data

Evaluating the results of an ICA extraction may be sometimes troublesome due to the fact that it is an *unsupervised* technique. In fMRI ICA data analysis, the evaluation is done by experts that may give a *score* to an independent spatial map and to his time-course according to the physiological plausibility of the spatio-temporal patterns, but, for a *quantitative* rather than a *qualitative* evaluation of results of a new ICA extraction, usually artificial signals are employed. In fact, by means of an artificial dataset, it is possible to:

- Evaluate the quality of the results by means of correlation with original sources and mixing coefficients.

- Compare different techniques in terms of performances.

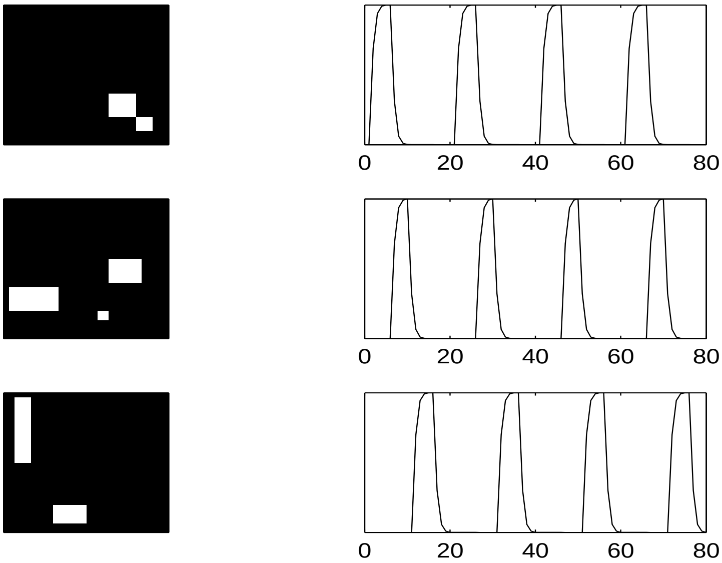


Figure 7.9: Simulated fMRI dataset. Left: spatial maps. Right: associated time courses

It has been therefore developed an artificial fMRI-like dataset and the algorithm has been tested on it, and results have been compared with FastICA extraction.
The simulated dataset consisted of three spatial maps with associated time-courses, fulfilling the ICA assumptions for fMRI data (discussed in [118]). The three spatio-temporal patterns are depicted in figure 7.9: active voxels of each spatial map (that consisted of 4096 points) are drawn in white, while temporal time courses of the spatial map are partially overlapping and are physiologically plausible, since they have been generated by convolving a fake experimental paradigm with three different estimates of the hemodynamic response function, estimated according to Boynton model ([31]).
To generate the whole dataset, the three spatial maps have been mixed linearly and i.i.d. Gaussian noise has been added. Standard deviation σ_n of the noise has been set to different levels, and performances have been evaluated at different noise levels. Performances evaluations are carried out by means of correlation between the recovered sources and the original ones, and between the recovered time courses and original ones.

The three artificial fMRI activations depicted in Figure 7.9 were mixed linearly and Gaussian white noise with standard deviation σ_m, ranging from 0 to 2, was superimposed. It has to be noted that artificial time courses range from 0 to 1 and active pixels have a unit intensity value.

To evaluate the effectiveness of the source separation, a correlation analysis was carried out. The correlation of the original map and the corresponding recovered map was computed, together with correlation between original time course and recovered one. Three different contrast functions have been tested: spatial autocorrelation, temporal autocorrelation and a linear combination of them. The three contrast functions considered are:

$$H_1(\mathbf{w}) = \begin{cases} H_{sp}(\mathbf{w}) & \text{if} H_{sp}(\mathbf{w}) \leq \kappa_1 \\ \kappa_1 & \text{if} H_{sp}(\mathbf{w}) \geq \kappa_1 \end{cases} \tag{7.16}$$

$$H_2(\mathbf{w}) = \begin{cases} H_t(\mathbf{w}) & \text{if} H_t(\mathbf{w}) \leq \kappa_2 \\ \kappa_2 & \text{if} H_t(\mathbf{w}) \geq \kappa_2 \end{cases} \tag{7.17}$$

$$H_3(\mathbf{w}) = H_1(\mathbf{w}) + \alpha H_2(\mathbf{w}) \tag{7.18}$$

where κ_1 has been set as 0.5, κ_2 as 0.4 and α, that weighs the two functions, has been set as 0.2 (since it has been seen that temporal autocorrelation is the less effective of the three contrast functions). In Figures 7.10, 7.11 and 7.12 the correlations between the recovered sources and the original ones are confronted with FastICA. The red color denotes the spatial autocorrelation constraint H_1, the green color is related to temporal autocorrelation constraint H_2, while the black line is related to spatio-temporal autocorrelation H_3.

Results of the simulation show that it is possible to recover the three sources also in extremely noisy environment, when a classical ICA extraction fails to separate correctly the sources. This is not surprising, since constraining the recovered sources to have an autocorrelation above a fixed threshold enhances points out the main difference between the three artificial activations and the noise, that has no autocorrelation both in space and time.

The poor result obtained with temporal constraint H_2 (that reflects also on the small difference between spatial H_1 and spatio-temporal H_3 constraints) may be due to the poor sample size (there were only 100 time points, while there were 4096 spatial points).

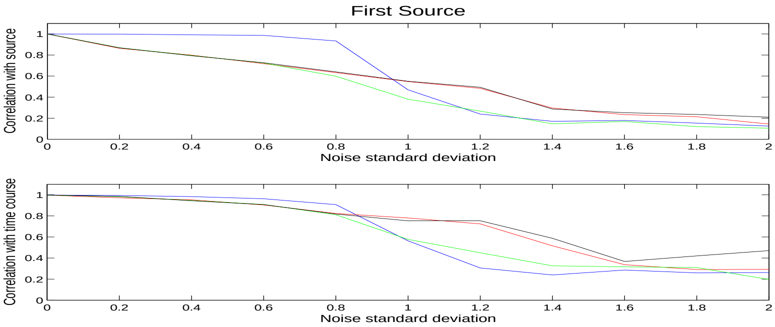


Figure 7.10: Correlation analysis for the first source. Blue line: FastICA. Red line: Spatial autocorrelation H_1. Green line: temporal autocorrelation H_2. Black line: spatio-temporal autocorrelation H_3

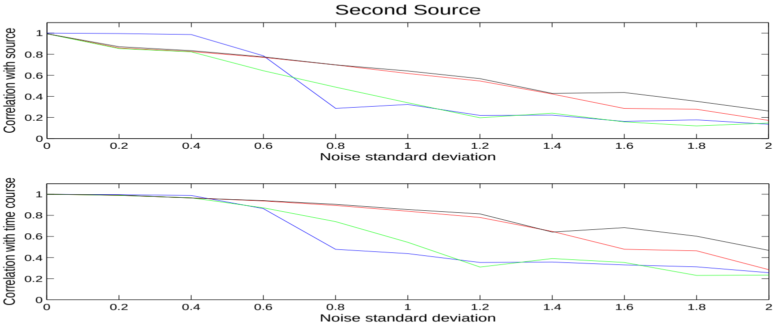


Figure 7.11: Correlation analysis for the second source. Blue line: FastICA. Red line: Spatial autocorrelation H_1. Green line: temporal autocorrelation H_2. Black line: spatio-temporal autocorrelation H_3

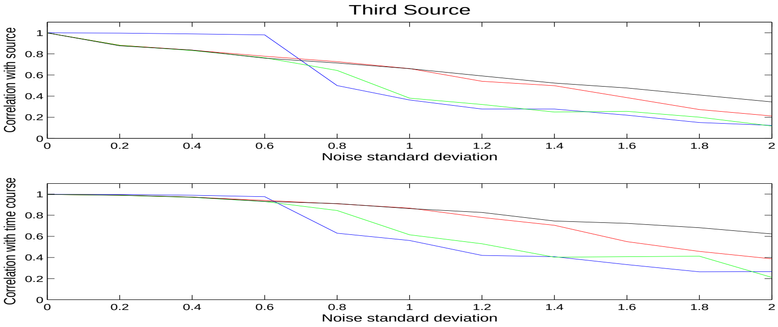


Figure 7.12: Correlation analysis for the third source. Blue line: FastICA. Red line: Spatial autocorrelation H_1. Green line: temporal autocorrelation H_2. Black line: spatio-temporal autocorrelation H_3

7.3.4 Resting State Experiment

To evaluate the effectiveness of source separation in fMRI experiments, usually resting state experiments are conducted, and superimposed activities are injected at different noise levels, to test if the technique employed is able to recover these activities in different conditions. In [53], both FastICA and INFOMAX have been tested on a simulated fMRI dataset superimposed to a resting state experiment; in that work it has been shown how the two most employed source separation algorithms perform almost similarly according to different noise levels. In particular, it has been observed that, when noise increases, ICA separating algorithm tend to deteriorate almost abruptly.
A resting state experiment has therefore been performed by a healthy volunteer. The whole brain was acquired on a 3T Siemens Allegra (Repetition Time 1.5s, Interslice time 46 ms, 32 slices, matrix 64×64, slice thickness 3mm, 210 volumes) at the University of Maastricht, department of Cognitive Neuroscience. The first two volumes were skipped due to T2* saturation effects. Data were preprocessed by means of linear de-trending and high pass filtering with BrainVoyager (www.brainvoyager.com). A single slice was selected and activations were superimposed at different noise levels.
Usually the level of noise in fMRI activations is expressed in terms of Contrast to Noise Ratio (CNR) [21]. In this case CNR was evaluated with the following procedure:

1. Consider the set S of spatial points were the activity is going to be injected.

2. Evaluate the standard deviation of the time course of all the time points in S and average it, obtaining σ_m.

3. The CNR is evaluated by considering the signal enhancement Δ_t due to the activity (i.e. the difference between the amplitude of the time course during the rest condition and the activity one), having:

$$\mathrm{CNR} = \frac{\Delta_t}{\sigma_n} \tag{7.19}$$

CNR is somehow similar to signal to noise ratio, and it can be considered as its tailored version for fMRI analysis.
The analysis was performed by means of the three contrast functions defined in

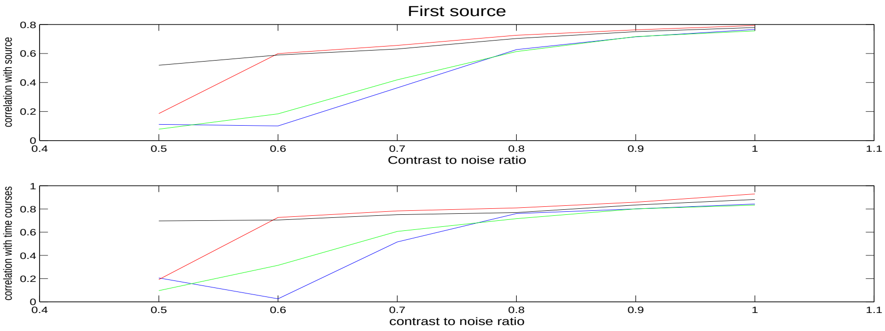


Figure 7.13: Resting state experiment, correlation for the first source. Blue line: FastICA. Red line: Spatial autocorrelation H_1. Green line: temporal autocorrelation H_2. Black line: spatio-temporal autocorrelation H_3

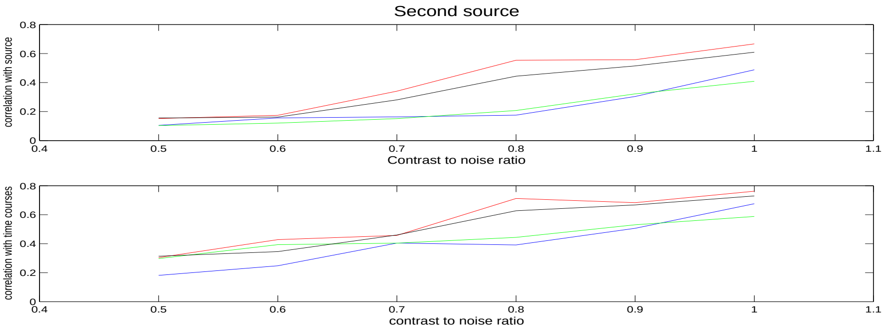


Figure 7.14: Resting state experiment, correlation for the second source. Blue line: FastICA. Red line: Spatial autocorrelation H_1. Green line: temporal autocorrelation H_2. Black line: spatio-temporal autocorrelation H_3

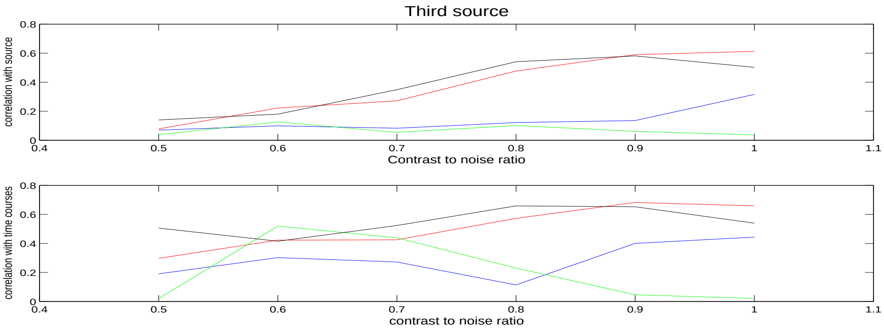


Figure 7.15: Resting state experiment, correlation for the third source. Blue line: FastICA. Red line: Spatial autocorrelation H_1. Green line: temporal autocorrelation H_2. Black line: spatio-temporal autocorrelation H_3

7.3.3, but this time they were not weighed in a constrained optimization fashion, and spatial one-lag autocorrelation has been computed *only* on the 5% most active voxels of each spatial map; this procedure aims at discarding the effects of noise peaks and makes the autocorrelation estimate more robust; on the other hand, it is now not possible anymore to express spatial autocorrelation as a quadratic form of the unmixing coefficients $\mathbf{w}$, since at each iteration the extremal values have to be computed. The weight of H_i, $i = 1, 2, 3$, has been chosen as a 1/10 of the level of negentropy for the sources, as described in section 6.5.1. Results of the extraction are shown in Figures 7.13, 7.14 and 7.15 (these figures should be interpreted as specular to the ones presented in previous section, since in this case the noise decreases while the x-axis increases).
Results are similar to those shown for the simulated dataset, showing an increased ability to recover spatially autocorrelated sources even with considerably low Contrast to Noise ratio. Moreover the three modified contrast functions make the algorithm converge to the right components if the contrast to noise level is greater than 2. Once again, temporal autocorrelation constraint has proved unable to modify ICA measure in a way as effective as spatial constraint. This maybe due to the poor sample size (this time 100 time-points were considered, while a single slice has 4096 points).

7.4 Discussion

It has been shown how it is possible to incorporate prior information about the nature of fMRI data into the ICA extraction in a very general and flexible way. In fact, by means of a modified contrast function, it is possible to account for both specific information (i.e. the knowledge of spatial localization of a source, or of its temporal time course) and general information about spatio-temporal regularities of the sources.
It is possible, by means of the proposed approach, to perform a semi-blind ICA analysis (like the one proposed by Calhoun et al. in [35]) by performing spatial ICA considering the contrast function $F = J_G + H$ using a *symmetric* approach and using a constraint H such that $H(\mathbf{W}) = H_1(\mathbf{w}_1) + H_2(\mathbf{w}_2) + \ldots + H_n(\mathbf{w})$, with H_i enforcing the similarity between the time course of the ith component and the i-th reference function up to a threshold (i.e. the confidence level proposed in [35]). In this way estimated sources are ordered according to the design matrix,

since, among all the permutations of independent components, the one that fulfills the constraints has the highest contrast function value and will be the solution of Simulated Annealing optimization.

The ICA-R (ICA with reference) approach, proposed by Lu et al. in [109] can be performed in a similar way, by considering *temporal* ICA in deflation approach and imposing the correlation constraint on the time course by means of additional function H (as seen in section 7.2.1).

However, the strength of the proposed approach do not lie only in the fact that it is particularly flexible and allows to perform substantially different analysis, but also in the fact that the contrast functions are not forced to be differentiable or expressed in closed form. This means that it is possible to employ more robust estimates, (like seen in section 7.3.4), that have a procedural definition.

Another important issue is the implementation of the multi-objective optimization that accounts for spatio-temporal regularities on the brain **cortex** rather than on the whole volume. In fact, since the brain surface is rugged, with its gyri and sulci, it is not possible to exploit correctly neighboring relationships by *only* considering volume neighborhood. By means of coregistration of anatomical and functional scans related to the same subject, it is possible, using BrainVoyager software, to exploit all the relationship among points of the volume, meaning that is possible to build a neighborhood matrix N_{cortex} that accounts for the spatial structure of the cortex voxels alone.

Due to the fact that the dataset obtained using all the cortex voxel is rather large, a different implementation of the algorithm is currently under study. In fact, since the ordering of the components is not interesting in this case (due to the fact that artifacts and in general many activities not related to stimulus have a large spatial autocorrelation and a high value of non Gaussianity and will tend to be extracted first with a global optimization technique), it is currently under development a hybrid approach, where a first optimization is performed on a differentiable approximation of the contrast function, and therefore global optimization is performed in a limited domain, performing a kind of “fine tuning”.

Another remarkable aspect is the choice of the weight λ in the optimization. In fact, a large value of λ tends to make the optimization a constrained one, while a relatively small one tends to make the additional term λH neglectable. A heuristic trade-off has been employed for the current analysis, but the choice of the optimal weight should be further investigated, also employing probabilistic techniques, like the Bayesian framework, that has been recently employed also in Independent Component Analysis ([23]).

Chapter 8

Use of prior information in ICA for MEG signals

8.1 Introduction

The algorithm presented in Chapter 6 has been applied also to MagnetoEncephalographic (**MEG**) recordings, in cooperation with "Fatebenefratelli" hospital in Rome. A comparison has been made between this modified version of ICA and classical ICA extraction. Prior information on one source, expressed in terms of reactivity to the stimulus has been exploited together with independence, and two new separating techniques have been implemented: Functional Component Analysis (FCA) and Functional Source Separation (FSS). Both these two techniques have proved more effective than classical ICA in recovering sources related to specific activity, for what concerns reactivity itself (that has been incorporated into the algorithm) and also for what concerns localization of the estimated sources (there was no constraint on the spatial pattern of the sources). In section 3.4 a brief background on MEG data analysis has been provided, therefore in section 8.2 the experiment analyzed will be directly addressed, together with a classical ICA analysis, while in sections 8.3.1 and 8.3.2 the two versions of the algorithm will be shown, together with the results of the analysis on the experiment.

8.2 A case study on ICA applied to a MEG experiment

8.2.1 Description of the experiment

An experiment has been conducted at hospital “Fatebenefratelli”in Rome on 15 healthy volunteers (7 females and 8 males, average age 31 ± 2 years), following the declaration of Helsinki procedures. The experimental paradigm was divided into four parts of 3 minutes each:

1. **Stimulation:**

   (a) The **median nerve** was stimulated electrically with two metallic plates with enough intensity to cause a *painless* contraction of the thumb. Electrical pulses lasted for 0.2 ms, with an inter-stimulus interval of 631 ms.

   (b) **Thumb** was stimulated by means of ring electrodes, with an intensity double with respect to the subjective perception threshold. Stimulation timing was the same as in the median nerve case.

   (c) **Little finger** stimulation was made as in the thumb case.

2. **Rest**

   (a) For a period of time of 3 minutes a recording with no activity was made on the subject.

During the experiment, the subject had his eyes open. Magnetic fields were recorded on the contralateral (with respect to the stimulated side) hemisphere (left rolandic region), by means of an MEG equipment with 28 channels. The central positioning was in C3 position (according to international positioning system 10-20 for EEG). MEG sensors were divided as follows:

- 16 axial gradiometers with a 1.8 cm diameter, on two concentric circles.
- 12 magnetometers with a 81 mm^2 area; 11 were on the outer circle, while one was used for balancing and noise suppression.

Global sensitivity of the equipment was approximately $5 - 7\mathrm{fT/Hz}^{1/2}$.
Recorded signals are filtered during acquisition in $0.16 - 250$ Hz, and sampled at a

frequency of 1000 Hz. Signals have been further filtered to reduce noise in frequencies not related with the experiment (muscular and cardiac artifacts) and malfunctioning channels recordings were removed.
The evaluation of MEG recordings aims at individuating signals related to the experimental paradigm, and at localizing spatially those activities within the cerebral cortex. The "interesting"signal is embedded in unstructured noise, related to measurement procedure, to cardiac and breathing artifacts and to baseline cerebral activity, therefore it is no easy task to retrieve the task-related activities by simple visual inspection of the recordings. Several techniques have been employed to evaluate the relevance of the measured activity, and to localize the extracted signal. The one we will refer to in the following is Evoked Activity (**EA**), and a related evoked activity index R_x. Moreover, the localization strategy employed will be discussed.

Evoked Activity

Since evoked response to a sensorial stimulation is embedded in brain baseline activity, that usually has a greater intensity, one way to point out a signal related to a somatosensor stimulus is to consider the periodicity of the experiment. In fact, since brain response to stimuli is temporally synchronized to them while baseline activity is not, it is possible, by means of averaging, to exclude most of the non task-related brain activity. The following procedure is usually employed:

1. The signal is divided in segments of length δ (typically shorter than interstimulus interval).

2. For every segment the signal is averaged, that is:

$$EA_X(t) = \frac{1}{n_r}\sum_{i=1}^{n_r} y(t+T_i), \quad \text{with} \quad t \in [-T_a, T_b] \tag{8.1}$$

   where n_r is the number of segments (i.e. the number of stimulations, therefore, if δ equals the interstimulus interval, then $n_r\delta$ is the length of the recording), $T_a + T_b = \delta$ and T_i is the ith stimulus time defined by the trigger signal. Subscript X means that the evoked activity can be evaluated for different stimulations; in the following of the discussion, X will be M,L and T to denote median nerve, little finger and thumb stimulation respectively.

From a statistical study on the experimental paradigm, the ideal Evoked Activity for the stimulations analyzed is depicted in Figure 8.1. It can be noted that often there

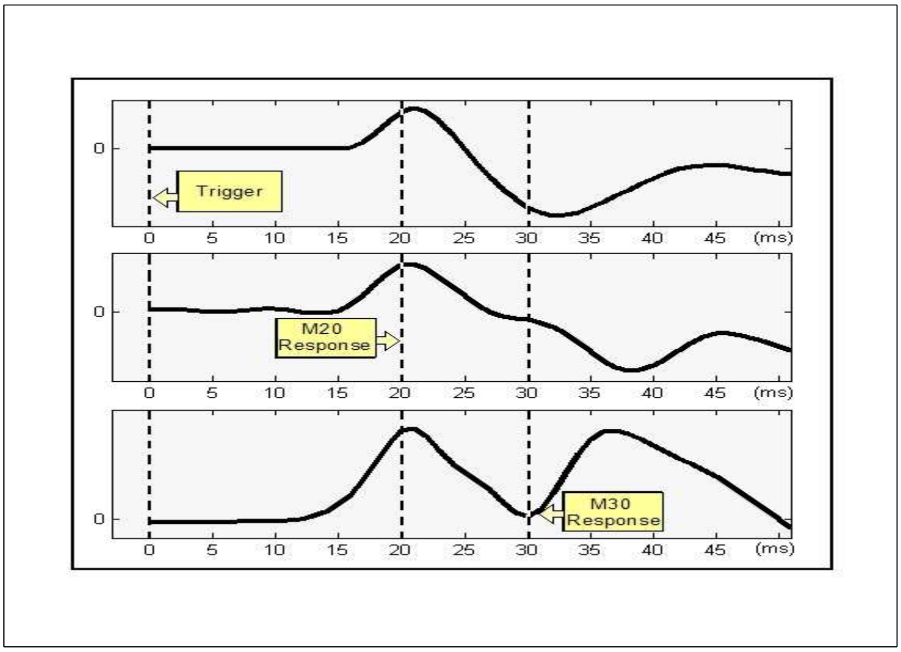


Figure 8.1: Typical Evoked Activity time-courses for healthy subjects

are two amplitude spikes, corresponding at about 20 and 30 ms after every stimulus: in the first case the active zone is the primary somatosensory area (connected with talamus), while in the second the activation is mainly related to inhibitory structures in somatosensor area and to primary motor area. It has to be noted, however, that amplitudes and latencies have an inter-subject variability that can be related to age and sex. The averaging process has the effect of removing uncorrelated noise, but for the experiment considered it was not effective for all subjects. In fact, consider Figure 8.2, relative to one case where averaging technique results were not satisfactory. In this case the stimulus presentation (*trigger event*) time point is at 30 ms. For the majority of the channels it is troublesome to point out the expected peak 20 ms after the stimulus. Moreover, the stimulus presentation itself is manifested almost in any recording as a *stimulus artifact*, since it is determined by an external cause, rather than from neuronal activity.
To express in a *quantitative* rather than in a *qualitatively* way the amount of evoked activity in the target intervals, a reactivity index R_x has been developed for the three stimulations. The following considerations can be made:

- The response to the somatosensory stimulus considered in this experiment is mainly concentrated in the time interval going from 20 to 50 ms after the

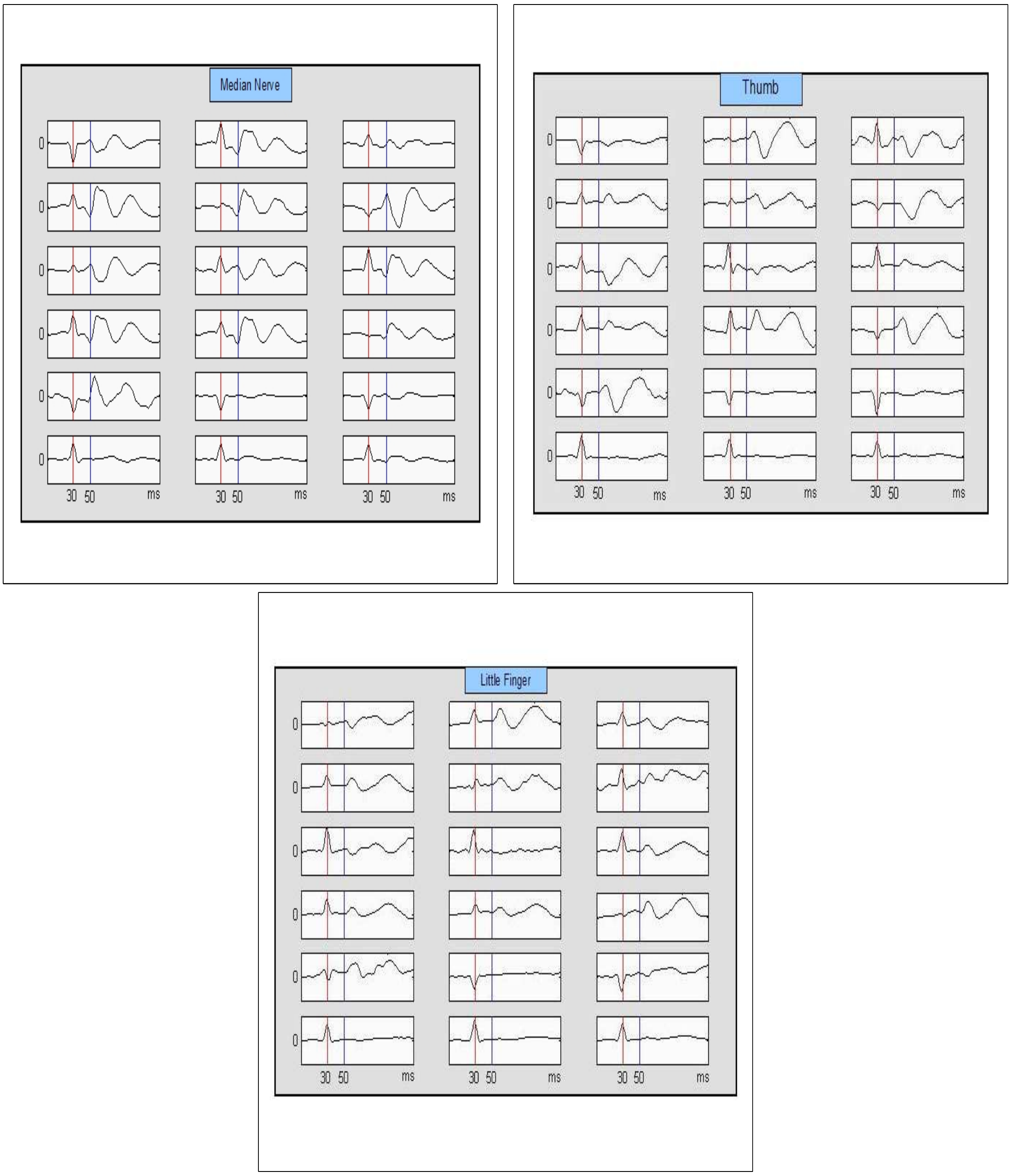


Figure 8.2: MEG recordings averages related to different stimulations

stimulus presentation (*trigger event*).

- In the time interval before the trigger event, evoked activity variability is related to noise, since the effects of previous stimulation have ended.

Therefore, by comparing the evoked activity before and after the stimulus presentation it is possible to give a “score”to the reactivity of the recording. The following procedure for evaluating reactivity index was adopted:

1. For each stimulation (M,T or L) the average of the evoked activity in the interval from 30 to 10 *ms before* the trigger event has been evaluated (considering the absolute value)

2. The activity in the interval from 20 to 40 *ms after* the stimulus presentation have been evaluated (considering as before the absolute values, since polarities of 20 *ms* and 30 *ms* responses may be inverted and their effect could be reduced).

3. For each stimulation, the Reactivity index is evaluated as follows:

$$R_X = \sum_{t=20}^{40} |EA_X(t)| - \sum_{t=-30}^{-10} |EA_X(t)| \tag{8.2}$$

By means of R_X index it is therefore possible to evaluate the amount of reactivity of a MEG signal, making it possible to perform an independent component extraction that may take reactivity into account, as a criterion for task-related components identification.

Localization

To address the localization problem, i.e. to solve the *inverse problem*, several strategies may be employed. In fact, the problem of individuating the sources of an observed electromagnetic field has no unique solution and for this reason it is necessary to acquire supplementary information ([69]), i.e. to define the parameters of the “forward problem”, in order to obtain position, intensity and direction of the modeled cerebral currents. Several methodologies have been proposed in literature: in [153] a single and multiple dipole approach has been proposed, while in [128] and [127] MUSIC algorithm has been employed. In [138] the LOw Resolution Emission Tomography Analysis (LORETA) method has been developed; in [170] a Synthetic

Aperture Magnetometry (SAM) methodology has been employed; for a review, see [11].
In our case the localization was performed by means of moving equivalent current dipole (ECD) model inside a homogeneous best-fitted sphere. This procedure looks for a single primary density current $\mathbf{J}^p(\mathbf{r})$ such that the whole measured magnetic field is generated by it. The single current dipole can be expressed as:

$$\mathbf{q} = \int_V \mathbf{J}^P(\mathbf{r}')d\nu \tag{8.3}$$

With the approximation of the head as a conductive sphere, it is possible to localize the current dipole that has generated the magnetic field.
Coordinates are expressed in a coordinate system defined on the basis of three anatomical landmarks (x-axis passing through the two preauricolar point directed rightward, the positive y-axis passing through the nasion and the positive z-axis consequently). Therefore goodness-of-fit is evaluated for each dipole estimate and by means of a threshold (typically 80 %) a single model localization is accepted or rejected.

FastICA analysis of MEG dataset

The 15 subjects' recordings were analyzed with classical averaging methods and with FastICA algorithm ([79]), and the results were evaluated in terms of reactivity to the three stimuli and in terms of localization. For what concerns the localization of averaged signals, the moving equivalent current dipole was considered only in the most significant time points (20 and 30 ms after the trigger event).

Stimulation	Time point	Explained Variance E_V	x	y	z
L	22	0.812	-41	7	108
T	21	0.969	-37	10	95
M	19	0.977	-33	9	101

Table 8.1: Localization and goodness-of-fit for the activity at 20 ms using an ECD model on averaged recordings

The dipole estimated on independent components is time-independent, therefore it has been associated to a whole independent source. The explained variance of the dipole modeling is denoted with E_V, and a component with an explained variance

Stimulation	Time point	Explained Variance E_V	x	y	z
L	33	0.923	-25	2	110
T	34	0.968	-30	8	95
M	30	0.992	-34	7	104

Table 8.2: Localization and goodness-of-fit for the activity at 30 ms using an ECD model on averaged recordings

greater than 80 % is considered acceptable. Results of averaged model localization are in Tables 8.1 and 8.2. For what concerns the ICA analysis, the results have been evaluated in a slightly different way. In fact for each independent component, that can be related to one of the stimuli, reactivity index, with respect to the three experimental paradigms is evaluated, together with kurtosis (to evaluate the distance from a Gaussian distribution), and the localization parameters (coordinates and explained variance). Results for a single subject analysis are shown in Table 8.3. In this case *only* three independent sources were localized satisfactorily (the ninth component, although it had to be accepted because of the explained variance above threshold, has been discarded because the dipole localization led to a physiologically inconsistent position).
Results on different subjects, however, where not fully satisfactory. In fact, for some subjects, the number of components that could be localized was different from the number of stimulations (i.e. three). Moreover, reactivity indexes were not satisfactory, meaning that it has not been possible to discriminate between different stimulations across different components. In particular, for 6 subjects out of 15 it was possible to identify correctly the independent sources related to thumb and little finger stimulation, while for 9 subjects the two stimulations were mixed in a single component. A more detailed analysis will be presented in section 8.3.3.

8.3 Use of prior knowledge in MEG ICA

To overcome the problems encountered in FastICA extraction new procedures have been developed. Specific information on reactivity to one of the stimuli has been incorporated into ICA extraction. In fact, to avoid problems encountered in performing analysis on some subjects with classical ICA, reactivity index has been

Component	R_M	R_T	R_L	Kurtosis	E_V	x	y	z
1	9.29	2.56	4.68	5.1	0.984	-27	9	110
2	0.567	0.692	0.27	1.8	0.273			
3	15.47	8.5	6.05	3.91	0.1			
4	12.16	8.91	3.59	4.99	0.920	-50	-23	83
5	-0.12	0.41	1.94	3.38	0.778			
6	4.27	6.46	3.67	4.36	0.971	-45	16	75
7	1.11	1.89	0.32	3.37	0.677			
8	2.4	2.6	0.63	3.17	0.171			
9	1.54	0.47	0.32	3.13	0.802			
10	2.45	2.95	0.97	3.07	0.016			
11	0.55	-0.54	0.83	3.07	0.100			
12	0.7	-0.15	1.16	3.05	0			
13	0.96	0.26	0.97	3.05	0.050			
14	0.85	1.84	0.8	3.05	0.010			
15	2.65	1.05	-0.03	3.04	0			
16	0.67	0.64	0.24	3.03	0.480			

Table 8.3: Localization and reactivity for the first 16 FastICA components

accounted for while performing analysis. Two new procedures have been employed on the same dataset, and compared with averaging techniques and classical ICA both in terms of localization and of reactivity. In section 8.3.1 the first of the two techniques will be explained, while in section 8.3.2 the second procedure will be discussed. In section 8.3.3 results will be discussed.

8.3.1 Functional Component Analysis (FCA)

To identify neural networks devoted to individual finger central representation, the “reactivity”to the stimulus was incorporated into the separation algorithm. The reactivity evaluation is carried out by means of R_X index defined in eq. 8.2, here presented again:

$$R_X = \sum_{t=20}^{40} |EA_X(t)| - \sum_{t=-30}^{-10} |EA_X(t)| \,. \tag{8.4}$$

As said before, the subscript X denotes that the reactivity can evaluated for different stimulations.
The functional information is included into the extraction by means of the algorithm presented in 6.3, in the constrained version. Therefore negentropy has been maximized in a domain defined by the function $H(\mathbf{w})$, defined in (6.13).
The additional function considered in this analysis is:

$$H_X(\mathbf{w}) = \varphi\left(R_X(\mathbf{w}), \kappa\right), \tag{8.5}$$

where

$$\varphi\left(R_X(\mathbf{w}), \kappa\right) = \begin{cases} R_X\left(\mathbf{w}\right)/\kappa & \text{if} \quad R_X\left(\mathbf{w}\right) \leq \kappa \\ \kappa & \text{elsewhere} \end{cases} \tag{8.6}$$

And κ is the *minimum* allowed reactivity to the stimulus.
The optimization of $F + \lambda H$ using a *deflation approach* is carried out by means of simulated annealing, as seen in Chapter 6, since H is not differentiable and therefore gradient techniques cannot be employed.
The choice of parameters λ and κ is crucial to the effectiveness of the separation, and so is the order of functional constraints enforced. The following choices were made:

- $\boldsymbol{\lambda}$: to avoid setting the weight parameter for each subject by means of subsequent extractions (using the procedure explained in 6.5.2), we performed a preliminary analysis on a subject, and therefore observed that the value of negentropy were such that a choice of $\lambda = 1000$ was suitable to make the constraint effective without having a reduced precision for negentropy representation.

- $\boldsymbol{\kappa}$: the threshold of the constraint parameter was defined for each subject employing the following procedure: an extraction without thresholding is performed on data, an the global maximum of reactivity R_{Xmax} is found; subsequently κ parameter is chosen as $0.95R_{Xmax}$. This choice is motivated by the fact that, together with the desire of having a high value of reactivity, there is also the need for maximum independence within the domain defined by the contrast function.

- **Extraction Order:** due to the fact that sources are partially spatially overlapped and thumb and median spatial representations are larger than little finger one, the weaker sources were extracted first. This procedure has been motivated by the observation that extracting median or thumb component first would tend to affect the little finger extraction. In Tables 8.4, 8.5 and 8.6 some results on a subject with three possible extraction orders are depicted. It can be noted that in the first two case (M-L-T and T-L-M) it is not possible to localize the little finger component, and its reactivity is considerably lower than in the last case (L-T-M). The reason for this difference between orderings lies in the fact that orthogonalization process tends to cancel the little finger component that is partially overlapped to the thumb component and that is considerably weaker.

Component	Constraint	R_M	R_T	R_L	Kurtosis	E_V	x	y	z
1	Median	**22.74**	11.88	6.97	4.33	0.986	-33	7	101
2	Thumb	1.85	**9.92**	3,78	4.1	0.903	-48	17	108
3	Little Finger	0.6	0.35	**6.48**	3.4	0.609			

Table 8.4: FCA extraction using the order M-T-L

Component	Constraint	R_M	R_T	R_L	Kurtosis	E_V	x	y	z
1	Thumb	17.47	**15.54**	2.37	4.05	0.988	-32	14	97
2	Little Finger	10.81	0.987	**9.79**	3.71	0.249			
3	Median	**9.05**	0.601	2.53	4.03	0.843	-35	15	110

Table 8.5: FCA extraction using the order T-L-M

Component	Constraint	R_M	R_T	R_L	Kurtosis	E_V	x	y	z
1	Little Finger	16.1	5.01	**10.5**	3.83	0.883	-28	9	116
2	Thumb	13	**14.7**	0.294	3.89	0.986	-38	19	96
3	Median	**9.08**	0.717	2.44	3.92	0.83	-34	14	109

Table 8.6: FCA extraction using the order L-T-M

The implemented procedure has been called Functional component analysis (FCA) to point out that its aim is to recover *functional* components with prior knowledge on some of their features (in the present case it is reactivity to the stimulus). It is to be noted that in this approach independence of components is only enforced in a strongly constrained solution set, so that its role is significantly less important in the optimization than that of the constraint. This is also visible in the fact that obtained components are indeed not independent. For this reason, it is more appropriate to consider such approach as something different even form a constrained ICA, which motivates the choice for a different name. By means of this technique it has been possible to extract thumb related component for all the subjects, while for 4 subjects out of 15 thumb component localization failed.

8.3.2 Functional Source Separation (FSS)

Functional Component Analysis (FCA) has been proved effective in recovering functionally related sources. However, problems related to the decorrelation procedure has been encountered. In fact, since the three stimulations of the experiment were such that functional areas of the brain overlapped and one of them was considerably weaker than the other, a new procedure has been implemented, by removing the orthogonality constraint. Functional Source Separation (FSS) is therefore simular to FCA, with the difference that extracted components are not forced to be orthogonal. From an ICA point of view, this is a really strong relaxation of the basic assumptions, since extracted signals are not independent anymore. However, for the experiment considered, stimulus-related brain activities showed features that would have been *removed* by an independent extraction procedure, incorporating two stimulations into a single component.

The non-orthogonal approach has already been introduced in literature ([176],[179] and [83]), both from the a theoretical point of view, as a generalization of joint diagonalization procedures to the case of a non-orthogonal diagonalization matrix, and in applications, particularly in image modeling, to estimate overcomplete ICA bases. In this work, after the extraction of a signal, the next component is searched in a space “quasi-orthogonal”to the initial space. In fact, it could in principle happen that the new component extracted lies (almost) in the subspace spanned by the previous ones, so that it would not be significant. Therefore, it is necessary to enforce the condition that the angle between the new component and the said subspace be large enough, or to check such condition *a posteriori*.

FSS was able to extract all stimulus-related components for all the 15 subjects, and all of these sources were validated successfully as far as localization is concerned. The results of the analysis for the same subject of Tables 8.4-8.6 are depicted in 8.7. It is to be noted that the order of extraction in this case is arbitrary , since extraction for each functional source is performed on the whole dataset, each time with a different additional contrast function H. To test if the extraction process found every time a different component, the angles between the basis vectors couples of the component related to little finger stimulation FS$_L$ and the one related to thumb FS$_T$ were computed across subjects. The median angle between them was 63 °, with the interquartile range of the distribution being [42°;79°] and the minimum 25°.

Component	Constraint	R_M	R_T	R_L	Kurtosis	E_V	x	y	z
1	Little Finger	16.21	5.07	**10.47**	3.83	0.912	-22	10	112
2	Thumb	17.49	**15.54**	2.38	4.05	0.989	-31	14	96
3	Median	**22.74**	11.89	6.98	4.34	0.986	-33	6	101

Table 8.7: FSS separation evaluation on a subject

The evoked activities EA of the sources extracted with FSS are depicted in Figure 8.3. It has to be noted that the scale is different for each signal, but this has been done only for graphical reasons, since sources have no intrinsic amplitude. More details on the analysis will be presented in the next section.

8.3.3 Discussion

A comparison between classical ICA, FCA and FSS has been carried out on the 15 subjects, in terms of spatial localization and of reactivity. The analysis focused on thumb and little finger stimulation, since it would not be relevant a comparison in terms of spatial localization or reactivity for the median and the fingers since median stimulation induces activity also in fingers area. For each technique only the components whose localization was successful have been considered, therefore the number of subjects considered varies from a technique to another. In particular, FastICA ([79]) was able to detect correctly the activations in 6 subjects out of 15. For the remaining subjects, an independent source IC$_{T;L}$that accounted both for thumb and little finger representation has been considered, as it was not possible to separate them in two different components. Results of localization and reactivity index are

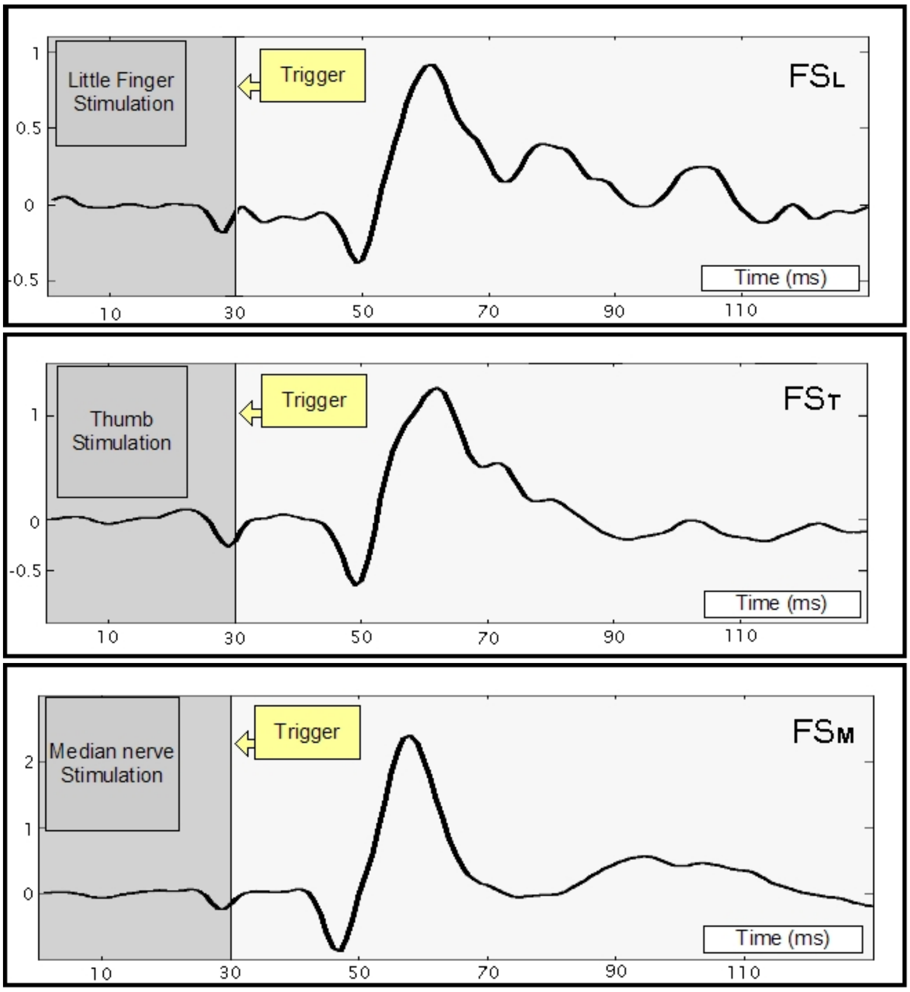


Figure 8.3: Evoked Activity for FSS components related to different stimulation accounted for in the algorithm

Sources		Sub.	Position (mm)				Reactivity		
			E_V	x	y	z	R_L	R_T	R_M
FSS	FS$_L$	15	0.95±0.04	-33±10	6±12.9	9±14	12.7±4.9	7.7±5.6	18.8±12
	FS$_T$	15	0.97±0.03	-38±10	10±13	90±10	6.3±5.1	13.4±4.8	18.3±12.6
FCA	FC$_L$	15	0.95±0.04	-33±10	5±12	99±13	12.7±4.9	7.7±5.5	18.9±12
	FC$_T$	11	0.95±0.05	-41±10	11±12	87±11	1.2±0.8	10.7±3.1	10.0±6.0
fast	IC$_L$	6	0.96±0.04	-36±12	9±23	96±11	7.5±5.9	4.4±5.1	11.4±11.3
ICA	IC$_T$	6	0.95±0.03	-41±14	16±43	79±20	2.5±3.9	6.3±2.7	11.1±14.2
	IC$_{T;L}$	9	0.93±0.08	-39±13	7±10	97±12	7.9±3.7	7.3±5.1	16.7±16.6

Table 8.8: **Spatial and reactivity characteristics of independent components**. Number of subject with successful localizations (> 80%, Successful cases), mean ± s.d., explained variance (E_V) and mean ± s.d., coordinates (x, y, z, millimetres) of ECDs; mean ± s.d. sources reactivity indexes to the three stimulations (R$_L$, R$_T$, R$_M$, pure numbers)

shown in Table 8.8. Localization obtained with classical averaging techniques in 20 and 30 ms after the trigger event are shown in Table 8.9.

To evaluate the level of residual finger response to the corresponding stimulation after sources extraction an index of discrepancy in the response was considered. The "discrepancy"index was defined as follows:

$$\text{Discr}_{R_X} = \frac{\sum_i (R_X^{MEG} - R_X^{MEG_{rec_Y}})^2}{\sum_i (R_X^{MEG})^2} \tag{8.7}$$

where R_X is defined as in (8.2), R_X^{MEG} is the reactivity index computed on MEG data during stimulation and $R_X^{MEG_{rec_Y}}$ is the reactivity index of reconstructed MEG data with the Y finger (Y=L,T) data during its corresponding stimulation; the index i runs upon the 4 channels of minimal and maximal amplitude at M20 and M30 latencies. In fact, the dipole field distributions generated at these peak latencies are well described by their minimum/maximum values ([163]). Discrepancy values for the three techniques are presented in Table 8.10, left panel, together with the results of ANOVA test for the corresponding contrast.

Discrepancy reactivity resulted significantly lower for the FSS procedure with respect to both its orthogonal version FCA and FastICA, indicating its more satisfactory performance in extracting activity of interest; instead, not significant difference in Discr$_{R_T}$ mean values was found between FCA and FastICA. Low mean discrepancy reactivity values for FSS (6% of residual response for the little finger

Sources		Subjects	Position (mm)			
			E_V	x	y	z
MEG Data	M20$_L$	12	0.94±0.06	-34±9	7±14	99±9
	M20$_T$	15	0.96±0.02	-42±8	11±11	91±10
	M30$_L$	11	0.97±0.02	-31±8	4±10	97±9
	M30$_T$	12	0.97±0.03	-33±8	6±13	89±12

Table 8.9: **Spatial and reactivity characteristics of averaged signals**. Number of subject with successful localizations ($>$ 80%, Successful cases), mean ± s.d., explained variance (E_V) and mean ± s.d., coordinates (x, y, z, millimetres) of ECDs.

Method	**Mean ± s.d.**		**Contrast**	p	
	Discr$_{R_L}$	Discr$_{R_T}$		Discr$_{R_L}$	Discr$_{R_T}$
FCA	0.06±0.07	0.22±0.21	FCA *vs* FSS	-	0.04
FSS	0.06±0.07	0.03±0.04	FCA *vs* FastICA	0.004	0.3
FastICA	0.28±0.27	0.32±0.28	FSS *vs* FastICA	0.004	0.001

Table 8.10: Discrepancy response levels. Mean ± s.d. for the two finger discrepancy indexes in the three algorithms. Results of ANOVA test (with Bonferroni correction for multiple comparison in the thumb case) are summarized for the corresponding contrast in the p values column.

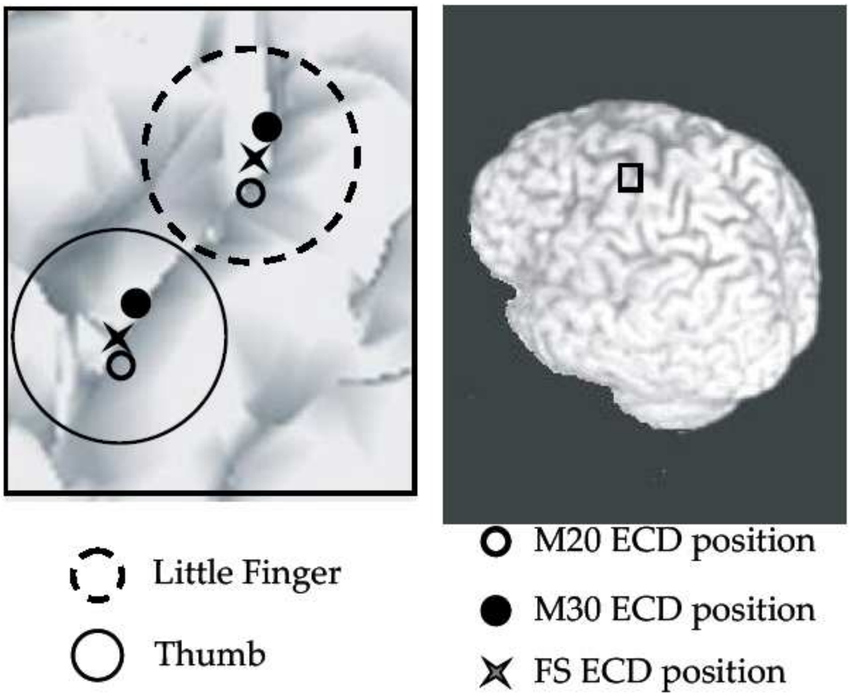


Figure 8.4: Positions in one representative subject of thumb (solid circle area) and little finger (dotted circle area) sources. M20 (empty circle), M30 (filled circle) and the extracted source with the FSS procedure (star).

and 3% for the thumb with respect to the original averaged MEG data) indicated that the two extracted finger sources described practically all the evoked response contained in the original data matrix.

To have a benchmark for finger somatosensory source position, known markers of signal arrival in the primary sensory cortex, occurring at around 20 and 30 ms from the stimulus were calculated by averaging original MEG channel signals and computing corresponding ECDs. In Figure 8.4 sources recovered with FSS and related to thumb and little finger are depicted, together with M20 and M30 localization. It could be noted from that FS_L and FS_T lie in-between their respective M20 and M30 positions, in agreement with the constraint time window definition (see equation (8.2)), that includes both M20 and M30 latencies.

Chapter 9

Conclusions and Perspectives

In this work, development and implementation of new Blind Source Separation (BSS) algorithms has been investigated. Such algorithms are capable of taking known features of the signals that should be retrieved into account, while preserving *blindness* of the approach to avoid introducing unwanted constraints that may cause artifacts or prevent the relevant solutions to be obtained.

While many applications of ICA are amenable to off-line computation, in some cases fast computation is required because of real-time constraints, or just because the large quantity of data would delay results excessively. For this reason, implementation of the FastICA algorithm on floating and fixed point DSP architectures has been investigated. The results obtained in Chapter 4 indicate that a floating-point architecture, due to the large amount of time needed for extract independent components, is not a suitable choice. On the other hand, a fixed-point architecture seems more able to extract independent sources in an amount of time that is comparable to that of ordinary general-purpose microprocessors. These results can be used as the basis for development of embedded solutions for ICA computations and of multi-processor multi-threading architectures, that may significantly outperform ordinary microprocessors in terms of speed and of portability.

A new relevant application on ICA is considered in this work concerning correlated equivalent-circuit-based statistical models (ECM) of electronic devices. In fact in Chapter 5 a linear decomposition of circuital correlated parameters in *independent* components has been compared with a decomposition in *uncorrelated* ones. This choice has been motivated by the fact that circuit parameters do not have in general a Gaussian distribution, and therefore higher order statistics contain information that is usually discarded while employing Principal Component Analysis. It has been shown that a ICA-based model simulation is more accurate than a PCA-based one in describing autocorrelations of the device population, and in dealing with the more sensitive elements of the scattering matrix, where the other methods fail to represent data correctly, but on the other hand some deteriorations in overall accuracy of S-parameters has been observed. This analysis has been performed by approximating the independent component distribution with log-normal ones, that exhibited some drawbacks in approximating data distribution. These deteriorations may be avoided by means of different approximations of independent components densities by means of Gaussian Mixture Models (GMM), that can be easily implemented in CAD simulation tools. Another important issue that needs to be addressed is the possibility of performing a non-linear mixing model, to account for the relationship between the physical parameters and the circuital ones; Non-linear ICA, that has been recently studied, could be promising in yield optimization design, providing a physical level variability estimation by means of circuital parameters observation.

New methodologies have been developed in this work to order to incorporate prior information in ICA analysis. In fact, in real world problems where ICA is usually employed, additional information on the structure of the data is generally available, but neglected by current BSS algorithms. The new methodology proposed is extremely flexible and it has been designed to account for specific and loose prior information on the sources. In fact, by means of an additional contrast function it is possible both to constrain some solutions of an ICA problem to lie in a particular region of the multidimensional optimization landscape pointed out by prior information and to account for loose information on all the components, by perturbing the classical ICA solution. In Chapter 6 the principles of this technique have been depicted together with some theorems that guarantee that the constrained optimum is reached by means of penalty terms. The choice of Simulated Annealing optimization comes from the need to deal with prior information of various kinds, that can therefore also be non-differentiable. Simulated Annealing has the double advantage of not requiring derivatives and of reaching the global optimum, even if it is slower than gradient-based techniques.

The effectiveness and flexibility of this approach has been proved on both fMRI and MEG data. In fMRI data analysis, spatial and temporal constraints have been added to independent component separation, allowing to recover a source directly avoiding post-selection. Moreover, the constrained approach has been investigated on a set of finger-tapping experiments, proving its effectiveness in extracting the target source by means of some knowledge on its time course, like correlation with the experimental paradigm or a specific frequency content. In this case it has been shown that it is possible to outperform classical ICA approach by enhancing a specific feature of one of the independent components.

Moreover, a new approach has been developed in order to take spatio-temporal regularities into account. Classical ICA methods do not consider ordering of the points in signals, while it is known that cerebral activities show regularities in both space and time. Therefore the proposed methodology has been employed with an additional contrast function regarding spatial and temporal autocorrelation, and it has been tested on a simulated fMRI dataset and on a real fMRI resting state experiment with superimposed artificial activations. It has been shown that, by means of this additional contrast function, source extraction is dramatically improved especially in high noise environments. However the choice of the weight of the additional contrast function needs further investigation, also in a Bayesian framework fashion, to determine which is the optimal weighting for prior information term. Another important issue that is going to be addressed is an implementation of this algorithm on the cerebral cortex, rather than on the whole volume. In fact, neighborhood relationships are better pointed out on the cortex, since two neighboring voxels in the volume may not represent neighboring cerebral areas.

Prior information has been also employed in MEG data analysis. In Chapter 8 it has been shown that constraining sources by means of a reactivity measure helps improving both localization and interpretability of the independent components and two new algorithms (FCA and FSS) have been proposed and compared to classical ICA algorithms.

The proposed methodology has proved to be flexible and accurate, as it has been applied successfully in both constrained and multi-objective case to fMRI and MEG data analysis. Many of the solutions proposed in literature can be obtained by means of the proposed approach, that is therefore more general and can account also for information that cannot be expressed by means of a differentiable function. In particular, the proposed algorithm can be employed also in audio signal separation (one example is the classical cocktail party problem), where very specific information is available on the sources. Moreover, investigations on overcomplete ICA (i.e. when there are more sources than sensors) are currently under way. In fact, the overcomplete case is ill-posed due to the non square mixing process, but available information may help in recovering the sources effectively.

Bibliography

[1] E.H.L. Aarts and P.J.M. van Laarhoven. *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers, 1987.

[2] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763, 1996.

[3] S.I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):1875–1883, 1998.

[4] S.I. Amari, T.P. Chen, and A. Cichocki. Stability analysis of adaptive blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

[5] S.I. Amari and A. Cichocki. *Adaptive Blind Signal and Image Processing - Learning algorithms and applications*. John Wiley & Sons, 2002.

[6] I. Angelov, H. Zirath, and N. Rorsman. A new empirical nonlinear model for HEMT and MESFET devices. *IEEE Transactions on Microwave Theory and Techniques*, pages 2258–2266, 1992.

[7] R. Anholt and R. Neidhard. Statistical analysis of GaAs MESFET S-parameter equivalent-circuit models. *International Journal of Microwave Millimeter-Wave Computer-Aided Engineering*, 1(3):263–270, 1991.

[8] B.A. Ardekani and I. Kanno. Statistical methods for detecting activated regions in functional MRI of the brain. *Magnetic Resonance in Imaging*, 16(10):1217–1225, 1998.

[9] K. Arfanakis, D. Cordes, V.M. Haughton, C.H. Moritz, M.A. Quigley, and M.E. Meyerand. Combining independent component analysis and correlation

analysis to probe interregional connectivity in fMRI task activation datasets. *Magnetic Resonance Imaging*, 18:921–930, 2000.

[10] K. Bacher, S. Massie, D. Hartzel, and T. Stewart. Present ability of commercial molecular beam epitaxy. In *International conference on Indium Phosphide and Related Materials*, pages 351–352, 1997.

[11] S. Baillet, J.C. Mosher, and R.M. Leahy. Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30, 2001.

[12] P.A. Bandettini, A. Jesmanowicz, E.C. Wong, and J.S. Hyde. Processing strategies for time-course data sets in functional MRI if the human brain. *Magnetic Resonance in Medicine*, 30:161–173, 1993.

[13] P.A. Bandettini and E.C. Wong. Effects of biophysical and physiologic parameters on brain activation-induced R2* and R2 changes: simulations using a deterministic diffusion model. *International Journal of Imaging System and Technology*, 6:133–152, 1995.

[14] P.A. Bandettini, E.C. Wong, R.S. Hinks, and R.S. Tikofsky. Time course epi of human brain function during task activation. *Magnetic Resonance in Medicine*, 25:390–398, 1992.

[15] P.A. Bandettini, E.C. Wong, A. Jesmanowicz, R.S. Hinks, and J.S. Hyde. Spin-echo and gradient-echo EPI of human brain activation using BOLD contrast: a comparative study at 1.5 T. *NMR in Biomedicine*, 7:12–20, 1994.

[16] J.W. Bandler, R. M. Biernacki, Q. Cai, S.H. Chen, S. Ye, and Q.-J Zhang. Integrated physics-oriented statistical modeling, simulation and optimization. *IEEE Transaction on Microwave Theory and Techniques*, 40(7):1374–1400, 1992.

[17] G. Barbati, C. Porcaro, F. Zappasodi, and F. Tecchio. An ICA approach to detect functionally different intra-regional neuronal signals in MEG data. *Clinical Neurophysiology*, 115:1220–1232, 2004.

[18] G. Barbati, C. Porcaro, F. Zappasodi, and F. Tecchio. An ICA approach to detect functionally different intra-regional neuronal signals in MEG data. In *Computational Intelligence and Bioinspired Systems Lecture Notes in Computer Science (LNCS 2512) , IWANN '05 Barcelona*, pages 1083–1090, 2005.

[19] A.K. Barros and A. Cichocki. Extraction of specific signals with temporal structure. *Neural Computation*, 13(9):1995–2003, 2001.

[20] K. Baudendistel, L.R. Schad, M. Friedlinger, F. Wenz, J. Schroeder, and W. Lorenz. Postprocessing of functional MRI data of motor cortex stimulation measured with a standard 1.5 T imager. *Magnetic Resonance Imaging*, 13:701–707, 1995.

[21] R. Baumgartner, L. Ryner, W. Richter, R. Summers, M. Jarmasz, and R. Somorjai. Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component. *Magnetic Resonance Imaging*, 18:89–94, 2000.

[22] C.F. Beckmann. Probabilistic ICA for fMRI. In *Proceedings of the IEEE Symposium on Biomedical Imaging: Macro to Nano*, 2004.

[23] C.F. Beckmann and S.M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on medical imaging*, 23(2), 2004.

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

[25] J.W. Belliveau, D. Kennedy, R.C. McKinstry, B.R. Buchbinder, R.M. Weiskoff, M.S. Cohen, J.M. Vevea, T.J. Brady, and B.R. Rosen. Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254:716–719, 1991.

[26] A. Belouchrani, K. Abed Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

[27] H. Berger. Über das elektroenkephalogram des menschem. *Archiv für Psychiatrie und Nervenkrankheiten*, 87:527–570, 1929.

[28] D.P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, 1982.

[29] B.B. Biswal and J.L. Ulmer. Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *Journal of Computer Assisted Tomography*, 23:265–271, 1999.

[30] A.M. Blamire, S. Ogawa, and K. Ugurbil. Echo Planar Imaging of the activated human visual cortex shows a time delay between stimulus and activation. In *Proceedings of the 11th annual meeting of the Society of Magnetic Resonance in Medicine*, page 1823, 1992.

[31] G. Boynton, S.A. Engel, G.H. Glover, and D.J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *The Journal of Neuroscience*, 16(13):4207–4221, 1996.

[32] V.D. Calhoun, T. Adali, V.B. McGinty, J.J Pekar, T.D. Watson, and G.D. Pearlson. fMRI activation in a visual-perception task: network and areas detected using the general linear model and independent components analysis. *NeuroImage*, 14:1080–1088, 2001.

[33] V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14:140–151, 2001.

[34] V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*, 13:43–53, 2001.

[35] V.D. Calhoun, T. Adali, J.J. Pekar, and G.D. Pearlson. Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ICA analysis. *Neuroimage*, 25(2):527–538, 2005.

[36] V.D. Calhoun, J.J. Pekar, V.B. McGinty, T. Adali, T.D. Watson, and G.D. Pearlson. Different activation dynamics in multiple neural systems during simulated driving. *Human Brain Mapping*, 12:172–193, 2002.

[37] J.-F. Cardoso, C. Jutten, , and P. Loubaton, editors. *Proceedings of the 1st Int. Workshop on Independent Component Analysis and Signal Separation.* 1999.

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

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

[40] J. Carroll, K. Whelan, S. Prichett, and D.R. Bridges. FET statistical modeling using parameter orthogonalization. *IEEE Transactions on Microwave Theory and Techniques*, 1(3):47–54, 1996.

[41] M. Castelo-Branco, E. Formisano, E. Backes, F. Zanella, S. Neuenschwander, W. Singer, and R. Goebel. Activity patterns in human motion-sensitive areas depend on the interpretation of the global motion. *Proceedings of the National Academy of Science USA*, 99:13914–13919, 2002.

[42] F. Centurelli, A. Di Martino, G. Scotti, P. Tommasino, and A. Trifiletti. A new procedure for non-linear statistical model extraction of GaAs FET integrated circuits. *International Journal of RF and Microwave Computer-Aided Engineering*, 13(5):348–356, 2003.

[43] A. Cichocki and R. Unbehauen. Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Trans. on Circuits and Systems*, 43(11):894–906, 1996.

[44] D. Cohen. Magnetoencephalography: Evidence of magnetic fields produced by alpha-rhythm currents. *Science*, pages 784–786, 1968.

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

[46] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[47] W.R. Curtice. A MESFET for use in the design of GaAs integrated circuits. *IEEE Transaction on Microwave Theory and Techniques*, 28(5), 1980.

[48] K.I. Diamantaras and S.Y. Yung. *Principal Component Neural Networks, Theory and Applications.* John Wiley & Sons, 1996.

[49] S. Dodel, M. Herrmann, and T. Geisel. Localization of brain activity – blind separation for fMRI data. *Neurocomputing*, 32-33:701–708, 2000.

[50] R. Dymond and D.G. Norris. Mechanism and echo time dependence of the fast response in fMR. *Magnetic Resonance in Medicine*, 38:1–6, 1997.

[51] J. Eriksson and V. Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

[52] T. Ernst and J. Ennig. Observation of a fast response in functional MR. *Magnetic Resonance in Medicine*, 33:636–647, 1994.

[53] F. Esposito, E. Formisano, E. Seifritz, R. Goebel, R. Morrone, G. Tedeschi, and F. Di Salle. Spatial independent component analysis of functional MRI time-series: to what extent do results depend on the algorithm used? *Human Brain Mapping*, 16:146–157, 2002.

[54] F. Esposito, T. Scarabino, A. Hyvärinen, J. Himberg, E. Formisano, S. Comani, G. Tedeschi, R. Goebel, E. Seifritz, and F. Di Salle. Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage*, 25:193–205, 2005.

[55] F. Esposito, E. Seifritz, E. Formisano, R. Morrone, T. Scarabino, G. Tedeschi, S. Cirillo, R. Goebel, and F. Di Salle. Real-time independent component analysis of fMRI time-series. *NeuroImage*, 20:2209–2224, 2003.

[56] E. Formisano, F. Esposito, N. Kriegeskorte, G. Tedeschi, F. Di Salle, and R. Goebel. Spatial independent component analysis of functional magnetic resonance imaging time-series: characterization of the cortical components. *Neurocomputing*, 49:241–254, 2002.

[57] E. Formisano, F. De Martino, F. Gentile, M. Balsi, F. Esposito, F. Di Salle, and R. Goebel. Classification of fMRI-independent components in a multidimensional feature space using least-square support vector machines. In *Human Brain Mapping (HBM) International Conference*, 2004.

[58] P.T. Fox and M.E. Raichle. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Science USA*, 83:1140–1144, 1986.

[59] O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Exploratory fMRI analysis by autocorrelation maximization. *NeuroImage*, 16:454–464, 2002.

[60] K.J. Frison. Modes of models: a critique on independent component analysis for fMRI. *Trends in Cognitive Sciences*, 2(10), 1998.

[61] K.J. Friston, J. Ashburner, C.D. Frith, J.B. Poline, J.D. Heather, and R. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 2:165–189, 1995.

[62] K.J. Friston, C. Frith, P.F. Liddle, R. Dolan, A.A. Lammertsma, and R.S.J. Frackowiak. The relationship between global and local changes in PET scans. *Journal of Cerebral Blood Flow and Metabolism*, 10:458–466, 1990.

[63] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C.D. Frith, and R. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[64] K.J. Friston, P. Jezzard, and R. Turner. The analysis of functional MRI time-series. *Human Brain Mapping*, 1:153–171, 1995.

[65] K.J. Friston, S. Williams, R. Howard, R. Frackowiak, and R. Turner. Moment related effects in fMRI time-series. *Magnetic Resonance in Medicine*, 35:346–355, 1996.

[66] M. Gaeta and J.L. Lacoume. Source separation without prior knowledge: the maximum likelihood solution. In *Proc. EUSPICO*, pages 621–624, 1990.

[67] I. Gerace, F. Cricco, and A. Tonazzini. An extended maximum likelihood approach for the robust blind separation of autocorrelated images from noisy mixtures. In *Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004*, pages 954–961. Springer, 2004.

[68] R.P. Giffard. Fundamentals for SQUID applications. In *SQUID'80: Proceedings of the 2nd International Conference on SQUID Devices and their Applications*, pages 445–471, 1980.

[69] C. Del Gratta, V. Pizzella, F. Tecchio, and G.L. Romani. Magnetoencephalography - a non invasive brain imaging method with 1 ms time resolution. *Reports on Progress in Physics*, 64:1759–1814, 2001.

[70] A. Grinvald and R.D. Frostig. High resolution optical imaging of functional brain architecture in the awake monkey. *Proceedings of the National Academy of Science USA*, 88:11559–11563, 1991.

[71] J.V. Hajanl, N. Saeed, E.J. Oatridge, I.R. Young, and G.M. Bidder. A registration and interpolation procedure for subvoxel matching of serially acquired MRI images. *Journal of Computer Assisted Tomography*, 19:289–296, 1995.

[72] H. Happy and A. Cappy. *HELENA: HEMT electrical properties and noise analysis, software and users' manual.* Artech House, London, U.K., 1993.

[73] J. Himberg, A. Hyvärinen, and F. Esposito. Validating the independent components of neuroimaging time-series via clustering and visualization. *Neuroimage*, 22:1214–1222, 2004.

[74] A.P. Holmes, J.B. Poline, and K.J. Friston. Characterizing brain images with the general linear model. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, and J.C. Mazziotta, editors, *Human Brain Function*, pages 59–84. Academic Press USA, 1997.

[75] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. of Education Psychology*, 24:417–441, 1933.

[76] A.M. Howseman, D.A. Porter, C. Hutton, O. Josephs, and R. Turner. Blood oxygenation level dependent signal time courses during prolonged visual stimulation. *Magnetic Resonance Imaging*, 16(1):1–11, 1998.

[77] X. Hu, T.H. Lee, and K. Ugurbil. Evaluation of the early response in fMRI in individual subjects using short stimulus duration. *Magnetic Resonance in Medicine*, 37:877–884, 1997.

[78] J.S. Hyde and B.B. Biswal. Functionally related correlation in the noise. In C.T. Moonen and P.A Bandettini, editors, *Functional MRI*, pages 263–275. Springer-Verlag, Berlin, 2000.

[79] A. Hyvärinen. Fast and robust fixed point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[80] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[81] J. Igual, L. Vergara, A. Camacho, and R. Miralles. Independent component analysis with prior information about the mixing matrix. *Neurocomputing*, 50:419–438, 2003.

[82] S. Ikeda and K. Tomaya. Independent Component Analysis for noisy data - MEG data analysis. *Neural Networks*, 13:1063–1074, 2000.

[83] M. Inki and A. Hyvärinen. Two methods for estimating overcomplete independent component bases. In *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, 2001.

[84] R.W. Liu J. Luo, X.T. Ling. Principal independent component analysis. *IEEE Transactions on Neural Networks*, 10(4):912–917, 1999.

[85] C.J. James and O.J. Gibson. Temporally constrained ICA: an application to artifact rejection in electromagnetic brain signal analysis. *IEEE Transactions on Biomedical Engineering*, 50(9):1108–1116, 2003.

[86] C.J. James and C.W. Hesse. Independent Component Analysi for biomedical signals. *Physiological Measurement*, pages 15–39, 2005.

[87] T.P. Jung, S. Makeig, M.J. McKeown, A.J. Bell, and T.W. LEE T.J. Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–1022, 2001.

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

[89] J. Karhunen and P. Pajunen. Blind source separation using least-squares type adaptive algorithms. In *Proc. of IEEE int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97)*, pages 3361–3364, 1997.

[90] J. Karhunen, P. Pajunen, and E. Oja. The nonlinear PCA criterion in blind source separation: Relations with other approaches. *Neurocomputing*, 22:5–20, 1998.

[91] S.G. Kim, W. Richter, and K. Ugurbil. Limitations of temporal resolution in functional MRI. *Magnetic Resonance in Medicine*, 37:631–636, 1997.

[92] Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *Proceedings of Int. Conf. on Neural Information Processing (ICONIP'98)*, volume 2, pages 895–898, 1998.

[93] K.H. 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, 1999.

[94] S. Krupenin, R.R. Blanchard, M.H. Somerville, J.A. del Alamo, K.G. Duh, and P.C. Chao. Physical mechanisms limiting the manufacturing uniformity of millimeter-wave power InP HEMT's. *IEEE Transactions on Electronic Devices*, 47(8):1560–1565, 2000.

[95] S.M. Kuo and B. H. Lee. *Real-Time Digital Signal Processing.* John Wiley & Sons Ltd, 2001.

[96] S. Kuroda, K. Imanishi, N. Harada, K. Hikosaka, and M. Abe. Highly uniform N-InAlAs/InGaAs HEMT's on a 3-in InP substrate using photochemical selective dry recess etching. *IEEE Electron Devices Letters*, 13(2):105–107, 1992.

[97] K.K. Kwong. Functional magnetic resonance imaging with echo planar imaging. *Magnetic Resonance Quarterly*, 11(1):1–20, 1995.

[98] K.K. Kwong, J.W. Belliveau, I.E. Goldberg, R.M. Weiskoff, B.P. Poncelet, D.N. Kennedy, B.E. Hoppel, M.S. Cohen, R. Turner, H.M. Cheng, T.J. Brady, and B.R. Rosen. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Science USA*, 89:5674–5679, 1992.

[99] N. Lange, S.C. Strother, J.R. Anderson, F.A. Nielsen, A.P. Holmes, T. Kolenda, R. Savoy, and L.K. Hansen. Plurality and resemblance in fMRI data analysis. *NeuroImage*, 10:282–303, 1999.

[100] N.A. Lassen. The metabolic and hemodynamic events secondary to functional activation. *Magnetic Resonance in Medicine*, 38:521–523, 1997.

[101] P.C. Lauterbur. Image formation by induced local interactions - examples employing nuclear magnetic-resonance. *Nature*, 242:190–191, 1973.

[102] 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):409–433, 1999.

[103] M. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2), 2000.

[104] R. Liao, J.L. Krolik, and M.J. McKeown. An information-theoretic criterion for intrasubject alignment of fMRI time series: motion corrected independent component analysis. *IEEE Transactions on Medical Imaging*, 24(1):29–44, 2005.

[105] X. Liao and L. Carin. A new algorithm for independent component analysis with or without constraints. In *Proceedings of Sensor Array and Multichannel Signal Processing Workshop*, pages 413–417, 2002.

[106] N.K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:150–157, 2001.

[107] M. Lowe and J. Sorensen. Quantitative comparison of functional contrast from BOLD-weighted spin-echo and gradient-echoplanar imaging at 1.5 Tesla and $H_2^{15}O$ PET in the whole brain. *Magnetic Resonance in Medicine*, 37:723–729, 1997.

[108] W. Lu and J.C. Rajapakse. Eliminating indeterminacy in ICA. *Neurocomputing*, 50:271–290, 2003.

[109] W. Lu and J.C. Rajapakse. Approach and applications of constrained ICA. *IEEE Transactions on Neural Networks*, 16(12):203–212, 2005.

[110] A.S. Lukic, M.N. Wemick, L.K. Hansen, J. Anderson, and S.C. Strother. A spatially robust ICA algorithm for multiple fMRI data sets. In *IEEE International Symposium on Biomedical Imaging*, pages 839–842, 2002.

[111] V.K. Madisetti and D.B. Williams. *Digital Signal Processing Handbook*. Chapman & Hall/CRC Press, 1999.

[112] S. Makeig, A.J. Bell, T.P. Jung, and T.J. Sejnowski. Independent component analysis of electroencephalographic data. *Advances of Neural Information Processing Systems*, 8:145–51, 1996.

[113] S. Makeig, S. Debener, J. Onton, and A. Delorme. Mining event-related brain dynamics. *Trends in Cognitive Science*, 8(5):204–210, 2004.

[114] D. Malonek and A. Grinvald. Interactions between electrical activity and cortical microcirculation revealed by optical imaging spectroscopy: implication for functional brain imaging. *Science*, 272:551–4, 1996.

[115] J. Mateos, D. Purdo, T. Gonzales, P. Tadyszak, F. Danneville, and A. Cappy. Influence of mole fraction on the noise performance of GaAs/Al$_x$Ga$_{1-x}$As HEMT's. *IEEE Transactions on Electronic Devices*, 45(9):2081–2083, 1998.

[116] M.J. McKeown. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *NeuroImage*, 11:24–35, 2000.

[117] M.J. McKeown, L.K. Hansen, and T.J. Sejnowski. Independent component analysis of functional MRI: What is signal and what is noise? *Current Opinion in Neurobiology*, 13(5):620–629, 2003.

[118] M.J. McKeown and T.J. Sejnowski. Independent component analysis of fmri data: examining the assumptions. *Human Brain Mapping*, 6:368–372, 1998.

[119] M.J. McKeown, T.J. Sejnowski, G.G. Brown, S.S. Kindermann, A.J. Bell, and T.J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6:160–188, 1998.

[120] M.J. McKeown, V. Varadarajan, S. Huettel, and G. McCarthy. Deterministic and stochastic features of fMRI data: implications for analysis of event-related experiments. *Journal of Neuroscience Methods*, 118:103–113, 2002.

[121] R.S. Menon and S.G. Kim. Spatial and temporal limits in cognitive neuroimaging with fMRI. *Trends in Cognitive Sciences*, 3:207–216, 1999.

[122] R.S. Menon, S. Ogawa, J.S. Strupp, P. Andersen, and K. Ugurbil. BOLD-based functional MRI at 4 Tesla includes a capillart bed contribution: echo-

planar imaging correlates with previous optimal imaging using intrinsic signals. *Magnetic Resonance in Medicine*, 33:453–459, 1995.

[123] P.P. Mitra, S. Ogawa, X. Hu, and K. Ugurbil. The nature of spatiotemporal changes in cerebral hemodynamics as manifested in Functional Magnetic Resonance Imaging. *Magnetic Resonance in Medicine*, 37:511–518, 1997.

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

[125] J.E. Moran, D.L. Drake, and N. Tepley. Method for MEG imaging. *Neurology and Clinical Neurophysiology*, 72, 2004.

[126] C.H. Moritz, B.P. Rogers, and M.E. Meyerand. Power spectrum ranked independent component analysis of a periodic fMRI complex motor paradigm. *Human Brain Mapping*, 18:111–122, 2003.

[127] J.C. Mosher and R.M. Leahy. Source localization using Recursively Applied and Projected (RAP) Music. *IEEE Transactions on Signal Processing*, 47:332–345, 99.

[128] J.C. Mosher, P.S. Lewis, and R.M. Leahy. Multiple dipole modeling and localization from spatiotemporal MEG data. *IEEE Transations on Biomedical Engineering*, 39:541–557, 1992.

[129] D. Obradovic and G. Deco. Information maximization and independent component analysis: Is there a difference? *Neural Computation*, 10(8):2085–2101, 1998.

[130] S. Ogawa, T. Lee, A.R. Kay, and D.W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Science*, 89:9868–9872, 1990.

[131] S. Ogawa, T.M. Lee, A.S. Nayak, and P. Glynn. Oxygenation sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14:68–78, 1990.

[132] S. Ogawa, D. Tank, R. Menon, J.M. Ellermann, S.G. Kim, K. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping using MRI. *Proceedings of the National Academy of Science USA*, 89:5951–5955, 1992.

[133] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17:25–45, 1997.

[134] W. Orrison, J.D. Lewine, J. Sanders, and M.Hartshorne. *Functional Brain Imaging.* Mosby, 1995.

[135] P. Pajunen and J. Karhunen, editors. *Proceedings of the 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation.* 2000.

[136] C. Papathanassiou and M. Petrou. Incorporating prior knowledge in ica. In *Proceedings of 14th International Conference in Digital Signal Processing, DSP 2002*, volume 2, pages 761–764, 2002.

[137] A. Papoulis. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, third edition, 1991.

[138] R.D. Pascual-Marqui, C. M. Michel, and D. Lehman. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, 18:49–65, 1995.

[139] L. Pauling and C.D. Coryell. The magnetic properties and structure of hemoglobin, oxyhemoglobin, and carbonmoxy-hemoglobin. *Proceedings of the National Academy of Science, USA*, 22:210–216, 1936.

[140] B. Pearlmutter and L. Parra. A context-sensitive generalization of independent component analysis. In *Proceedings of the International Conference on Neural Information Processing ICONIP '96*, 1996.

[141] B.A. Pearlmutter and L.C. Parra. Maximum likelihood blind source separation: A context sensitive generalization of ICA. *Advances in Neural Information Processing Systems*, 9:613–619, 1997.

[142] K. Pearson. On lines and planes of closest fit to system points in space. *Philosophical Magazine*, 2:559–572, 1901.

[143] K. Petersen, L. K. Hansen, T. Kolenda, and E. Rostrup. On the independent components of functional neuroimages. In *Third International Conference on Independent Component Analysis and Blind Source Separation*, pages 615–620, 2000.

[144] K.M. Petersson, T.E. Nichols, J. Poline, and A.P. Holmes. Statistical limitations in functional neuroimaging, Part I: non-inferential methods and statistical models. *Philosophical Transaction of The Royal Society B: Biological Sciences*, 354:1239–1260, 1999.

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

[146] J. Purviance, D. Criss, and D. Monteith. FET model statistics and their effects on design centering and yield prediction for microwave amplifiers. In *Proceedings of IEEE MTT Symposium*, volume 1, pages 315–318, 1998.

[147] M.A. Quigley, V.M. Haughton, J. Carew, D. Cordes, C. H. Moritz, and M. E. Meyerand. Comparison of independent component analysis and conventional hypothesis-driven analysis for clinical functional MR image processing. *American Journal of Neuroradiology*, 23:49–58, 2002.

[148] J.C. Rajapakse, A. Cichocki, and K. Siwek. Non independent strategies for blind source separation in functional mri. In *Proceedings of the 9th International Conference on Neural Information Processing, ICONIP 02*, volume 4, pages 1654–1660, 2002.

[149] T. Ristaniemi and J. Joutsensalo. On the performance of blind source separation in CDMA downlink. In J.-F. Cardoso, C. Jutten, , and P. Loubaton, editors, *Proceedings of the 1st Int. Workshop on Independent Component Analysis and Signal Separation*, pages 437–441, 1999.

[150] B.R. Rosen, J.W. Belliveau, H.J. Aronen, D. Kennedy, B.R. BuchBinder, A. Fischman, M. Gruber, J. Glas, R.M. Weisskoff, and M.S. Cohen. Susceptibility contrast imaging of cerebral blood volume: human experience. *Magnetic Resonance in Medicine*, 22(2):293–299, 1991.

[151] S.M. Ross. *Introduction to Probability Models*. Academic Press, sixth edition, 1997.

[152] F. Di Salle, E. Formisano, D.E.J. Linden, R. Goebel, S. Bonavita, A. Pepino, F. Smaltino, and G. Tedeschi. Exploring brain function with magnetic resonance imaging. *European Journal of Radiology*, 30:84–94, 1999.

[153] M. Scherg and P. Berg. Use of prior knowledge in brain electromagnetic source analysis. *Brain Topography*, 4:143–150, 1991.

[154] J.V. Stone. Blind source separation using temporal predictability. *Neural Computation*, 13(7):1559–1574, 2001.

[155] J.V. Stone and J. Porrill. Regularisation using spatiotemporal independence and predictability. Computational Neuroscience Report 201, University of Sheffield, 1999.

[156] J.V. Stone, J. Porrill, N.R. Porter, and I.W. Wilkinson. Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. *Neuroimage*, 15(2):407–421, 2002.

[157] K. Suzuki, T. Kiryu, and T. Nakada. Fast and precise independent component analysis for high field fMRI time series tailored using prior information on spatiotemporal structure. *Human Brain Mapping*, 15:54–66, 2001.

[158] M. Svensén, F. Kruggel, and H. Benali. ICA of fMRI group study data. *NeuroImage*, 16:551–563, 2002.

[159] J.F. Swidzinsky and K. Chang. Nonlinear statistical modeling and yield estimation technique for use in monte carlo simulations. *IEEE Transactions on Microwave Theory and Techniques*, 48(12):2316–2324, 2000.

[160] A. Taleb and C. Jutten. On underdetermined source separation. In *Proc. ICASSP 99*, volume 3, pages 1445–1448, 1999.

[161] K.L. Tan, P.H. Liu, D.C. Streit, R. Dia, A.C. Han, A. Freudental, J. Velebir, K. Stolt, J. Lee, M. Biedenbender, R. Lai, H. Wang, and B. Allen. A manufacturable high performance 0.1-μm pseudomorphic AlGaAs/InGaAs HEMT process for W-band MMICs. In *Gallium Arsenide Integrated Circuit (GaAs IC) Symposium*, pages 251–254, 1992.

[162] A.C. Tang, M.T. Sutherland, and C.J. McKinney. Validation of SOBI components from high-density EEG. *Neuroimage*, 25:539–553, 2004.

[163] F. Tecchio, F. Zappasodi, P. Pasqualetti, and P.M. Rossini. Neural connectivity in hand sensorimotor area: an evaluation by evoked fields morphology. *Human Brain Mapping*, 24:99–108, 2005.

[164] B. Thirion and O. Faugeras. Dynamical components analysis of fMRI data through kernel PCA. *NeuroImage*, 20:34–49, 2003.

[165] L. Tong and Y.F. Huang R.W. Liu, V.C. Soon. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38(5):499–509, 1991.

[166] R. Turner, D. Le Bihan, C.T. Moonen, D. Despres, and J. Frank. Echo planar time course MRI of cay brain oxygenation changes. *Magnetic Resonance in Medicine*, 22:159–166, 1991.

[167] R. Vigario and E. Oja. Independence: a new criterion for the analysis of the electromagnetic fields in the global brain? *Neural Networks*, 13:891–907, 2000.

[168] R. Vigario, J. Sarela, K. Jousmaki, M. Hamalainen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47:589–593, 2000.

[169] A. Villinger, J. Planck, C. Hock, L. Schleninkofer, and U. Dirnagl. Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience Letters*, 154:101–104, 1993.

[170] J. Vrba and S.E. Robinson. Signal processing in magnetoencephalography. *Methods*, 25(2):249–271, 2001.

[171] L. Wanhammar. *DPS Integrated Circuits*. Academic Press, 1999.

[172] R.M. Weiskoff, J.L. Boxerman, C.S. Zuo, and B.R. Rosen. Endogenous susceptibility contrast: Principles of relationship between blood oxygenation and MR signal change. In *Functional MRI of the Brain. Berkley: SMRM*, pages 143–151, 1993.

[173] P. Woods, S.R. Cherry, and J.C. Mazziotta. Rapid automated algorithm for aligning and reslicing PET images. *Journal of Computer Assisted Tomography*, 16:620–633, 1992.

[174] S. Xie, Z. He, and Y. Fu. A note on Stone's conjecture of blind signal separation. *Neural Computation*, 17(2):321–330, 2005.

[175] A. Yeredor. Blind separation of gaussian sources via second-order statistics with asymptotically optimal weighting. *IEEE Signal Processing Letters*, 7(7):197–200, 2000.

[176] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with applications in Blind Source Separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, 2002.

[177] E.A. De Yoe, J. Neitz, P.A. Bandettini, and E.C. Wong. Time-course of event related MR signal enhancement in visual and motor cortex. In *Proceedings of the 11th annual meeting of the Society of Magnetic Resonance in Medicine*, page 1824, 1992.

[178] X. Zhao, D. Glahn, L.H. Tan, N. Li, J. Xiong, and J.-H Gao. Comparison of TCA and ICA techniques for fMRI data processing. *Journal of Magnetic Resonance Imaging*, 19:397–402, 2004.

[179] A. Ziehe, M. Kawanabe, S. Harmeling, and K.R. Müller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its applications to Blind Source Separation. *Journal of Machine Learning Research*, 5:801–818, 2004.

[180] A. Ziehe and K. Muller. TDSEP – an efficient algorithm for blind separation using time structure. In *Proceedings of Int. Conf. on Artificial Neural Networs (ICANN'99)*, pages 675–680, 1997.

[181] L. Zukhov, D. Weinstein, and C. Johnson. Independent component analysis for EEG source localization. *IEEE Eng. Med. Biol. Mag.*, 19(3):87–96, 2000.