Analysis of Functional Magnetic Resonance Imaging Time Series by Independent Component Analysis

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat)

von Dipl.-Statistikerin Mandy Sohr

geb. am 20.02.1980 in Lutherstadt Wittenberg

genehmigt durch die Fakultät für Mathematik

der Otto-von-Guericke-Universität Magdeburg

Gutachterin: Frau Prof. Dr. Waltraud Kahle

Gutachter: Herr Priv.Doz. Dr. Siegfried Kropf

eingereicht am: 26.02.2007

Verteidigung am: 22.06.2007

Contents

List of Figures

List of Tables

Index of Symbols

X	random variable
x	realization of random variable
$\mathbf{X} = (X_1, \ldots, X_N)$	vector of random variables
$\mathbf{x} = (x_1, \ldots, x_N)$	vectors of realizations of random variables
X_t	stochastic process $X_t = \{X(t), t \in T \subseteq \mathbf{R}\}$
$(\mathbf{R}, \mathfrak{B}, F)$	probability space
$\mathbf{R}$	sample space, basic set of elements, usually the set of real numbers
$\mathbf{N}$	space of natural numbers
$\mathfrak{B}$	σ-algebra of subsets of $\mathbf{R}$
F	probability measure
$(\mathbf{R}^N, \mathfrak{B}^N, F_i)$	N-dimensional probability space
$\mathbf{R}^N$	N-dimensional sample space, basic set of elements
$\mathfrak{B}^N$	σ-algebra of subsets of $\mathbf{R}^N$
F_i	probability measures, $i = 1, 2$
H_0	null hypothesis
H_1	alternative hypothesis
$F(x) = P(X \leq x)$	cumulative distribution function of a random variable X
$f(x) = F'(x)$	probability density function of a random variable X
$\phi(x)$	probability density function of standard gaussian distributed random variable
$\Phi(x)$	probability distribution function of standard gaussian distributed random variable

Index of Symbols

$E\{X\} = \mu_x$	expectation of a random variable X
$Var(X) = \sigma_x^2$	variance of a random variable X
$\mathbf{C_X} = Cov(\mathbf{X})$	covariance matrix of a random vector $\mathbf{X}$
ρ_{xy}	correlation coefficient between two random variables X and Y
$I(f_1(\mathbf{x}) : f_2(\mathbf{x}))$	information of two density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$
$H(f(x))$	differential entropy of a random variable X
$J(f(x))$	negentropy of a random variable X
$\|a\|$	absolute value of a
$\|\|\mathbf{X}\|\|$	Euclidian length of a vector $\mathbf{X}$
$\det(\mathbf{A})$	determinant of a matrix $\mathbf{A}$
$\mathrm{rg}(\mathbf{A})$	rank of a matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	trace of a matrix $\mathbf{A}$
$\mathbf{A}^T$	transpose of a matrix $\mathbf{A}$
$\mathbf{I}$	Identity matrix
$\mathbf{J}$	Jacobian matrix, matrix of partial derivatives
$K(x)$	Kernel function
$L(x)$	Likelihood function
$\mathbf{X} = [x_{i,t}]_{i=1,...,N,t=1,...,T}$	data matrix of observations
$\mathbf{S} = [s_{j,t}]_{j=1,...,M,t=1,...,T}$	matrix of independent source signals
$\mathbf{A} = [a_{i,j}]_{i=1,...,N,j=1,...,M}$	mixing matrix
$\mathbf{W} = [w_{j,i}]_{j=1,...,M,i=1,...,N}$	unmixing matrix
$\mathbf{Y} = [y_{j,t}]_{j=1,...,M,t=1,...,T}$	estimated source signals
$\mathbf{U} = [u_{i,j}]_{i=1,...,N,j=1,...,M}$	eigenvector matrix of $\mathbf{X}\mathbf{X}^T$
$\mathbf{V} = [v_{t,j}]_{t=1,...,T,j=1,...,M}$	eigenvector matrix of $\mathbf{X}^T\mathbf{X}$
$\mathbf{\Lambda} = [\lambda_{j,j}]_{j=1,...,M}$	diagonal matrix of eigenvalues
N	number of mixtures
M	number of estimates
T	number of time points

Index of Symbols

τ	lag in autocorrelation function
$c(\tau)$	autocovariance function
$\rho(\tau)$	autocorrelation function
ω	frequency
$IF(\omega)$	intensity function of ω
$FT(\omega)$	Fourier transform
κ_s	length of stimulation block (in seconds) ($\kappa_s > 0,\ \kappa_s \in \mathbf{N}$)
κ_r	length of resting block (in seconds) ($\kappa_r > 0,\ \kappa_r \in \mathbf{N}$)
κ_p	number of phases ($\kappa_p > 0,\ \kappa_p \in \mathbf{N}$)
κ_b	number of fMRI images recorded during one block ($\kappa_b > 0,\ \kappa_b \in \mathbf{N}$)
κ_B	total length the experiment (in images) ($\kappa_B > 0,\ \kappa_B \in \mathbf{N}$)
κ_{TR}	length of one image ($\kappa_{TR} > 0,\ \kappa_{TR} \in \mathbf{R}$)
κ_T	total length of the experiment (in seconds) ($\kappa_T > 0,\ \kappa_T \in \mathbf{R}$)
κ_a	signal amplitude ($\kappa_a \in \mathbf{R}$)
κ_m	variation of signal amplitude between stimulation blocks ($\kappa_m \in \mathbf{R}$)
κ_n	variation of signal amplitude within stimulation blocks ($\kappa_n \in \mathbf{R}$)
κ_c	temporal shift between two HRFs ($\kappa_c \in \mathbf{N}$)

Index of Abbreviations

AC	auditory cortex
ACF	autocorrelation function
ANOVA	analysis of variance
BA	Brodmann area
BOLD effect	blood oxygen level dependent effect
BSS	blind source separation
CDF	cumulative distribution function
EEG	electroencephalography
EPI	echo planar imaging
FM	frequency modulation
fMRI	functional magnetic resonance imaging
GLM	general linear model
GUI	graphical user interface
HRF	hemodynamic response function
ICA	independent component anaylsis
iff	if and only if
MEG	magnetoencephalography
PCA	principal component anaylsis
PDF	probability density function
RMSE	root mean squared error
ROI	region of interest
sICA	spatial ICA
SVD	singular value decomposition
tICA	temporal ICA
VOI	volume of interest

Abstract

Functional magnetic resonance imaging (fMRI) gained a lot of interest in medical and human research in the last years. FMRI is a noninvasive method used to study human brain functions by localizing activated brain areas. There are a lot of interesting questions in neurobiology, one of these is the processing of learning-related processes in the human brain and how these processes can be described and analyzed. To analyze fMRI data different hypothesis-based and data-based methods can be used. The Independent Component Analysis (ICA) is an information-theoretic statistical and computational technique used to identify hidden factors of observed multivariate data. The mathematical background of ICA is investigated in this thesis and relations to other methods like Principal Component Analysis (PCA) are elaborated. The advantages of ICA in comparison to classical methods for analyzing fMRI data under the aspect of learning-related processes are investigated in real fMRI data as well as in simulations studies. Thereby dynamic changes in the fMRI time series are systematically analyzed and described.

Zusammenfassung

Die funktionelle Kernspintomographie (fMRT) hat in den letzten Jahren sehr an Bedeutung in der Medizin wie auch in der Forschung gewonnen. Mit dieser Methode können unter anderem Hirnaktivitäten des menschlichen Gehirns untersucht und lokalisiert werden. In der Neurobiologie ergeben sich hieraus viele interessante Fragestellungen z. B. wie im Gehirn bestimmte Lern- und Gedächtnisprozesse verarbeitet werden. Für die Analyse von fMRT-Daten können verschiedene hypothesenbasierende und hypothesengenerierende Verfahren angewendet werden. Die Independent Component Analysis (ICA) ist ein informationstheoretisches statistisches Verfahren, um zugrunde liegende Faktoren in beobachteten multivariaten Daten zu identifizieren. Der mathematische Aspekt der ICA wird in dieser Arbeit beleuchtet sowie Zusammenhänge zu anderen Verfahren wie der Principal Component Analysis (PCA) hergestellt. Die Vorteile der ICA gegenüber klassischen Analyseverfahren von fMRT-Daten unter dem Aspekt von Lernprozessen werden an realen Datensätzen sowie in Simulationsstudien erarbeitet. Hierbei werden systematisch die dynamischen Veränderungen in den fMRT-Zeitreihen untersucht und analysiert.

1 Introduction

Independent component analysis (ICA) is a statistical and computational technique for revealing hidden factors that are sets of random variables, measurements, or signals. ICA defines a generative model for the observed multivariate data, which is given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of unknown latent variables, whereas the mixing system is also unknown. The latent variables are assumed to be mutually independent, and they are called the independent components of the observed data [Comon, 1994, Hyvärinen et al., 2001*b*].

In general, the method of ICA can be applied to blind source separation (BSS) problems, where the measurements are given as a set of parallel signals. Examples for BSS problems can be found everywhere, for instance mixtures of simultaneous speech signals that have been picked up by several microphones, brain waves recorded by multiple sensors, interfering radio signals arriving at a mobile phone, or parallel time series obtained from some industrial process. These examples are taken from Samarov and Tsybakov, 2004. The concept of ICA was described in different publications like Jutten and Hérault, 1991, Comon, 1994, or Hyvärinen et al., 2001*b*.

ICA can be applied to a variety of problems like finding hidden factors in financial time series or reducing noise in natural images. ICA was also already successfully applied to neurobiological studies to obtain sources of neuronal activation. Thereby this method was applied to electroencephalography (EEG) data and magnetoencephalography (MEG) data [Makeig et al., 1997, Vigario et al., 2000], furthermore, to data obtained from functional magnetic resonance imaging (fMRI) studies [McKeown et al., 1998*b*]. FMRI is a method for noninvasively studying human brain functions by localizing activated brain regions as a consequence of responding to a stimuli. The signal, an fMRI time series, is recorded for anatomical coordinates in the brain which are called voxels. The use of ICA for fMRI time series was motivated for several reasons. A mathematical reason is that the time series obtained by fMRI studies are supposed to be linear mixtures of realizations of different stochastic processes and the voxel values are considered as random variables. The

stochastic processes might be the neuronal responses to presented stimuli, processes related to heart beat or breathing of the subjects, motion artifacts, and noise caused by the tomograph. Since the exact temporal behavior of such signals is not always predictable, the ICA, a method without any hypothesis about the expected time courses, is used to extract these different signals [McKeown et al., 1998*b*]. This method was already used to reduce head motion-induced variations from the fMRI signal [Liao et al., 2006] or to reduce the noise of neuronal fMRI responses [Thomas et al., 2002]. Moreover, some studies demonstrated the use of ICA for clinical fMRI processing by comparing ICA to conventional hypothesis driven analysis [Quigley et al., 2002].

The work was cooperated with the Leibniz-Institute for Neurobiology in Magdeburg, a center for learning and memory research. The special-laboratory for noninvasive brain imaging of this institute investigates the role and functional organization of the auditory cortex (AC) in humans. Their aim is to study the AC by examining the processing of fundamental stimulus properties in combination with specific tasks with neuroimaging methods like EEG, MEG, and fMRI. The special-lab is particularly interested in the processing of cognitive tasks and working memory (WM) processes in the human AC. The basic concept of WM refers to "a brain system that provides temporary storage and manipulation of the information necessary for cognitive tasks" [Baddeley, 1992]. This means that some information must be shortly maintained, recalled, and compared with test items to previously instructed rules. In the special-lab some fMRI studies involving WM tasks were performed as described in Brechmann et al., 2007. In this study frequency modulated (FM) tones were presented as stimuli in a sequence. For each stimulus tone, the subjects had to decide by key pressing whether it matched a tone two back in the tone sequence according to FM direction and frequency. Interesting findings of these studies were that the activation of the subjects strongly depends on the task and the activation is success-dependent as discussed in Gaschler-Markefski et al., 2003 and Sohr et al., 2003. This was shown be correlating the amount of fMRI activation to the task performance of the subjects. Thereby, we interestingly found a positive correlation between amount of activated voxels and task performance of the subjects proposing that the positive direction of the correlation of activity in left AC with the performance of subjects is a correlate of memory maintenance.

This was in contrast to previous found correlations in the special-laboratory, namely negative correlations. For instance in a directional categorization of similar FM stimuli as used in the present study a negative correlation with performance was found on the right side [Brechmann and Scheich, 2005]. This negative correlation was interpreted as a sign of variable proficiency in solving this classification task, namely that a high performance

in processing depends on a restriction to specialized neurons that clearly distinguish the direction of FM at the expense of less specific neurons which may be involved in an initial, less proficient stage of the experiment.

These findings are particularly relevant because correlations of performance with BOLD activation are considered signs of specific involvement of brain structures in a task [Ohl and Scheich, 2005]. Moreover, positive or negative correlations of BOLD activity are found in different tasks which may allow different interpretations of the underlying process [Jonides, 2004].

It was previously shown in a study by Barch et al., 1997, that the fMRI activation is success-dependent, and the dependence of activation and the task was already shown by Braver et al., 1997, to name just a few publications. But it should be pointed out that these studies investigated only the parietal and the frontal cortex in humans. An involvement of the AC in WM studies was so far not demonstrated by fMRI studies. This lead us to a further motivation for utilizing ICA for fMRI time series, namely, the fMRI activation is success-dependent and may therefore differ from subject to subject. Therefore, the time series of fMRI activation themselves, and not the amount of activation, should be investigated and analyzed in detail. Based on the previous studies of the special-lab it is aimed to describe dynamic, spatio-temporal changes in neuronal response of subjects in fMRI studies involving cognitive learning processes in repeated sessions. ICA should help to detect possibly dynamic changes in fMRI times series which cannot be detected with classical methods for analyzing fMRI time series like general linear models (GLM).

In the literature there are some publications discussing the shape of the neuronal response of time series, see for instance Seifritz et al., 2003 or Harms and Melcher, 2003. They described the waveshape of the signals to be composed of transient and sustained activations. But how the fMRI signal would behave under the aspect of learning related processes of the subjects was so far not investigated.

Aim of this work is to describe the mathematical coherence to the information-theoretic method of ICA. Thereby the relation of ICA and principal component analysis (PCA) for nongaussian and gaussian random variables, respectively, is pointed out. Furthermore, information-theoretic relations to statistics are elaborated. The ICA is used to detect learning related dynamic changes in fMRI time series. On the one hand this is supported by simulation studies, where different models for learning related changes are introduced. On the other hand the detection of dynamic changes in the fMRI signal is verified by an fMRI

study investigating an auditory WM study with repeated sessions of the subjects.

The content of the thesis is organized as follows:

Chapter 2 introduces fundamental terms and statistical definitions. Probability spaces and random variables for the univariate and multivariate case and stochastic processes are defined. Statistical independence is described and the relation between independent and uncorrelated random variables is demonstrated. Moreover, measures of nongaussianity and independence are introduced, namely kurtosis as well as information-theoretic functions like information, differential entropy, and negentropy. These measures are used in ICA algorithms.

Chapter 3 describes the principles of MRI and fMRI. FMRI is a method for noninvasively measuring signal changes in the brain that are due to changes of neural activity. This chapter attends to fMRI time series which are considered as realizations of stochastic processes. Hypothesis-based and data-based methods for analyzing fMRI data are summarized.

Chapter 4 describes classical methods for analyzing fMRI data in more detail. These methods include general linear models (GLM) and principal component analysis (PCA). Since the observed fMRI data are time series, some methods for analyzing time series are summarized in this chapter as well.

Chapter 5 introduces the method of ICA. This chapter addresses the definition of ICA, its identifications and restrictions as well as ambiguities of ICA estimates are given. The relationship between PCA and ICA is pointed out. Classical ICA algorithms from the literature like the Jutten-Hérault algorithm, maximum likelihood estimations, information maximization algorithms, FastICA algorithm, decorrelation approaches, and nonparametric ICA algorithm are described. In a following section, performance testing methods of the estimates of ICA algorithms are pointed out. Finally, a literature overview depicts the application of ICA to fMRI data at the end of that chapter.

Chapter 6 describes performing and results of simulation studies. The aim of simulation studies is to test whether source signals can be estimated from linearly mixed signals using different ICA algorithms from Chapter 5. The signal of the hypothetical neuronal response is modelled and since it is assumed that the task performance of subjects is reflected in the neuronal response, this signal is varied in different parameters like the signal amplitude, an amplitude increase or decrease within or between stimulation blocks. Additionally, further signals that might contribute to an fMRI measurement are modelled and mixed linearly

with each other. Different ICA algorithm are used to perform a decomposition of the mixed signals into the independent source signals. The performance of the estimation is validated by error indices to draw conclusions about the performance of the estimated independent components for different parameters. Among all the tested ICA algorithms, the FastICA algorithm outperformed the other algorithms. This algorithm showed good estimates even if dynamic changes within the time series are modelled.

Chapter 7 describes an auditory fMRI study with repeated sessions investigating a WM task. The data is analyzed with ICA. Furthermore, temporal changes in the signals within or between repeated sessions are investigated. These results of ICA are compared with classical correlation analysis results. The results revealed that almost the same areas are involved for all subjects but showing different dynamic time courses. We also showed that a hypothesis-based method like GLM is not always the best approach to investigate task-related activations. In this special case the ICA results outperformed the results of GLM analysis.

Appendices A.1 - A.3 compile relevant properties and characteristics of the information-theoretic functions of Section 2.3. Additionally, Appendix A.4 describes the approximation of information-theoretic functions through higher-order cumulates since the information-theoretic functions are more theoretical functions than practically used functions.

2 Fundamentals

In this chapter fundamental terms and definitions are described. First, probability spaces and random variables for the univariate and multivariate case are defined and stochastic processes are introduced. Second, the term statistical independence is defined and the relation between independent and uncorrelated random variables is demonstrated. In one subsection measures of independence and nongaussianity are introduced, namely information-theoretic functions like information, differential entropy, and negentropy, and kurtosis as measure of nongaussianity. Additionally, the information and differential entropy for some distributions like the uniform distribution as well as the univariate and multivariate gaussian distribution are given in examples.

2.1 Probability Spaces, Random Variables, and Stochastic Processes

The following notation is used: X is used for a univariate random variable and x is used for a specific value of X. Just as well, $\mathbf{X}$ is used for a random vector $\mathbf{X} = (X_1, \ldots, X_N)$ and $\mathbf{x} = (x_1, \ldots, x_N)$ is used for a realization of $\mathbf{X}$.

Definition 2.1 *Considering the Euclidian space* $\mathbf{R}$*, a* σ*-algebra* $\mathfrak{B}$ *of Borelian subsets of* $\mathbf{R}$ *and a probability measure* P*, this triple* $(\mathbf{R}, \mathfrak{B}, P)$ *is called probability space.*

$\mathbf{R}^N$ is the sample space and in case of realizations it describes the real-valued space of variables.

Definition 2.2 *Let* X *be a univariate random variable then* $F(x) = P(X \leq x)$ *is the cumulative distribution function. If* $F(x)$ *is absolutely continuous, the random variable* X *is said to be continuous. If* $F(x)$ *is differentiable and* $f(x) = F'(x)$ *is the derivative of the distribution function, then* $f(x)$ *is the probability density function for* X *(*$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$*).*

Definition 2.3 *Let us consider a set of random variables* X_t*. The variables depend on a time parameter* $t \in T \subseteq \mathbf{R}$*. We call* $X_t = \{X(t), t \in T \subseteq \mathbf{R}\}$ *a stochastic process* [Fisz,

1980]. *For a specific t, $X(t)$ is a random variable with distribution*

$$F(x,t) = P(X(t) \leq x). \tag{2.1}$$

According to Papoulis, *1991, the function $F(x,t)$ will be called the first-order distribution of the process X_t. Its derivative with respect to x:*

$$f(x,t) = \frac{\partial F(x,t)}{\partial x}, \tag{2.2}$$

is the first-order density of X_t.

Definition 2.4 *Let $\mathbf{X}$ be a multivariate random vector $(X_1, \ldots, X_N)$ with distribution function $F(\mathbf{x}) = P(X_1 \leq x_1, \ldots, X_N \leq x_N)$. The multivariate random vector $\mathbf{X}$ is called continuous if a real-valued function $f(\mathbf{x}) = f(x_1, \ldots, x_N)$ exists and let $f(\mathbf{x}) = F'(\mathbf{x})$ be the derivative of the distribution function, that*

$$F(\mathbf{x}) = \int_{\mathbf{R}^N} f(x_1, \ldots, x_N)\, \mathrm{d}x_1 \ \ldots\ \mathrm{d}x_N. \tag{2.3}$$

The function $f(\mathbf{x}) = f(x_1, \ldots, x_N)$ is called the multivariate density function of $(X_1, \ldots, X_N)$ with the following properties

$$f(\mathbf{x}) \geq 0 \ \forall\ x_1, \ldots, x_N \in \mathbf{R}^N \tag{2.4}$$

and

$$\int_{\mathbf{R}^N} f(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1. \tag{2.5}$$

For references of multivariate distribution and density functions see Hartung and Elpelt, 1995. Let us now consider the multivariate vector $\mathbf{X} = (X_1, \ldots, X_N)$ and we further define two complementary hypothesis H_1 and H_2.

Theorem 2.1 *If H_i, $i = 1, 2$, are two complementary hypotheses that $\mathbf{X}$ is from the statistical population with the probability measure F_i, and $\mathbf{x} \in \mathbf{R}^N$ is an observation, then it follows from Bayes' theorem* [Kullback, 1959] *that*

$$P(H_i|\mathbf{x}) \ = \ \frac{P(H_i) f_i(\mathbf{x})}{P(H_1) f_1(\mathbf{x}) + P(H_2) f_2(\mathbf{x})}, \quad i = 1, 2. \tag{2.6}$$

where $f_i(\mathbf{x}) \neq 0, \ \ i = 1, 2$, are the probability densities and $P(H_i), \ \ i = 1, 2$, are the prior probabilities of H_1 and H_2 respectively. From this it follows that

$$\frac{P(H_1) f_1(\mathbf{x})}{P(H_1|\mathbf{x})} = \frac{P(H_2) f_2(\mathbf{x})}{P(H_2|\mathbf{x})}, \tag{2.7}$$

and hence

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{P(H_2)/P(H_1)}{P(H_2|\mathbf{x})/P(H_1|\mathbf{x})}, \tag{2.8}$$

where $P(H_i|\mathbf{x})$ is the posterior probability of H_i, or the conditional probability of H_i given $\mathbf{X} = \mathbf{x}$. From that equation it is obtained

$$\log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \log \frac{P(H_1|\mathbf{x})}{P(H_2|\mathbf{x})} - \log \frac{P(H_1)}{P(H_2)} \tag{2.9}$$

This equation is needed later for the definition of information I (see Equation 2.22).

2.2 Independence and Correlation

In the following the independence of random variables should be considered. Consider two random variables X_1 and X_2. The joint probability density of X_1 and X_2 is denoted by $f(x_1, x_2)$. Basically, if the variables are independent then the information of the value of X_1 does not give any information on the value of X_2, and vice versa.

Let us further denote by $f^{(1)}(x_1)$ the marginal probability density function of X_1, i.e. the probability density function of X_1 when it is considered alone:

$$f^{(1)}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2 \tag{2.10}$$

and analogously for x_2, with

$$f^{(2)}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_1 \tag{2.11}$$

the marginal probability density function of X_2.

Definition 2.5 *The random variables X_1 and X_2 are independent if and only if (iff) the joint density function is the product of the marginal distributions*

$$f(x_1, x_2) = f^{(1)}(x_1) f^{(2)}(x_2). \tag{2.12}$$

The Definition 2.5 can be extended to N multivariate random variables $\mathbf{X} = (X_1, \ldots, X_N)$ with multivariate density function $f(\mathbf{x})$ and marginal densities $f^{(i)}(x_i),\ i = 1, \ldots, N$:

Definition 2.6 *The random variables $X_1, \ldots, X_N$ of a random vector $\mathbf{X} \in \mathbf{R}^N$ are independent iff the joint density function is the product of the marginal independent distributions*

$$f(\mathbf{x}) = \prod_{i=1}^{N} f^{(i)}(x_i). \tag{2.13}$$

A weaker form of independence is uncorrelatedness. Independent variables are different from uncorrelated variables because uncorrelated variables are only partly independent.

Definition 2.7 *Suppose we have a two dimensional random vector ($N = 2$) with elements X_1 and X_2, then the random variables X_1 and X_2 are said to be uncorrelated, if their covariance c_{x_1,x_2} is zero:*

$$c_{x_1,x_2} = E\{X_1 X_2\} - E\{X_1\}E\{X_2\} = 0, \tag{2.14}$$

where $E\{X\}$ is the expectation of a random variable X, calculated for continuous random variables as

$$E\{X\} = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x. \tag{2.15}$$

In the more general setting with arbitrary N we have a similar construct.

Definition 2.8 *The covariance matrix $\mathbf{C_X}$ with elements c_{ij}, $i, j = 1, \ldots, N$ of an N-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_N)$ is defined by*

$$c_{ij} = Cov(X_i, X_j) = E\{(X_i - \mu_i)(X_j - \mu_j)^T\}, \tag{2.16}$$

where μ_i, $i = 1, \ldots, N$ is the expectation of the random variable X_i. The elements $c_{ii} = \sigma_i^2$, $i = 1, \ldots, N$ are the variances of the random variables X_i. $X_1, \ldots, X_N$ are called uncorrelated if all c_{ij} are zero, for reference see Hartung *and* Elpelt*, 1995.*

If the variables are independent, they are uncorrelated, but uncorrelatedness does not imply independence, see Hyvärinen et al., 2001*b*. In this context the correlation coefficient ρ will be defined.

Definition 2.9 *Given N random variables X_i, $i = 1, \ldots, N$, the correlation coefficient ρ_{ij}, $i, j = 1, \ldots, N$ is defined by*

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} = \frac{c_{ij}}{\sigma_i \sigma_j}, \tag{2.17}$$

where $Var(X_i)$ and $Var(X_j)$ are the variances of X_i and X_j. The variance $Var(X)$ of a continuous random variable X is defined as

$$Var(X) = \int_{-\infty}^{\infty} (x - E\{X\})^2 f(x)\, \mathrm{d}x = E\{(X - E\{X\})^2\}. \tag{2.18}$$

To show the relation of independent and uncorrelated random variables, consider the following 2-dimensional example. Let S_1 and S_2 be two random variables. S_1 and S_2 are independent variables with a uniform distribution, i.e. knowing the value of S_1 gives no information about what the corresponding value of S_2 might be. The sample distribution of

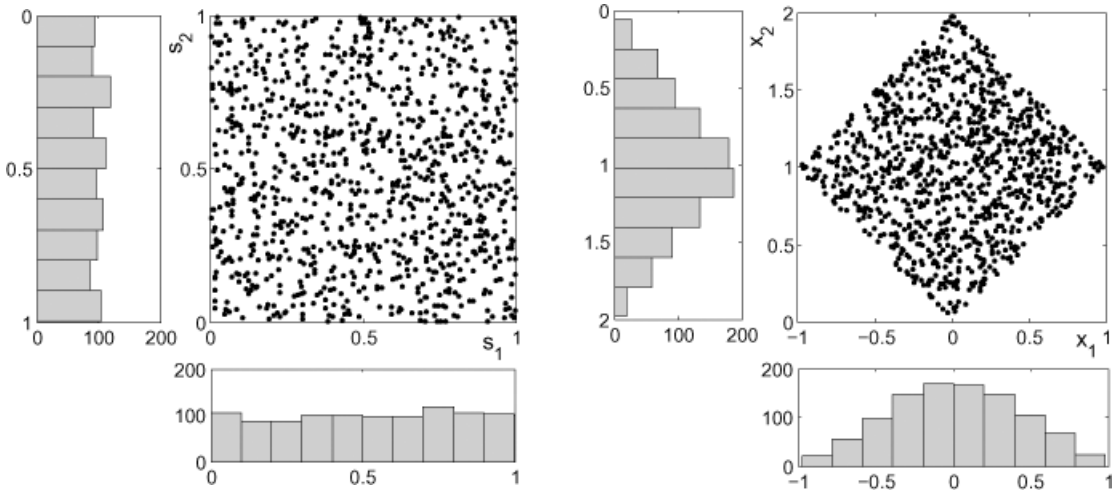

Figure 2.1: Independent versus uncorrelated random variables

the two variables is shown in the left plot of Figure 2.1. Suppose then two random variables X_1 and X_2 which are linear mixtures of the variables S_1 and S_2. In this case $X_1 = S_1 - S_2$ and $X_2 = S_1 + S_2$. The distribution of X_1 and X_2 is shown in the right plot of Figure 2.1. These mixtures X_1 and X_2 are uncorrelated, since $Cov(X_1, X_2) = Cov(S_1 - S_2, S_1 + S_2)$ $= E\{S_1^2 + S_1S_2 - S_2S_1 - S_2^2\} - E\{S_1 - S_2\}E\{S_1 + S_2\} = E\{S_1^2\} - E\{S_2^2\} = 0$. However, they are not independent. Consider therefore a point at $X_1 = 0.8$. Knowing the location $X_1 = 0.8$ gives information about X_2 since it is constrained to be within a limited range. Thus X_1 and X_2 are uncorrelated but not independent.

Estimates of the marginal densities are shown as histograms. The histograms of X_1 and X_2 go closer to a gaussian distribution than the histograms of S_1 and S_2 which are uniformly

distributed. This property comes from the central limit theorem, which states that any linear mixture of two independent random variables is more approximately gaussian than the original variables.

Theorem 2.2 (Fisz, 1980, (Lindeberg-Feller)) *Given a sequence* $\{X_i\}(i = 1, \ldots, N)$ *of independent random variables with* $F_i(x)$, μ_i, *and* $\sigma_i \neq 0$ *the distribution function, mean and standard deviation of the random variable* X_i, *further we have* $\sigma_N = \sqrt{\sum_{i=1}^{N} \sigma_i^2}$. *Under this condition it holds that* $\lim_{N\to\infty} \max_{1\leq i\leq N} \frac{\sigma_i}{\sigma_N} = 0$, *and the sequence of distribution functions* $F_N(z)$ *of the standardized random variables* Z_N *with* $Z_N = \frac{\sum_{i=1}^{N}(X_i-\mu_i)}{\sigma_N}$ *fulfills the condition*

$$\lim_{N\to\infty} F_N(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} \mathrm{d}z \tag{2.19}$$

if for each $\varepsilon > 0$

$$\lim_{N\to\infty} \frac{1}{\sigma_N^2} \sum_{i=1}^{N} \int_{|x-\mu_i|>\varepsilon\sigma_N} (x-\mu_i)^2 dF_i(x) = 0. \tag{2.20}$$

2.3 Measures of Independence and Nongaussianity

ICA uses statistical independence of the source signals as criterion for the estimation of independent components (see Chapter 5). Therefore, general quantitative measures of statistical independence of random variables X_i are needed. Furthermore, it is assumed that the source signals should be nongaussian distributed, therefore, the kurtosis as measure of nongaussianity will be introduced as well.

2.3.1 Measuring Independence by Information-Theoretic Functions

In this section information-theoretic measures of independence as information, differential entropy and negentropy will be described [Cover and Thomas, 1991]. The information theory is a part of the mathematics describing random variables by exceeding the classical measures of expectation and variance.

Definition 2.10 *The expected logarithm under the hypothesis* H_1 *of the likelihood ratio,* $E\{\log[f_1(\mathbf{x})/f_2(\mathbf{x})]\}$, *is defined as the information* $I(1:2)$ *for discrimination in favor of*

hypothesis H_1 stating that $f_1(\mathbf{x})$ is the joint density function ($f(\mathbf{x})$) against hypothesis H_2 that $f_2(\mathbf{x})$ is the product of marginal densities ($\prod_{i=1}^{N} f_i(x_i)$). This means that, under H_1 the variables in $\mathbf{X} \in \mathbf{R}^N$ are dependent but under H_2 the variables are independent. Deriving from Equation (2.9)*,*

$$I(1:2) = I(f_1(\mathbf{x}) : f_2(\mathbf{x})) = \int_{\mathbf{R}^2} f_1(\mathbf{x}) \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \, \mathrm{d}\mathbf{x}. \tag{2.21}$$

$$= \int_{\mathbf{R}^2} f_1(\mathbf{x}) \log \frac{P(H_1|\mathbf{x})}{P(H_2|\mathbf{x})} \, \mathrm{d}\mathbf{x} - \log \frac{P(H_1)}{P(H_2)}. \tag{2.22}$$

This definition of information is according to the KULLBACK-LEIBLER-divergence [KULLBACK, 1959]. In the literature, there occur different terms for information like mutual information of mean information, as well. Moreover, there are different notations, where $I(1:2) = I(f_1(\mathbf{x}) : f_2(\mathbf{x}))$.

The information $I(1:2)$ is a natural measure of dependence between random variables. The information will always be positive and will equal zero only when the components are independent (see Appendix A.1). Using the information as measure, it takes into account the whole dependence structure of the variables. In ICA it is aimed to recover source signals that are as independent of each other as possible, not just uncorrelated as used in principal component analysis (PCA) (see Chapter 5.2).

Example 2.1 *Information of a two-dimensional random variable* [KULLBACK, 1959, p. 8]: Suppose that the sample space $\mathbf{R}^N$ is the Euclidean space $\mathbf{R}^2$ with elements X_1 and X_2. Under H_1 the variables X_1 and X_2 are dependent with probability density $f(x_1, x_2)$, but under H_2, X_1 and X_2 are independent, with respective probability densities $f^{(1)}(x_1)$ and $f^{(2)}(x_2)$. Regarding the hypotheses H_1 and H_2, the information $I(1:2)$ can be written as:

$$I(1:2) = \int_{\mathbf{R}^2} f(x_1, x_2) \log \frac{f(x_1, x_2)}{f^{(1)}(x_1) f^{(2)}(x_2)} \, \mathrm{d}\mathbf{x}. \tag{2.23}$$

□

Example 2.2 *Information of the central bivariate gaussian density* [KULLBACK, 1959, p. 8]: Assume a central bivariate gaussian density, i.e. with means $\mu_{X_1} = \mu_{X_2} = 0$, furthermore with variances $\sigma^2_{X_1}$ and $\sigma^2_{X_2}$ and a correlation coefficient ρ ($|\rho| \leq 1$). The hypothesis H_1 implies the bivariate gaussian density

$$\phi(x_1, x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{x_1^2}{\sigma^2_{X_1}} - 2\rho \frac{x_1 x_2}{\sigma_{X_1}\sigma_{X_2}} + \frac{x_2^2}{\sigma^2_{X_2}} \right) \right\}, \tag{2.24}$$

and the hypothesis H_2 implies that the joint density is the product of the marginal gaussian densities $\phi^{(1)}(x_1)$ and $\phi^{(2)}(x_2)$ given by

$$\phi^{(i)}(x_i) = \frac{1}{\sigma_{X_i}\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2\sigma_{X_i}^2}\right) \quad i = 1, 2. \tag{2.25}$$

The information of the bivariate gaussian density is given by

$$\begin{aligned} I(1:2) &= \int_{\mathbf{R}^2} \phi(x_1, x_2) \log\left(\frac{1}{\sqrt{1-\rho^2}} \cdot \right. \\ &\quad \left. \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_{X_1}^2} - 2\rho\frac{x_1 x_2}{\sigma_{x_1}\sigma_{X_2}} + \frac{x_2^2}{\sigma_{X_2}^2}\right) + \frac{x_1^2}{2\sigma_{X_1}^2} + \frac{x_2^2}{2\sigma_{X_2}^2}\right\}\right) \mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{R}^2} \phi(x_1, x_2) \log\left(\frac{1}{\sqrt{1-\rho^2}}\right) \mathrm{d}\mathbf{x} + \int_{\mathbf{R}^2} \phi(x_1, x_2) \left\{-\frac{1}{2(1-\rho^2)}\right. \\ &\quad \left. \cdot \left(\frac{x_1^2}{\sigma_{X_1}^2} - 2\rho\frac{x_1 x_2}{\sigma_{X_1}\sigma_{X_2}} + \frac{x_2^2}{\sigma_{X_2}^2} - (1-\rho^2)\frac{x_1^2}{\sigma_{X_1}^2} - (1-\rho^2)\frac{x_2^2}{\sigma_{X_2}^2}\right)\right\} \mathrm{d}\mathbf{x} \\ &= -\frac{1}{2}\log(1-\rho^2) + \int_{\mathbf{R}^2} \phi(x_1, x_2) \left\{-\frac{1}{2(1-\rho^2)}\left(\rho^2\frac{x_1^2}{\sigma_{X_1}^2} - 2\rho\frac{x_1 x_2}{\sigma_{X_1}\sigma_{X_2}} + \rho^2\frac{x_2^2}{\sigma_{X_2}^2}\right)\right\} \mathrm{d}\mathbf{x}, \end{aligned}$$

where $\int_{\mathbf{R}^2} \phi(x_1, x_2)\frac{x_1^2}{\sigma_{X_1}^2}\mathrm{d}\mathbf{x} = 1$, as well as $\int_{\mathbf{R}^2} \phi(x_1, x_2)\frac{x_2^2}{\sigma_{X_2}^2}\mathrm{d}\mathbf{x} = 1$,and $\int_{\mathbf{R}^2} \phi(x_1, x_2)\frac{x_1 x_2}{\sigma_{X_1}\sigma_{X_2}}\mathrm{d}\mathbf{x} = \rho$.

Therefore,

$$\begin{aligned} I(1:2) &= -\frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left(\rho^2 - 2\rho^2 + \rho^2\right) \\ &= -\frac{1}{2}\log(1-\rho^2). \end{aligned} \tag{2.26}$$

In this case, the information is a function of the correlation coefficient ρ only, and the information ranges from 0 to ∞ as $|\rho|$ ranges from 0 to 1. □

Further properties of information $I(1:2)$ like additivity can be found in Appendix A.1. The use of information can also be motivated using the concept of entropies. The entropy of a random variable is a measure of the average uncertainty in the random variable [Cover and Thomas, 1991]. In the context of information coding it describes roughly the minimum necessary code length to transmit a large number of observations of a random variable most efficiently [Cover and Thomas, 1991]. In general, for continuous random variables the entropy is called differential entropy, see again Cover and Thomas, 1991.

Definition 2.11 *The differential entropy $H(f(x))$ of a continuous random variable X with a density function $f(x)$ is defined as*

$$H(f(x)) = -\int_{-\infty}^{\infty} f(x) \log f(x) \, \mathrm{d}x. \tag{2.27}$$

Cover and Thomas, 1991, write that the entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable. Therefore, they showed that the entropy is related to the expectation value of the random variable. Based on Definition 2.11 it can be seen that the differential entropy is the expectation of $-\log f(x)$

$$H(f(x)) = E\{-\log f(x)\}. \tag{2.28}$$

Cover and Thomas, 1991, further showed that a low differential entropy implies that the random variable has a small variance, and a high differential entropy indicates that the random variable is widely dispersed. The differential entropy of a random variable is related to the information that the observation of the variable gives. The more 'random', i.e. unpredictable and unstructured the variable is, the larger its differential entropy [Hyvärinen et al., 2001*b*, p. 182].

Example 2.3 *Differential entropy of a uniform distributed random variable* [Cover and Thomas, 1991, p. 225]: Consider a random variable X uniformly distributed in $[0, a]$ with $0 < a < \infty$, that is, its density is given by

$$f(x) = \begin{cases} \frac{1}{a} & 0 \leq x \leq a \\ 0 & \text{elsewhere} \end{cases}. \tag{2.29}$$

Then its differential entropy is

$$H(f(x)) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \mathrm{d}x = \log a \quad \text{for } 0 \leq x \leq a. \tag{2.30}$$

Note that for $a < 1$, $\log a < 0$ and the differential entropy is negative. □

Example 2.4 *Differential entropy of a gaussian distributed random variable* [Cover and Thomas, 1991, p. 225]: Let the random variable X be gaussian distributed with gaussian density function $\phi(x) = (1/\sqrt{2\pi\sigma^2}) \cdot e^{-x^2/2\sigma^2}$, the entropy is given by

$$H(\phi(x)) \quad = \quad -\int_{-\infty}^{\infty} \phi(x) \log \phi(x) \mathrm{d}x$$

$$
\begin{aligned}
&= -\int_{-\infty}^{\infty} \phi(x) \left[-\frac{x^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2} \right] \mathrm{d}x \\
&= \frac{E(X^2)}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \\
&= \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2). \qquad (2.31)
\end{aligned}
$$

□

As it can be seen from these two examples: The larger the variance of the random variable is, indicated by a large a in case of uniform distributed random variables and by a large σ in case of gaussian distributed random variables, the larger the differential entropy of the random variable.

The definition of differential entropy of a single random variable X can be extended to several random variables $X_1, \ldots, X_N$.

Definition 2.12 *The entropy of a set $X_1, \ldots, X_N$ of random variables with joint density $f(x_1, \ldots, x_N)$ is defined as*

$$
H(f(x_1, \ldots, x_N)) = -\int_{\mathbf{R}^N} f(x_1, \ldots, x_N) \log f(x_1, \ldots, x_N) \, \mathrm{d}x_1 \ldots \mathrm{d}x_N. \qquad (2.32)
$$

Example 2.5 *Differential entropy of a multivariate gaussian distribution* [Cover and Thomas, 1991, p. 230]: Let $X_1, \ldots, X_N$ have a multivariate gaussian distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and covariance matrix $\mathbf{C}$ and a joint density function $\phi(\mathbf{x})$ with $\mathbf{x} = (x_1, \ldots, x_N)$,

$$
\phi(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^N |\det \mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \qquad (2.33)
$$

where $|\det \mathbf{C}|$ denotes the determinant of the covariance matrix $\mathbf{C}$. The elements of $\mathbf{C}$ are given by $c_{ij}, i, j = 1, \ldots, N$ and $c^{ij}, i, j = 1, \ldots, N$ are the elements of $\mathbf{C}^{-1}$. Then the differential entropy is given by

$$
\begin{aligned}
H(\phi(\mathbf{x})) &= -\int_{\mathbf{R}^N} \phi(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \log\left((\sqrt{2\pi})^N |\det \mathbf{C}|^{1/2} \right) \right] \mathrm{d}\mathbf{x} \\
&= \frac{1}{2} E\left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - \mu_i) c^{ij} (x_j - \mu_j) \right\} + \frac{1}{2}\log\left((2\pi)^N |\det \mathbf{C}| \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} E\left[(x_j-\mu_j)(x_i-\mu_i)c^{ij}\right] + \frac{1}{2}\log\left((2\pi)^N|\det\mathbf{C}|\right) \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} c_{ji}c^{ij} + \frac{1}{2}\log(2\pi)^N|\det\mathbf{C}|,
\end{aligned}
$$

since it holds that $\sum_{i=1}^{N} c_{ji}c^{ij} = \begin{cases} 1 & \text{for} \quad i=j \\ 0 & \text{for} \quad i\neq j \end{cases}$, $\mathbf{C}\mathbf{C}^{-1}=\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, then the differential entropy of a multivariate gaussian distribution is given by

$$
H(\phi(\mathbf{x})) = \frac{N}{2} + \frac{1}{2}(\log(2\pi)^N|\det\mathbf{C}|). \tag{2.34}
$$

□

Theorem 2.3 *Besides defining information using* Kullback-Leibler*-divergence* (see Equation 2.21)*, the information can also be defined using differential entropies:*

$$
I(1:2) = \sum_{i=1}^{N} H(f(x_i)) - H(f(\mathbf{x})), \tag{2.35}
$$

Proof: Using Equations (2.21) and (2.27) one can write

$$
\begin{aligned}
I(1:2) &= \int_{\mathbf{R}^N} f(\mathbf{x})\log\frac{f(\mathbf{x})}{\prod_{i=1}^{N} f_i(x_i)}\mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{R}^N} f(\mathbf{x})\left(\log f(\mathbf{x}) - \sum_{i=1}^{N}\log f_i(x_i)\right)\mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{R}^N} f(\mathbf{x})\log f(\mathbf{x})\mathrm{d}\mathbf{x} - \sum_{i=1}^{N}\int\ldots\int \log f_i(x_i) f(x_1,\ldots,x_N)\mathrm{d}x_1\ldots\mathrm{d}x_N \\
&= -H(f(\mathbf{x})) + \sum_{i=1}^{N} H(f(x_i))
\end{aligned}
$$

■

Cover and Thomas, 1991 interpreted the information by using the interpretation of differential entropy as code length. The terms $H(f(x_i))$ give the lengths of codes for the x_i when these are coded separately, and $H(f(\mathbf{x}))$ gives the code length when $\mathbf{x}$ is coded as a random vector, i.e. all the components are coded in the same code. Information thus shows what code length reduction is obtained by coding the whole vector instead of the separate components.

Theorem 2.4 *The differential entropy $H(f(\mathbf{x}))$ of a random vector $\mathbf{X} \in \mathbf{R}^N$ for fixed covariance matrix $\mathbf{C}$ is maximized with respect to $f(\mathbf{x})$ when $f(\mathbf{x})$ is a multivariate gaussian density ($f(\mathbf{x}) = \phi(\mathbf{x})$). For any other distribution with the same covariance matrix $\mathbf{C}$, the differential entropy is strictly smaller* [Hyvärinen, 1999*c*].

Proof: [Cover and Thomas, 1991, p. 234] Let the random vector $\mathbf{X} \in \mathbf{R}^N$ have a multivariate gaussian distribution with zero mean (see Appendix A.2, translation of differential entropy) and covariance matrix $\mathbf{C} = E\{\mathbf{X}\mathbf{X}^T\}$, then $H(f(\mathbf{x})) = \frac{N}{2} + \frac{1}{2}(\log(2\pi)^N |\det \mathbf{C}|)$, according to Equation (2.34).

Let $g(\mathbf{x})$ be any density satisfying $\int\limits_{\mathbf{R}^N} g(\mathbf{x}) x_i x_j \, \mathrm{d}\mathbf{x} = c_{ij}$, for all i, j. Let $\phi_{\mathbf{C}}(\mathbf{x})$ be the density of a multivariate gaussian random vector $\mathbf{X}$ as given in (2.33), where $\boldsymbol{\mu}$ is set to zero. Note that $\log \phi_{\mathbf{C}}(\mathbf{x})$ is a quadratic form and $\int\limits_{\mathbf{R}^N} x_i x_j \phi_{\mathbf{C}}(\mathbf{x}) \mathrm{d}\mathbf{x} = c_{ij}$.

Since the information can be defined using differential entropies, see Theorem 2.3, and the information is greater equal zero, see the proof of 'Information is greater equal zero' (Theorem A.1), it follows that,

$$0 \quad \leq \quad \int\limits_{\mathbf{R}^N} g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\phi_{\mathbf{C}}(\mathbf{x})} \mathrm{d}\mathbf{x}.$$

Because $\int\limits_{\mathbf{R}^N} g(\mathbf{x}) \log g(\mathbf{x}) \, \mathrm{d}\mathbf{x} = -H(g(\mathbf{x}))$, we can write

$$= \quad -H(g(\mathbf{x})) - \int\limits_{\mathbf{R}^N} g(\mathbf{x}) \log \phi_{\mathbf{C}}(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Since $g(\mathbf{x})$ and $\phi_{\mathbf{C}}(\mathbf{x})$ yield the same moments of the quadratic form $\log \phi_{\mathbf{C}}(\mathbf{x})$, since $\int\limits_{\mathbf{R}^N} g(\mathbf{x}) x_i x_j \, \mathrm{d}\mathbf{x} = c_{ij}$ and $\int\limits_{\mathbf{R}^N} x_i x_j \phi_{\mathbf{C}}(\mathbf{x}) \mathrm{d}\mathbf{x} = c_{ij}$, we can substitute $\int\limits_{\mathbf{R}^N} g(\mathbf{x}) \log \phi_{\mathbf{C}}(\mathbf{x}) \mathrm{d}\mathbf{x}$ by $\int\limits_{\mathbf{R}^N} \phi_{\mathbf{C}}(\mathbf{x}) \log \phi_{\mathbf{C}}(\mathbf{x}) \mathrm{d}\mathbf{x}$ and write

$$\begin{aligned} 0 \quad &\leq \quad -H(g(\mathbf{x})) - \int\limits_{\mathbf{R}^N} \phi_{\mathbf{C}}(\mathbf{x}) \log \phi_{\mathbf{C}}(\mathbf{x}) \mathrm{d}\mathbf{x} \\ 0 \quad &\leq \quad -H(g(\mathbf{x})) + H(\phi_{\mathbf{C}}(\mathbf{x})). \end{aligned} \tag{2.36}$$

As a result, we get

$$H(g(\mathbf{x})) \quad \leq \quad H(\phi_{\mathbf{C}}(\mathbf{x})), \tag{2.37}$$

■

Of all distributions with the same variance, the gaussian distribution maximizes the differential entropy. Consequently, the differential entropy of the gaussian distribution gives a good bound on the entropy in terms of the variance of the random variable [Cover and Thomas, 1991, p. 234].

Further properties of differential entropy like differential entropy of a transformed random variable can be found in Appendix A.2.

Another important information-theoretic function is the negentropy (negative entropy) $J(f(x))$ measuring the nongaussianity of a random variable X [Comon, 1994]. The negentropy can be regarded as a normalized version of the differential entropy.

Definition 2.13 *Regard a random variable X with assumed density $f(x)$. Further regard the gaussian density $\phi(x)$ assuming to have the same expectation and variance as $f(x)$, then the negentropy $J((f(x))$ is defined as*

$$J(f(x)) = H(\phi(x)) - H(f(x)). \tag{2.38}$$

The negentropy is always nonnegative, and zero only for gaussian random vectors. The proof that the negentropy will always be nonnegative can be found in Appendix A.3. Negentropy has the additional interesting property that it is invariant for linear transformations. This proof can also be found in Appendix A.3. Moreover, since information-theoretic functions are more theoretically that practically used functions, approximations of information-theoretic functions by kurtosis through Taylor expansions can be found in Appendix A.4.

2.3.2 Measuring Nongaussianity by Kurtosis

Since it is assumed that the source signals in ICA should have nongaussian distributions, a classical measure of nongaussianity will be introduced, namely the kurtosis. The kurtosis of a random variable X is defined by

$$kurt(X) = E\{X^4\} - 3(E\{X^2\})^2. \tag{2.39}$$

If it is assumed that X has zero-mean and variance equal to one the kurtosis simplifies to $E\{X^4\} - 3$, which shows that the kurtosis is a normalized version of the fourth moment $E\{X^4\}$. If X is gaussian distributed, the fourth moment equals $3(E\{X^2\})^2$. Thus

kurtosis is zero for a gaussian random variable, and consequently kurtosis can be used as a measure of gaussianity. The kurtosis can be either positive or negative for nongaussian variables. Random variables that have negative kurtosis are called subgaussian, and those with positive kurtosis are called super-gaussian.

3 Functional Magnetic Resonance Imaging

3.1 Functional Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a method used to visualize the inside of living organisms. It is primarily used to demonstrate pathological or physiological alterations of living tissues and is a commonly used form of medical imaging.

Magnetic resonance tomography (MRT) is based on the principle of magnetic spin resonance. Atomic nuclei with odd spin numbers (essentially, an unpaired proton and neutron) have a spin and therefore a magnetic dipole moment. Since the hydrogen nuclei has a relatively large magnetic moment and often appears in biological systems, this hydrogen nuclei is used for imaging in humans. In a magnetic field the spins within the tissue all arrange in one of two opposite directions, namely parallel or antiparallel to the magnetic field. Since the parallel arrangement is energetically more efficient the dipoles most frequently arrange in parallel direction. The magnetic dipole moments of the nuclei then precess around the axial field. Precessing is the slow movement of the nuclei around their axes. The frequency with which the dipole moments precess is called the Lamor frequency. This frequency increases proportional to the strength of the magnetic field. Common used field strength for human research ranges from 0.5 to 3 Tesla. For comparison, the average earth's magnetic field is only around 50 μTesla.

MRI measures the signals arising in tissue by relaxation processes after radiofrequency pulses (RF pulse) are applied. Therefore, the tissue in the magnetic field is briefly exposed to RF pulses in a plane perpendicular to the magnetic field, causing some of the magnetically aligned hydrogen nuclei to assume a temporary non-aligned high-energy state. Images can then be created from the acquired data using discrete fourier transforms resulting in gray values. To understand MRI contrasts in the tissue the time constants involved in the relaxation processes after the RF pulse must be considered. As the high-energy nuclei relax

and realign, they emit energy at rates which are recorded to provide information about their environment in the tissue. The realignment of spins with the magnetic field is termed longitudinal relaxation and the time required for a certain percentage of the tissue nuclei to realign is termed $T1$. This time is about 1 *sec*. $T2$-weighted imaging relies upon local dephasing of spins following the application of the transverse energy pulse. This is the transverse relaxation time typically <100 *msec* for tissue. An important variant of the $T2$ technique is called $T2^*$ imaging. $T2^*$ imaging employs a spin echo technique, in which spins are refocused to compensate for local magnetic field inhomogeneities. Applications of $T2^*$ imaging include functional MRI (fMRI). In the brain, $T1$-weighting causes fiber tracts like nerve connections to appear white, accumulations of neurons to appear gray, and cerebrospinal fluid to appear dark. The contrast of "white matter," "gray matter" and "cerebrospinal fluid" is reversed using $T2$ or $T2^*$ imaging.

MRI can be used to investigate mainly the anatomy of brain structures, but besides looking at structural scans of the brain only, MRI can additionally be used to look at functional activities of the brain, which is referred to as fMRI.

FMRI measures signal changes in the brain that are due to changes of neural activity. In fMRI scans the brain is scanned at low resolution but at a rapid rate, typically once every 2-3 seconds. Increases in neural activity cause changes in the magnetic resonance signal via a mechanism called the BOLD (blood oxygen level-dependent) effect. Activated brain regions need more energy and thus consume more oxygen and glucose. The neuronal system overcompensates for this by increasing the amount of oxygenated hemoglobin (blood with a high level of oxygen) relative to deoxygenated hemoglobin (blood with a low level of oxygen) in that regions. Because deoxygenated hemoglobin attenuates the magnetic resonance signal, the response leads to a signal increase that is related to the neural activity, see Ogawa et al., 1990, Friston, 1996, and Lange, 1996, for summaries. Thus, by recording slight changes in capillary blood oxygenation, fMRI intends to measure neuronal activity indirectly in response to designed stimuli [Lange, 1996].

Consequently, fMRI reveals which parts of the brain are active in solving certain tasks. The type of scanning technique most commonly used is echo planar imaging (EPI), which allows for fast measurement of the signal. The spatial resolution of the activation is about 1-5 millimeters. The temporal resolution, meaning the time distance between two data points, is only about 2-3 seconds or even more seconds. This means, that fMRI has a good spatial resolution but poor temporal resolution, compared with methods that measure neuronal activity more directly such as electroencephalography (EEG) or magnetoencephalography

(MEG).

One of the major advantages of fMRI over other brain mapping techniques is the possibility to look at the relationship between brain anatomy and function of areas in the brain noninvasively. However, the BOLD fMRI signal is an indirect measure of the underlying neuronal activity and it reflects the sum of the activities of a large group of neurons only.

3.2 FMRI Time Series

Most fMRI experiments are designed as blocked-designed experiments in alternating stimulus blocks and resting blocks. In stimulus blocks the sensory impulses, i.e. visual or auditory stimuli, are presented. The stimulus block may last about 20 to 40 seconds followed by a resting block of at least 20 seconds. The stimulation causes neural activity in some regions of the brain, which leads to different gray values at a given spatial location of the fMRI image measured at corresponding time points of stimulation. During the resting block the BOLD-signal returns to baseline, whereby the increase and decrease of the signal is time-delayed to the stimulation protocol.

This time course of the fMRI signal is known as the hemodynamic response function (HRF), which is the response to a temporary increase in neuronal activity. The HRF goes through different stages. In general, the fMRI signal is characterized by a delayed increase after the onset of stimulation, reaching a plateau level after about 6 seconds, and decreases slowly to baseline after the offset of stimulation in about 10 - 15 seconds. Sometimes an undershoot below baseline is found, see Heeger and Ress, 2002 and Jezzard, 2001 for further explanations. Figure 3.1 shows an exemplary HRF normalized in $[0, 1]$, where gray blocks indicate stimulation and white blocks indicate resting.

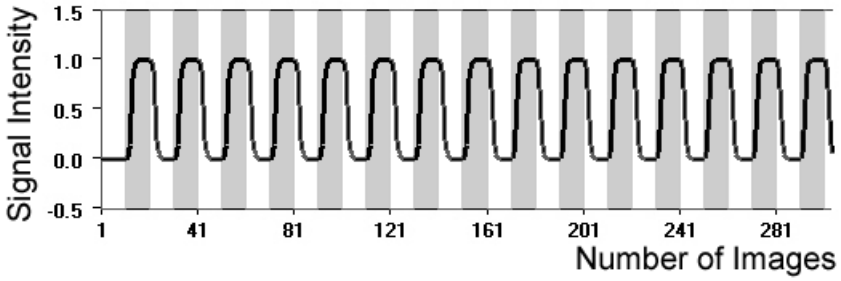

Figure 3.1: Time course of hemodynamic response function

Using a block design has one major advantage over event-related designs, where only single events are measured: The increase in fMRI signals in response to a stimulus is additive. Meaning that the amplitude of the HRF increases when multiple stimuli are presented in rapid succession.

During the fMRI measurement multiple slices of the brain are recorded, thereby each slice consists of a number of pixels. Consequently, the fMRI images are composed of three-dimensional data points (voxels). For each voxel $\mathbf{v}^i$, $i = 1, \ldots, N$, where N is the total number of voxels, at the anatomical coordinate (v_x^i, v_y^i, v_z^i) in the human brain, we observe the gray value $x_i(t)$ for each time point t $(t = 1, \ldots, T)$. It is assumed that the images are recorded at equidistant time points. The gray value can be represented as a function of the induced brain signal.

These signals can be seen as realizations of stochastic processes. The measured fMRI signal has a temporal and spatial structure at several time and length scales, which can be analyzed by different signal processing strategies that emphasize either the spatial or the temporal aspects [Lange et al., 1999].

3.3 FMRI Time Series Regarded as Stochastic Processes

The voxel values may be treated as continuous random variables and the fMRI time series can be considered as a single observation of a stochastic process. Since time series may have some structure, inferences about properties of the underlying processes are made from a single realization, i.e. a single observation at each time. The gray value x_i for every voxel is a random variable at every time point t. The process X_t is a discrete-time and continuous-state process. Friston et al., 1991 suppose that voxel intensities are generated by a continuous, stationary and isotropic stochastic process. 'Continuous' means that the process involves continuous random variables, 'stationary' means that the properties of the process do not depend on its location in either time or space [Beyer et al., 1988], and 'isotropic' means that the process does not prefer a particular directional orientation in space. But Turner and Donald, 2005 point out that fMRI time series are not stationary over time and exhibit spatial correlations. They claim that fMRI time series are structured and often affected by trends and variations. They studied the temporal stationarity and spatial consistency of fMRI noise using ICA and revealed nonstationary processes.

3.4 Analyzing fMRI Data

Statistical methods are used to get activation clusters in the cortex caused by one or different experimental conditions. In fMRI, hypothesis-based and data-based methods are used to investigate different BOLD signals from the acquired time series. Hypothesis-based methods enable detection and signal characterization through an estimation procedure that is repeated identically at each individual voxel [Bandettini et al., 1993, Friston, 1996]. The general linear model (GLM) [Worsley and Friston, 1995, Friston, 1996] belongs to the hypothesis-based methods. It works by cross correlating the time course of each voxel with a reference function modelling the hypothetical HRF of the signal [Bandettini et al., 1993]. This method tests if a voxel is affected by the cognitive task or not. Finally a t-test verifies whether there exist significant differences between stimulation and resting condition. Further details of this method can be found in the next chapter. The GLM has a lot of restrictions concerning accuracy and discrimination power [Baumgartner et al., 2000]. Moreover, hypothesis-based approaches are often univariate, i.e. they test each single brain voxel independently. They neither exploit at all nor fully characterize the co-activation of different voxels. This lack of consideration of spatial interactions inspired the utilization of data-based methods for detection and estimation of spatial activation and temporal dynamics of the brain.

Data-based techniques have been verified and adopted for functional connectivity pattern analysis of distributed regions in the brain during cognitive tasks, such as human memory [Fletcher et al., 1996] and resting state [van de Ven et al., 2004]. This task is accomplished by estimating suitable second- or higher-order statistics and considering relationships among subsets of brain voxels or time points. These data-based methods comprise: principal component analysis (PCA) [Bullmore et al., 1996], independent component analysis (ICA) [McKeown et al., 1998*a*], and cluster algorithms [Filzmoser et al., 1999, Goutte et al., 1999].

These methods are often multivariate and try to aggregate the voxels in spatio-temporal patterns of activity based on a common time course and a common spatial distribution of a given effect. In general, the relationships between voxel time courses are estimated in the spatial covariance of the measured signals. In other words, data-based methods try to find common hidden characteristics in the data often assuming that neighboring voxels are not really independent samples. These methods are useful when the expected neuronal activation cannot be determined in advance, e.g. by presentation of complex stimuli or by temporal displacements, which cannot be found with simple cross correlations. The char-

acteristics of hypothesis-based and data-based methods are summarized in Table 3.1. The GLM, PCA, as well as classical time series analyzing methods will be described in the next chapter.

Table 3.1: Hypothesis-based and data-based methods for analyzing fMRI data

Hypothesis-based methods	Data-based methods
• require *a priori* knowledge of the time course of the hemodynamic response • assume homogeneity of variance of the signals across different brain regions • allow tests of statistical significance within an assumed data and noise model	• require minimal space and time assumptions • explore time courses and spatial distributions of the data • provide no noise model for statistical testing • reveal unforseen activation (time-varying, site-dependent)

4 Classical Methods for Analyzing fMRI Time Series

As mentioned in the previous chapter the fMRI data can be analyzed by hypothesis-driven methods, like the general linear model (GLM) or data-driven methods, like the principal component analysis (PCA). These two methods are considered in more detail in this chapter. Moreover, classical time series analyzing methods are introduced in the last section of this chapter.

4.1 General Linear Model

The general linear model (GLM) is a hypothesis-driven analytical method for detecting activations in fMRI time series. The method can be seen as an extension of linear to multiple regressions. The usage of GLM for fMRI data was motivated by Friston et al., 1995 and Worsley and Friston, 1995. The GLM for a time series of a voxel $\mathbf{v}^i$ $(i = 1, \ldots, N)$ can be written as

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ldots & \vdots \\ a_{T1} & \cdots & a_{Tk} \end{pmatrix} \begin{pmatrix} \gamma_{i1} \\ \vdots \\ \gamma_{ik} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix} = \mathbf{A}\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \tag{4.1}$$

where $\mathbf{x}_i$ is a column vector of the observed response, i.e. the time series of voxel $\mathbf{v}^i$. In general the time series is mean corrected. The matrix $\mathbf{A}$ is a design matrix whose columns are the covariates, i.e. predictors of different experimental conditions. The vector $\boldsymbol{\gamma}_i$ is a column vector of parameters defining the contribution of each column of the design matrix to the model. The errors $\boldsymbol{\varepsilon}_i$ are assumed to be independent and identically gaussian distributed with unit variances.

The multivariate model for N voxels can be written as

$$\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N) = \mathbf{A}(\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_N) + (\boldsymbol{\varepsilon}_1 \cdots \boldsymbol{\varepsilon}_N) = \mathbf{A}\boldsymbol{\Gamma} + \mathbf{E}. \tag{4.2}$$

For each voxel $\mathbf{v}^i$ the same matrix $\mathbf{A}$ is assumed. The parameters γ for one voxel can be estimated by minimizing the sum of squares of the observed errors (residuals) e (the index i was omitted)

$$\begin{aligned} e^T e &= (\mathbf{x} - \mathbf{A}\gamma)^T(\mathbf{x} - \mathbf{A}\gamma) \\ &= \mathbf{x}^T\mathbf{x} + \gamma^T\mathbf{A}^T\mathbf{A}\gamma - 2\gamma^T\mathbf{A}^T\mathbf{x} \;\rightarrow\; \min. \end{aligned}$$

With

$$\frac{\partial(e^T e)}{\partial\gamma} = 2\mathbf{A}^T\mathbf{A}\gamma - 2\mathbf{A}^T\mathbf{x} = 0 \tag{4.3}$$

the following equation is obtained

$$\mathbf{A}^T\mathbf{A}\gamma = \mathbf{A}^T\mathbf{x}. \tag{4.4}$$

This equation provides the estimates of the parameters γ by least squares estimation:

$$\hat{\gamma} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}, \tag{4.5}$$

if $(\mathbf{A}^T\mathbf{A})^{-1}$ exists. Now statistical inferences about the effects of interest are addressed, e.g. the significance of the regression coefficients for the predefined reference waveforms in $\mathbf{A}$. The null hypothesis can be formulated that the effects embodied in $\mathbf{A}$ are not present. This can be tested with the t statistic using linear compounds or contrasts of the parameters estimates $\hat{\gamma}$. The contrasts are given in a row vector $\mathbf{c} = [c_1 \;\cdots\; c_k]$. This is a set of weights that sum to zero. In some special cases it might be interested to analyze only the influence of one predictor then there is only one '1' in the contrast vector. The t statistic is formulated as

$$t = \frac{\mathbf{c}^T\hat{\gamma}}{\sqrt{\hat{\sigma}^2\mathbf{c}^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c}}}, \tag{4.6}$$

where $\hat{\sigma}^2 = (\mathbf{x} - \mathbf{A}\hat{\gamma})^T(\mathbf{x} - \mathbf{A}\hat{\gamma})/(T - r)$, with r is the rank of matrix $\mathbf{A}$ ($r = rg(\mathbf{A})$). The rank of a matrix $\mathbf{A}$ is the maximum number of columns or rows of $\mathbf{A}$ which are linearly independent. The statistic t is t-distributed with $T - r$ degrees of freedom.

Having detected voxels significantly activated by one or more effects, these voxels are then color coded in the anatomical brain map.

4.2 Principal Component Analysis

In many situations where there is a large number of variables in the database it is very likely that subsets of variables are highly correlated with each other. Applied to fMRI data there

is a large number of voxels N suggesting that neighboring voxels and voxels from functionally connected regions in the brain are correlated to each other. Principal component analysis (PCA) was introduced by HOTELLING, 1936, as a mathematical procedure that transforms a number of possible correlated variables into a smaller number of uncorrelated variables called principal components, assuming gaussian distribution of the variables. The objective of PCA is to reduce the dimensionality (number of variables) of a dataset but retain most of the original variability in the data.

The mathematical background for PCA is the following. PCA is a method for data reduction based on second-order statistics, i.e. variances and covariances of the random variables under the assumption of gaussian distributed random variables. Consider a $N \times T$ data matrix $\mathbf{X}$, where N is the number of possibly correlated characteristics, i.e. possibly correlated voxels. $\mathbf{X}_0$ is the standardized data matrix having empirical mean zero and variances equal to one for each row. The empirical correlation matrix $\mathbf{R}$ of N observed characteristics is given by

$$\mathbf{R} = \frac{1}{N-1}\mathbf{X}_0\mathbf{X}_0^T = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{12} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{1N} & \cdots & \cdots & 1 \end{bmatrix}. \tag{4.7}$$

Having a standardized data matrix $\mathbf{X}_0$, the correlation matrix $\mathbf{R}$ equals to the covariance matric $\mathbf{C}$. PCA transforms the correlation matrix in that way that N observed correlated variables are represented by N latent uncorrelated components of which M ($M \leq N$) essential components are chosen. These uncorrelated components are called principal components (PC) and are linear combinations of the original data which are defined by the following formula:

$$\begin{aligned} PC_1 &= \mathbf{w}_1^T\mathbf{X}_0 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1N}X_N \\ PC_2 &= \mathbf{w}_2^T\mathbf{X}_0 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2N}X_N \\ &\vdots \\ PC_M &= \mathbf{w}_M^T\mathbf{X}_0 = w_{M1}X_1 + w_{M2}X_2 + \cdots + w_{MN}X_N, \end{aligned} \tag{4.8}$$

where $X_1, \ldots, X_N$ are row vectors of the standardized data matrix $\mathbf{X}_0$ and w_{ji} are weighting coefficients. The weighting coefficient vectors $\mathbf{w}_1, \ldots, \mathbf{w}_M$ are chosen such that they satisfy the following conditions: The first principal component PC_1 is a linear combination $\mathbf{w}_1^T\mathbf{X}_0$ that maximizes $Var(\mathbf{w}_1^T\mathbf{X}_0)$ with $||\mathbf{w}_1|| = 1$. The second principal component PC_2 is a linear combination $\mathbf{w}_2^T\mathbf{X}_0$ that maximizes $Var(\mathbf{w}_2^T\mathbf{X}_0)$ with $||\mathbf{w}_2|| = 1$ and

$Cov(\mathbf{w}_1^T\mathbf{X}_0, \mathbf{w}_2^T\mathbf{X}_0) = 0$ which means that PC_2 is orthogonal to PC_1. The j-th principal component PC_j $(j = 1, \ldots, M)$ is a linear combination $\mathbf{w}_j^T\mathbf{X}_0$ that maximizes $Var(\mathbf{w}_j^T\mathbf{X}_0)$ with $||\mathbf{w}_j|| = 1$ and $Cov(\mathbf{w}_k^T\mathbf{X}_0, \mathbf{w}_j^T\mathbf{X}_0) = 0 \; \forall \; k < j$. Thereby Cov and Var are the empirical covariance and variance derived from the T repeated observations.

This says that the principal components are those linear combinations of the original variables which maximize the variance of the linear combination and which have zero covariance (and hence zero correlation) with the previous principal components.

In other words, the extraction of principal components amounts to a variance maximizing rotation of the original variables. It can be shown that PCA corresponds to an eigenvalue problem where the coefficient vector $\mathbf{w}_j$ of component PC_j corresponds to the standardized eigenvector to the j largest eigenvalues of the empirical correlation matrix R $(PC_j = \mathbf{w}_j/\sqrt{\lambda_j})$, where λ_j $(j = 1, \ldots, M)$ are the eigenvalues, sorted in decreasing order. The proportion of variance of the standardized characteristics explaining by the j-th component PC_j is given by

$$\gamma_j = \frac{\lambda_j}{\sum_{j=1}^{M} \lambda_j}. \tag{4.9}$$

An interesting question remains, namely, how many components to retain? Or, how to choose M? A criterion proposed by Kaiser, 1960, states to retain only components with eigenvalues greater than 1. This methods is the one most widely used. A graphical method is the *scree plot* first proposed by Cattell, 1966. The eigenvalues are drawn in a line plot. Cattell suggest to find the place where the smooth decrease of eigenvalues appears to level of to the right of the plot, i.e. there is an obviously deviation in the course of the eigenvalues.

PCA is a method that finds the representation using only the information contained in the covariance matrix of the data. The use of second-order techniques is to be understood in the context of the classical assumption that the random variables are of a gaussian distribution. If a random variable has gaussian distribution, its distribution is completely determined by its first- and second-order moments, where the expectation is already eliminated by the standardization here. Thus it is useless to include any other information for gaussian random variables.

PCA is directly related to another common used technique, the singular value decomposition (SVD). In PCA the principal components are usually calculated from the correlation matrix

which is equal to the covariance matrix for standardized data. In SVD the data matrix $\mathbf{X}$ itself and not the correlation matrix R is decomposed. The SVD for j ($j = 1, \ldots, M$) uncorrelated sources of the data matrix $\mathbf{X}$ ($\mathbf{X} = [x_{i,t}]_{i=1,\ldots N, t=1,\ldots,T}$) is decomposed as

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \tag{4.10}$$

where $\mathbf{U} = [u_{i,j}]_{i=1,\ldots N, j=1,\ldots,M}$ and $\mathbf{V} = [v_{t,j}]_{t=1,\ldots T, j=1,\ldots,M}$ are orthogonal matrices. The columns of $\mathbf{U}$ contain the eigenvectors of the inner product matrix ($\mathbf{X}\mathbf{X}^T$), which investigates the inter-relationships between voxels. The matrix $\mathbf{\Lambda} = [\lambda_{j,j}]_{j=1,\ldots M}$ is a diagonal matrix with nonnegative elements λ_j in decreasing order, the eigenvalues or singular values. The matrix $\mathbf{V}$ is then computed by

$$\mathbf{V} = \mathbf{X}^T\mathbf{U}\mathbf{\Lambda}^{-1}. \tag{4.11}$$

The columns of the matrix $\mathbf{V}$ contain the eigenvectors of $\mathbf{X}^T\mathbf{X}$ which investigates inter-relationships between the measurements at different time points t.

Detailed introductions to PCA can be found in Hartung and Elpelt, 1995, Jackson, 1991 and Rao, 1964. Applications of PCA to fMRI data and BSS are found in Andersen et al., 1999, Friston et al., 1993, and Mutihac and van Hulle, 2004.

4.3 Time Series Analyzing Methods

The fMRI observations can be regarded as realizations of stochastic processes $X_t = \{X(t), t \in T \subseteq \mathbf{R}\}$. Stochastic processes are described in Fisz, 1980, Beyer et al., 1988, and Papoulis, 1991. At every time point t corresponding to a record of a magnetic resonance image a gray value x_i for every voxel $\mathbf{v}^i = (v_x^i, v_y^i, v_z^i)$ ($i = 1, \ldots, N$) is observed. Therefore, the fMRI observations can be considered as time series $x_i = (x_i(1), \ldots, x_i(T))^T$, $i = 1, \ldots, N$.

To begin with time series one can think about the behavior of an observed time series $x(t)$ as being made up of various components. These components might be a trend component $x_{trend}(t)$, a seasonality component $x_{season}(t)$, and some irregular component $u(t)$. In a time series any or all of these components might be present. To explain these components in more detail, many time series exhibit a tendency to increase or to decrease over quite long periods of time which can be identified as trend. Many business or economic time series consists of quarterly or monthly observations which can be described as seasons, i.e. patterns repeated from year to year. Seasonal patterns in time series occur in general regularly and oscillatory. In addition, many business and economic time series appear to

exhibit oscillatory, or cyclic, patterns unconnected to the seasonal behavior which are not necessarily regular. This behavior can be described by a cyclical component but we will omit this component in our case. The final component is the irregular component which is induced by the multitude of factors influencing the behavior of a time series and whose pattern looks rather unpredictable on the basis of past experience. Further descriptions of these components can be found in SCHLITTGEN and STREITBERG, 1995 and NEWBOLD, 1995. It is now assumed that the observed time series $x(t)$ is represented as a sum of its components by an additive model:

$$x(t) = x_{trend}(t) + x_{season}(t) + u(t). \tag{4.12}$$

In some circumstances the series might be viewed as the product of its components. But we are assuming the additive model for our observed time series.

Now, temporal statistics can be used to characterize the observed time series which will be described in this chapter.

4.3.1 Stationary Process

The characteristic of stationary of a process is very important, since the stationarity of a process or time series is a prerequisite condition for time series characteristics like autocorrelation and autocovariance function, or frequency analysis.

In most cases, stationarity of a time series can be achieved by removing the trend, seasonal, and other cyclic components from the time series. Then the remaining part of the time series should be stationary. A white noise process is said to be stationary.

Since the underlying fMRI time series are considered as realizations of stochastic processes (see Section 3.3), it should be investigated if the time series are stationary. Therefore, the stationarity is defined here.

Definition 4.1 *A stochastic process X_t is called strict-sense stationary if its statistical properties are invariant to a shift of the origin. This means $\forall\, n$ and $\forall\, t_1, \ldots, t_n$, $P(X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n) = P(X(t_1 + \tau) \leq x_1, \ldots, X(t_n + \tau) \leq x_n)$ for any constant τ. A stochastic process is called weakly stationary if its mean is constant: $E\{X_t\} = \mu$ and its autocorrelation function $\rho(t, t + \tau)$ depends only on τ.* [SCHLITTGEN and STREITBERG, 1995]

Regard that, because the autocorrelation function only depends on τ, this implies also constant variance.

With stationarity many properties of time series are invariant of time, i.e. they are not only valid for some part of the time series. The time series has the same expectation and variance at all time points. To make a non-stationary time series to a stationary time series is one of the first tasks in time series analysis. Methods for this are creation of differences or transforming the time series through logarithms (see Schlittgen and Streitberg, 1995).

4.3.2 Autocovariance and Autocorrelation Function

A way of characterizing time series is given by their autocovariance and autocorrelation function (see Schlittgen and Streitberg, 1995, and Diggle, 1995). The empirical autocovariance function $c(t, t+\tau)$ of a time series $x(t)$ is

$$c(t, t+\tau) = \frac{1}{T}\sum_{t=1}^{T}\{(x(t)-\bar{x})(x(t+\tau)-\bar{x})\} \quad \text{for } \tau \geq 0, \tag{4.13}$$

where $\bar{x}$ is the empirical mean of the time series $x(t)$. The autocovariance function is symmetric. For a stationary process the statistical values like expectation and variance are not time-dependent anymore. The autocovariance function is not dependent on the location of the time points but dependent on the time difference τ.

The autocorrelation function $\rho(t, t+\tau)$ is defined as

$$\rho(t, t+\tau) = \frac{c(t, t+\tau)}{c(t, t)}, \tag{4.14}$$

where $c(t, t)$ is the variance of $x(t)$. The graph of the autocorrelation $c(t, t+\tau)$ versus τ is known as the correlogram.

4.3.3 Test for White-Noise Process

After all the trend and seasonal components have been removed from the time series (see Schlittgen and Streitberg, 1995 for methods to remove trend and seasonal components), it is tested if the remaining part is a white noise process. Since white noise processes are stationary, see Schlittgen and Streitberg, 1995. The runs test was introduced by Wald and Wolfowitz (described in Bradley, 1968). A description of the test can be found in Bortz et al., 2000 and Sachs, 1999. The null hypothesis tests if the adjusted time series $u(t)$ has sufficient runs. For the test, the mean or the median can be used as statistics. The empirical mean (or median) m is subtracted from the data:

$x_{t,m} = x(t) - m$, $t = 1, \ldots, T$. The variables $x_{t,m}$ can have two states, either positive or negative. The positive and negative $x_{t,m}$ are counted. T_1 is the number of positive $x_{t,m}$ and T_2 is the number of negative $x_{t,m}$, $T_1 + T_2 = T$. The test statistic r is the number of runs r, that is the number of consecutive sequences of identical states. The statistic W follows asymptotically a gaussian distribution

$$W = \frac{r - \mu_r}{\sigma_r} \sim \Phi(0, 1), \tag{4.15}$$

where

$$\mu_r = 1 + \frac{2T_1T_2}{T} \tag{4.16}$$

and

$$\sigma_r^2 = \frac{2T_1T_2(2T_1T_2 - T)}{T^2(T-1)}. \tag{4.17}$$

The runs test can be used to see if observations occur randomly, i.e. unstructured in time. The null hypothesis is rejected if W exceeds the $\alpha/2$-quantile or the $1-\alpha/2$-quantile of the standard gaussian distribution. This means that there are to many or to less runs.

4.3.4 Test for Gaussian Distribution

To compare the empirical cumulative distribution function of a random variable X with a theoretical distribution of the population, the Kolmogorov-Smirnov test is provided, see Chakravarti et al., 1967 and Bortz et al., 2000 for explanations. Here it shall be tested if the time series are of a gaussian distribution as it is used in the analysis of ICA (see later in Chapter 5), i.e. the hypothesis is tested if the cumulative distribution function $F(x)$ of the random variable X comes from a gaussian distribution $\Phi(x)$

$$H_0 : F(x) = \Phi(x). \tag{4.18}$$

The absolute maximum distance between the empirical distribution function $F_N(x)$ and $\Phi(x)$ is used as test statistic G

$$G = \max_x |F_N(x) - \Phi(x)|. \tag{4.19}$$

If the null hypothesis is valid, G is of a Kolmogorov-Smirnov distribution. If the test statistic G is greater than the predefined critical value, the null hypothesis H_0 will be rejected. The critical value is obtained from corresponding statistical tables.

It should be mentioned that there are other test of testing of gaussian distribution like the Shapiro-Wilks test, conducted by regressing the quantiles of the observed data against that of the best-fitting normal distribution [Shapiro and Wilks, 1983]. The Shapiro-Wilks test is more powerful compared with the Kolmogorov-Smirnov test, i.e. the Shapiro-Wilks test discards the null hypothesis more often.

4.3.5 Frequency Analysis

Another way of characterizing time series is in terms of the frequency analysis, which includes the periodogram and the Fourier transformation. The periodogram is a summary description based on a representation of an observed time series as a superposition of sinus waves of various frequencies [Diggle, 1995]. The periodogram or spectrum is a function $IF(\omega)$ of the frequency ω. For a frequency ω the intensity is given of how strong harmonic waves of this frequency occur in the time series [Schlittgen and Streitberg, 1995]. The intensity function $IF(\omega)$ is given by

$$IF(\omega) = T \cdot |C(\omega) + iS(\omega)|^2 \tag{4.20}$$

with

$$C(\omega) = \frac{1}{T}\sum_{t=1}^{T}(x(t) - \bar{x})\cos(2\pi\omega t) \tag{4.21}$$

and

$$S(\omega) = \frac{1}{T}\sum_{t=1}^{T}(x(t) - \bar{x})\sin(2\pi\omega t). \tag{4.22}$$

The function $C(\omega)$, $S(\omega)$, and $IF(\omega)$ can be derived from a Fourier transform $FT(\omega)$ of the mean-reduced measured time series $x(t) - \bar{x}$, $(t = 1, \ldots, T)$. The empirical Fourier transform is given by [Schlittgen and Streitberg, 1995]:

$$\begin{aligned} FT(\omega) &= \sum_{t=1}^{T}(x(t) - \bar{x})e^{i2\pi\omega t} \\ &= \sum_{t=1}^{T}(x(t) - \bar{x})\cos(2\pi\omega t) + i\sum_{t=1}^{T}(x(t) - \bar{x})\sin(2\pi\omega t). \end{aligned} \tag{4.23}$$

$C(\omega)$ and $S(\omega)$ are then the real and imaginary part of $FT(\omega)$, whereas $IF(\omega)$ is its absolute value, see Schlittgen and Streitberg, 1995. In case of discrete observations frequencies between 0 and 0.5. At that frequencies ω, where the function $FT(\omega)$ shows spikes, underlying periods P of the time series can be detected by $P = 1/\omega$.

4.3.6 Histograms and Probability Density Estimation

Finally the histograms and the probability density estimations are considered, even if they are not related to time series analysis because they assume independency of the time points. It is assumed that the random sample $X(1), \ldots, X(T)$ of a continuous distribution X is given, whereas the unknown density $f(x)$ should be estimated. It is aimed to evaluate the structure of the data like modality, symmetry or skewness. Histograms of the source signals and the observed signals may draw conclusions about the distribution of the random sample. Alternatively, kernel probability density estimations are used to describe the distribution of the signals. In ICA decomposition (see later Chapter 5), in general, the nongaussian distributions are assumed.

Histograms are a graphical representation based on a decomposition of the data $x = (x(1), \ldots, x(T))$ in P disjunct number for each interval of constant length h (see Silverman, 1986). The density estimation $\hat{f}(x)$ of the histogram can be written as

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} Q_p(x, x(t)), \tag{4.24}$$

where h is the classwidth, and $Q_p(x, x(t))$ is an indicator function with

$$Q_p(x, x(t)) = \begin{cases} 1 & \text{if } x \text{ belongs to the same class as } x(t) \\ 0 & \text{elsewhere} \end{cases} . \tag{4.25}$$

In other words, the indicator function counts the observations in each interval and the relative frequencies are computed.

The kernel density estimator can be written as

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} K\left(\frac{x - x(t)}{h}\right), \tag{4.26}$$

where $K(x)$ is the kernel function dependent on the bandwidth h. The kernel function fulfills the characteristics of a density function

$$\int_{-\infty}^{+\infty} K(x)\mathrm{d}x = 1 \quad \text{and} \quad K(x) \geq 0. \tag{4.27}$$

If, for example, the kernel function $K(x)$ is defined as

$$K(x) = \frac{1}{2} Q(|x| \leq 1), \tag{4.28}$$

where $Q(|x| \leq 1)$ is the indicator function for the event $|x| \leq 1$, then the density estimation is given as

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} \frac{1}{2} Q\left(\left| \frac{x - x(t)}{h} \right| \leq 1 \right), \tag{4.29}$$

see SILVERMAN, 1986, for further descriptions. The kernel density function is symmetric about zero and unimodal like the gaussian kernel with a standard gaussian density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{4.30}$$

Further estimates might be based on the Epanechnikov-kernel or the triangular density, see SILVERMAN, 1986. As for histograms, the bandwidth h plays an important role for kernel density estimators. For small bandwidths the density estimations show a rough structure and the density might be undersmoothed. On the other side, for large bandwidths the density estimation is very smooth and important structures in the data might get lost.

The methods described in this chapter are used later for comparing the ICA method to these time series methods for fMRI data and simulated data.

5 Independent Component Analysis

Independent component analysis is a method for blind signal separation formed on the basis of assumed statistical independence of the source signals. The problem of blind source separation or blind signal separation (BSS) appears in many contexts. Blind source separation is a class of explorative tools originally developed for the analysis of images and sound. BSS has received wide attention in various fields such as speech enhancement, geophysical data processing, data mining, wireless communications, image processing, and biomedical signal analysis and processing (EEG, MEG, fMRI). The method is called 'blind' because it aims to recover source signals from mixtures with unknown coefficients. The most simple situation occurs for two speakers speaking simultaneously. Imagine that the mixture of their voices reaches two microphones, and one wants to separate both sources such that each detector registers only one voice. The problem is called the cocktail party problem which can also be extended to N people standing around and chatting with each other. This mixture of signals is recorded by N microphones. Again, the aim is to extract the voices of the speaker (the sources) from the mixture of speech signals without knowing the sources and the mixture process assuming that the voices are independent of each other.

In this project the problem of BSS is applied to the field of functional magnetic resonance imaging (fMRI), especially to fMRI time series, For the fMRI time series it is assumed that the measured signal of neuronal activity are mixed linearly with multiple other signals like noise or movement artifacts, contributing to the measurement. The aim of blind signal separation in fMRI is to detect the intrinsic signals, i.e. the neuronal activity, from the mixed signals measured during the fMRI study. ICA is a statistical approach of transforming multidimensional data into components that are as independent of each other as possible.

5.1 Definition of ICA

The observed signals are assumed to be a linear mixture of realizations of stochastic processes. The signals $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))^T, t = 1, \ldots, T$ have to be decomposed into a set of independent signals, the independent components. In the blind source separa-

tion problem, the underlying mixture model generates the observed vector of signals $\mathbf{x}(t)$ from a vector of sources $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_M(t))^T, t = 1, \ldots, T$ $(M \leq N)$ by

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

where the function f can be any linear or nonlinear function. But in the following we will only consider the linear case which means that the observed signals are generated by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{5.2}$$

where $\mathbf{A} = [a_{ij}]_{i=1,\ldots,N;j=1,\ldots,M}$ is the unknown constant mixing matrix. In matrix notation, the equation is given by

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{5.3}$$

where $\mathbf{X} = [x_{it}]_{i=1,\ldots,N;t=1,\ldots,T}$ is the matrix of observed signals and $\mathbf{S} = [s_{jt}]_{j=1,\ldots,M;t=1,\ldots,T}$ is the matrix of the source signals. As mentioned in Section 2.3, statistical dependence within a fixed number of components can be quantified by means of their information. The information takes into account the whole dependence structure of the variables. Finding a transformation that minimizes the information between the components is a natural way of estimating the independent components [Comon, 1994]. In other words, the independence of signals can be found by minimization of the information of the signals.

Without knowing the source signals $\mathbf{s}(t)$ and the mixing matrix $\mathbf{A}$, ICA aims to recover the original sources from the observations $\mathbf{x}(t)$ by a linear transformation. This is roughly equivalent to estimating the mixing matrix $\mathbf{A}$. Thus, ICA decomposition can be defined as a transformation

$$\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t), \tag{5.4}$$

where $\hat{\mathbf{s}}(t)$ are the estimates of the source signals $\mathbf{s}(t)$ and $\mathbf{W} = [w_{j,i}]_{j=1,\ldots,M;i,\ldots,N}$ is the estimated 'unmixing' matrix. For the matrix $\mathbf{W}$, which is supposed to be orthogonal, it holds that $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, where $\mathbf{I}$ is an $N \times N$ identity matrix. The matrix $\mathbf{W}$ is determined such that the information of the independent components $\hat{\mathbf{s}}(t)$ is minimized. Moreover, the matrix $\mathbf{W}$ should be determined such that $\mathbf{A}\mathbf{W} \approx \mathbf{I}$, which means that $\mathbf{W} \approx \mathbf{A}^{-1}$. In the case that the mixing matrix $\mathbf{A}$ is not square, we use the Moore-Penrose pseudo-inverse as inverse matrix. Often only the quadratic case is investigated, where the number of sources M equals the number of observations N. For more signals than sources $(N > M)$, the quadratic case can be obtained by performing a principal component analysis in advance

that $\mathbf{A}$ is quadratic ($M = N$), whereas it has to be regarded that the number of independent components M is mostly unknown and has to be estimated.

It is impossible to obtain the original sources $\mathbf{s}(t)$ in a unique way because 5.4 contains some redundancies. The estimates $\hat{\mathbf{s}}(t)$ can be determined up to a permutation of indices, a multiplicative constant and the sign [Oja, 1998, Amari et al., 1996]. The reason is that both the signals $\mathbf{s}(t)$ and the matrix $\mathbf{A}$ are unknown. Any scalar multiplier in one of the sources $s_j(t)$ could always be cancelled by dividing the corresponding column a_j in $\mathbf{A}$ by the same scalar. The order of the independent components $\hat{\mathbf{s}}(t)$ cannot be determined, too. The reason is that again both the source signals $\mathbf{s}(t)$ and the matrix $\mathbf{A}$ are unknown, and the order of the terms in the sum in (5.2) can freely be changed, and one can call any of the independent components the first one [Oja, 1998]. Thus,

$$\begin{aligned} \hat{\mathbf{s}}(t) &= \mathbf{W}\mathbf{x}(t) && (5.5)\\ &= \mathbf{W}\mathbf{A}\mathbf{s}(t) && (5.6)\\ &= \mathbf{R}\mathbf{P}\mathbf{s}(t), && (5.7) \end{aligned}$$

where $\mathbf{R} = [r_{j,j}]_{j=1,\ldots,M}$ is the scaling matrix with coefficients only in the diagonal ($r_{jj} \neq 0,\ -\infty < r < \infty$) and $\mathbf{P} = [p_{j,j}]_{j=1,\ldots,M}$ is the permutation matrix. It should be mentioned that there still is an error term, if the estimation of independent components was not perfect. In fMRI data, for instance, the independent components might be ordered according to correlating the signal to a reference time course.

Before we introduce different ICA algorithms we want to consider the aspect of densities of the signals. The densities of the signals are needed for the computations in some ICA algorithms. It must be pointed out, that the densities of the signals can only be regarded as pseudo-densities because they are not densities in the intrinsic sense. Authors introducing ICA algorithms with the densities of the signals like Bell and Sejnowski or Hyvärinen hypothesize that the density of a signal is the intensity in the state space. Thereby the state of the signal at time t depends on $t - 1$.

5.1.1 Identification and Restriction of ICA Algorithms

The identification of the ICA model is widely discussed in Comon, 1994. The basic assumption is the independence of the source signals, this should be explained with a little example. Consider two random variables S_1 and S_2, the source signals, which are supposed to be independent of each other, with $E\{S_1\} = E\{S_2\} = 0$ and $Var\{S_1\} = Var\{S_2\} = 1$. These two

random variables are mixed by a mixing matrix $\mathbf{A}$, which could look like $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ \rho & 1 \end{pmatrix}$. Consequently, the random variables X_1 and X_2 are given by $X_1 = S_1$ and $X_2 = \rho S_1 + S_2$ ($E\{X_1\} = E\{X_2\} = 0$, $Var\{X_1\} = 1$ and $Var\{X_2\} = 1 + \rho^2$). The covariance of X_1 and X_2 is $Cov(X_1, X_2) = E\{S_1(\rho S_1 + S_2)\} = \rho$, so that the covariance matrix $\mathbf{C}_X$ is given by $\mathbf{C}_X = \begin{pmatrix} 1 & \rho \\ \rho & 1 + \rho^2 \end{pmatrix}$. The unknown source signals should be estimated of the observed signals, i.e. X_1 and X_2. This means that an unmixing matrix $\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ must be determined. The estimates are then given by $\hat{\mathbf{S}} = \mathbf{W} \cdot \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. In the ideal case $\mathbf{W} = \mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 \\ -\rho & 1 \end{pmatrix}$. The covariance matrix of the estimated source signals $\mathbf{C}_{\hat{\mathbf{S}}}$ is given by

$$\mathbf{C}_{\hat{\mathbf{S}}} = \mathbf{W}\mathbf{C}_{\mathbf{X}}\mathbf{W}^T = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 + \rho^2 \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{pmatrix},$$

$$\begin{aligned} c_{\hat{\mathbf{S}}}(11) &= w_{11}^2 + 2\rho w_{11} w_{12} + (1 + \rho^2) w_{12}^2 \\ c_{\hat{\mathbf{S}}}(12) &= w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22} \\ c_{\hat{\mathbf{S}}}(21) &= w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22} \\ c_{\hat{\mathbf{S}}}(22) &= w_{21}^2 + 2\rho w_{21} w_{22} + (1 + \rho^2) w_{22}^2. \end{aligned}$$

To obtain independency of the source signals, the coefficients w_{11}, w_{12}, w_{21}, and w_{22} must be estimated that way that

$$(w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22})^2 \rightarrow \min . \tag{5.8}$$

For this, the partial derivatives are set to zero

$$\begin{aligned} \frac{\partial}{\partial w_{11}} &= 2[w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22}](w_{21} + \rho w_{22}) = 0 \\ \frac{\partial}{\partial w_{12}} &= 2[w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22}](\rho w_{21} + (1 + \rho^2) w_{22}) = 0 \\ \frac{\partial}{\partial w_{21}} &= 2[w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22}](w_{11} + \rho w_{12}) = 0 \\ \frac{\partial}{\partial w_{22}} &= 2[w_{11} w_{21} + \rho(w_{12} w_{21} + w_{11} w_{22}) + (1 + \rho^2) w_{12} w_{22}](\rho w_{11} + (1 + \rho^2) w_{12}) = 0. \end{aligned}$$

This is only possible, if $w_{11}w_{21} + \rho(w_{12}w_{21} + w_{11}w_{22}) + (1 + \rho^2)w_{12}w_{22} = 0$. If $w_{12} = 0$, e.g. $S_1 = X_1$, then $w_{11}w_{21} + \rho w_{11}w_{22} = 0$. Here, w_{11} can be chosen arbitrarily, like $w_{11} = 1$, then we have $w_{21} + \rho w_{22} = 0$ which is given with $w_{22} = 1$ and $w_{21} = -\rho$. Consequently $\mathbf{W}$ can be determined as $\mathbf{W} = \begin{pmatrix} 1 & 0 \\ -\rho & 1 \end{pmatrix}$.

Since some elements of the matrix $\mathbf{W}$ are arbitrary chosen, this demonstrates that the independent components are estimated up to the sign, some factors, and the components are estimated without ordering.

Some further fundamental restrictions for the identifiability of the ICA algorithm (in addition to the basic assumption of stochastically independence of the source signals) has to be imposed, see HYVÄRINEN, 1999*c*. The source signals $\mathbf{s}(t)$ should have nongaussian distributions, since higher-order cumulates like kurtosis as they are used in many ICA algorithms are zero for gaussian variables, and consequently the algorithms are not able to optimize the solution. For gaussian distributed source signals the problem can be reduced to a PCA estimation. The number of observed linear mixtures N must be at least as large as the number of independent components M, i.e., $N \geq M$.

5.1.2 Preprocessing the Data

Solving the ICA problem is simplified if the observed mixture vectors $\mathbf{x}(t)$ are first preprocessed without loss of generality of the ICA estimation. The variables are centered by subtracting the mean of the data over time or space, see therefore later Section 5.5. This results in zero-mean vectors $\mathbf{x}_0(t)$. Furthermore, the variables are made uncorrelated and have unit variances this can be done by a singular value decomposition (SVD), see Section 4.2. The standardization results in vectors $\mathbf{z}(t) = \mathbf{Z}\mathbf{x}_0(t)$, which all have mean zero and equal unit variances. The matrix $\mathbf{Z}$ is given by $\mathbf{Z} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^T$, where $\mathbf{\Lambda} = [\lambda_{i,i}]_{i=1,\ldots,N}$ is a diagonal matrix with the eigenvalues of the data matrix $(\mathbf{X}\mathbf{X}^T)$, and $\mathbf{V} = [v_{t,i}]_{t=1,\ldots,T,i=1,\ldots,N}$ is a matrix with the corresponding eigenvectors of $\mathbf{X}^T\mathbf{X}$ at the columns (see Equation (4.11)). At this step the dimension reduction can be performed by selecting only the M $(M \leq N)$ most interesting eigenvectors.

Consequently the estimated independent components $\hat{\mathbf{s}}(t)$ are zero-mean and uncorrelated as well.

In the following sections, especially in Section 5.3 (Algorithms for ICA) it is assumed that the observed signals $\mathbf{x}(t)$ are preprocessed, without denoting them as $\mathbf{z}(t)$.

5.2 Relation between PCA and ICA

In this section we will look closer to the relation between independent component analysis (ICA) and principal component analysis (PCA). ICA can be regarded as an extension to PCA for nongaussian random variables. This means that ICA generalizes the characteristics of a PCA and can be applied to the data even if they are not gaussian distributed. In the other way around ICA algorithm would not work with gaussian distributed random variables since skewness and kurtosis are zero for gaussian variables and the optimization of the ICA algorithms would not work anymore.

Theorem 5.1 *In the case of gaussian random variables the problem of solving ICA is reduced to a PCA.*

Proof: The different ICA algorithms (see Section 5.3) are based on information-theoretic measures and higher-order moments, like skewness or kurtosis which would be zero for gaussian variables. If the sources $\mathbf{s}(t) = (s_1(t), \ldots, s_M(t))^T$ are gaussian distributed, their probability density function $\phi(\mathbf{s})$ according to (2.33) is given by

$$\phi(\mathbf{s}) = \frac{1}{(\sqrt{2\pi})^M |\det \mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{s}-\mu)^T \mathbf{C}^{-1}(\mathbf{s}-\mu)}, \tag{5.9}$$

where $\mathbf{C}$ is the covariance matrix of the signals. Among all densities having a given covariance matrix $\mathbf{C}$, the gaussian density is the one which has the largest differential entropy, see Theorem 2.4:

$$H(\phi(\mathbf{s})) \geq H(f(\mathbf{s})). \tag{5.10}$$

The entropy obtained for a multivariate gaussian density according to (2.34) is

$$\begin{aligned} H(\phi(\mathbf{s})) &= -\int_{\mathbf{R}^M} \phi(\mathbf{s}) \log \phi(\mathbf{s}) \, \mathrm{d}\mathbf{s} \\ &= \frac{1}{2} \left(M + M \log(2\pi) + \log |\det \mathbf{C}| \right). \end{aligned} \tag{5.11}$$

Regarding the entropy for gaussian densities and the assumed density for the sources, the negentropy J is defined as (see (2.38))

$$J(f(\mathbf{s})) = H(\phi(\mathbf{s})) - H(f(\mathbf{s})). \tag{5.12}$$

As we will see from Equations (2.21) and (5.12), the information can be written as:

$$I(f(\mathbf{s})) = J(f(\mathbf{s})) - \sum_{j=1}^{M} J(f(s_j)) + \frac{1}{2} \log \frac{\prod_{j=1}^{M} c_{j,j}}{|\det \mathbf{C}|}, \tag{5.13}$$

where $c_{j,j}$ are the diagonal elements of the covariance matrix $\mathbf{C}$. The relation between negentropy and information can be seen in the following proof:

$$\begin{aligned} J(f(\mathbf{s})) - \sum_{j=1}^{M} J(f(s_j)) &= H(\phi(\mathbf{s})) - H(f(\mathbf{s})) - \left(\sum_{j=1}^{M} H(\phi(s_j)) - \sum_{j=1}^{M} H(f(s_j)) \right) \\ &= I(f(\mathbf{s})) + H(\phi(\mathbf{s})) - \sum_{j=1}^{M} H(\phi(s_j)). \end{aligned} \tag{5.14}$$

Equation (5.13) is then obtained by replacing the gaussian entropies by their values given in (5.11) and

$$H(\phi(s_j)) = \frac{1}{2} \left(M + M \log(2\pi) + \sum_{j=1}^{M} \log c_{j,j} \right) \tag{5.15}$$

and finally, the information is expressed as in (5.13).

In case of gaussian distributed signals $f(\mathbf{s}) = \phi(\mathbf{s})$ and $f(s_j) = \phi(s_j)$ the term $J(f(\mathbf{s})) - \sum_{j=1}^{M} J(f(s_j))$ vanishes in equation (5.14), and Equation (5.13) is reduced to

$$I(f(\mathbf{s})) = \frac{1}{2} \log \frac{\prod_{j=1}^{M} c_{j,j}}{|\det \mathbf{C}|}. \tag{5.16}$$

The information of the signals is solely described by their second-order statistics contained in the covariance matrix $\mathbf{C}$ which is equivalent to the problem of PCA. ■

Because negentropy is invariant for linear transformations (see Theorem A.6 in Appendix A.3), finding an invertible transformation $\mathbf{W}$ that minimizes the information is roughly equivalent to finding directions in which the negentropy is maximized.

From Equation 5.16 it is also obviously that for a diagonal covariance matrix $\mathbf{C}$ or for $\mathbf{C} = \mathbf{I}$, this equation becomes $I(f(\mathbf{s})) = 0$. Consequently an optimum is reached, since variables that have an information of zero are independent of each other as required by ICA.

The similarity between the independent components in ICA and the latent variables in PCA or SVD can also be seen regarding Equations (5.2) and (4.10). This generates the following analogy between PCA and ICA variables, see Lin et al., 2003:

$$\mathbf{U}_{\cdot j} = \frac{\mathbf{A}_{\cdot j}}{|\mathbf{A}_{\cdot j}|} \tag{5.17}$$

$$\mathbf{V}_{.j}^T = \frac{\mathbf{S}_{j.}}{||\mathbf{S}_{j.}||} \tag{5.18}$$

$$\mathbf{\Lambda}_{jj} = \sqrt{\frac{||\mathbf{A}_{.j}\mathbf{S}_{j.}||}{||\mathbf{U}_{.j}\mathbf{V}_{.j}^T||}}. \tag{5.19}$$

One restriction for the identification of the ICA model is that the source signals are not gaussian distributed (see Chapter 5), else the problem of solving ICA decomposition is reduced to solving PCA decomposition.

To summarize, PCA is an effective method for reduction of dimensionality, while ICA is an effective method for extraction of independent features in the data. PCA can be used as a method to determine the number of independent components in advance, given by the eigenvalues and eigenvectors.

5.3 Algorithms for ICA

In the following some important algorithms are presented for the estimation of independent components. To completely identify the mixing matrix and the nongaussian source signals it is needed to go beyond mere covariance measurements, see Section 5.2 for comparison. The ICA algorithms use the information contained in the observed signals $\mathbf{x}(t)$.

Principles for estimating the independent components $\hat{\mathbf{s}}(t)$ from the observations $\mathbf{x}(t)$ can be based on different conditions like estimating by maximization of nongaussianity, by maximum likelihood estimation or by minimization of information. These principles will be described in the following subsections. Thereby two requirements of the source signals are made, namely the source signals are nongaussian distributed and independent of each other. These two requirements are the leading principles of the independent component estimation.

For the optimization of the estimation two general classes of algorithms can be used. The algorithms can be divided into algorithms using higher order statistics or decorrelations, respectively. The most widespread higher order algorithms are the Infomax algorithm by Bell and Sejnowski and the FastICA algorithm by Hyvärinen, whereas the most widespread decorrelation method is that of Molgedey and Schuster, which will be described later.

In general, all the algorithms are aimed to estimate the unmixing matrix $\mathbf{W}$ and compute

the independent components according to (5.4). For describing the algorithms the number of observed signals is restricted to equal the number of estimated signals ($N = M$) and consequently the matrix $\mathbf{W}$ is an $N \times N$ matrix, this might be achieved by performing a PCA in advance to reduce dimensionality. Note that in a typical fMRI data set the number of voxels N is much larger than the number of time points T ($N >> T$). Therefore, the number of possible estimated independent components M is restricted by the number of time points T, because they limit the size of the matrix X.

In practical use of ICA, firstly, initial values for the elements of the unmixing matrix $\mathbf{W}$ are constituted, mostly they are randomly chosen or the identity matrix $\mathbf{I}$ is chosen. The initial values are improved iteratively in the different ICA algorithms. After having estimated the unmixing matrix $\mathbf{W}$, e.g. by NEWTON iteration, the independent components are computed by (5.4).

5.3.1 Jutten-Hérault Algorithm

The pioneering work of blind separation of sources by JUTTEN and HÉRAULT, 1991 was inspired by neural networks. The algorithm was based on cancelling the nonlinear cross-correlations to obtain independent components. Using a recursive fully interconnected neural network with learning abilities, they propose a blind identification procedure, based on the use of higher order moments. The algorithm is based on the restriction that if the random variables are independent, the cross-correlations are zero, under the assumption that the random variables have symmetric densities. The nondiagonal terms of the matrix $\mathbf{W}$ are updated according to

$$\Delta w_{j,i} \propto g_1(\mathbf{s}_j) g_2(\mathbf{s}_i), \text{ for } i \neq j, \tag{5.20}$$

where $g_1(.)$ and $g_2(.)$ are some odd nonlinear functions. For instance the odd nonlinear functions $g_1(x) = x^3$ and $g_2(x) = \tanh^{-1}(x)$ are proposed. These functions g are used as estimates for the unknown source densities. The components are computed at every iteration as $\hat{\mathbf{s}}(t) = (\mathbf{I} + \mathbf{W})^{-1}\mathbf{x}(t)$. The diagonal terms $\hat{w}_{j,j}$ are set to zero. The signals $\hat{s}_j(t)$ then give, after convergence, estimates of the independent components. This algorithm converges only under rather severe restrictions [DELFOSSE and LOUBATON, 1995], therefore this algorithm has not much relevance in practical use. Later much more efficient algorithms were introduced. Further details can be found in JUTTEN and HÉRAULT, 1991. Moreover, the general framework for ICA introduced by JUTTEN and HÉRAULT is most clearly stated in COMON, 1994. Furthermore, COMON, 1994 introduced the concept of ICA and proposed cost functions related to the minimization of information between the sensors.

5.3.2 Algorithms for Maximum Likelihood Estimation

Maximum likelihood estimation is a fundamental method in statistical approximation and can also be applied to ICA estimation. The interpretation of the maximum likelihood estimation is to take those parameter values as estimates that give the highest probability for the observations.

For the random mixture vector $\mathbf{x}$ ($\mathbf{x} = \mathbf{As}$) the density $f(\mathbf{x})$ can be formulated as

$$f(\mathbf{x}) = f(\mathbf{s})|\det \mathbf{A}^{-1}|, \tag{5.21}$$

see Papoulis, 1991 or Hyvärinen et al., 2001*b* for the density of a transformation. Since the estimate of $\mathbf{s}$ is denoted as $\hat{\mathbf{s}}$, and $\hat{\mathbf{s}} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{Wx}$, Equation (5.21) can be written as

$$f(\mathbf{x}) = f(\mathbf{Wx})|\det \mathbf{W}|. \tag{5.22}$$

This equation can be rewritten with the densities $f_i(\hat{s}_i)$ of the independent components and $\mathbf{w}_i$ the column vectors of the unmixing matrix as

$$f(\mathbf{x}) = \left(\prod_{i=1}^{N} f_i(\mathbf{w}_i^T \mathbf{x}) \right) |\det \mathbf{W}|. \tag{5.23}$$

Assuming T observations $(\mathbf{x}(1), \ldots, \mathbf{x}(T))^T$, the likelihood function L for the random vector $\hat{\mathbf{s}}$ as a function of $\mathbf{W}$ is obtained as the product of this density evaluated at the T points

$$L(\mathbf{W}) = \prod_{t=1}^{T} \left(\prod_{i=1}^{N} f_i(\mathbf{w}_i^T \mathbf{x}(t)) \right) |\det \mathbf{W}|. \tag{5.24}$$

The log-likelihood takes the form [Pham et al., 1992]:

$$\log L(\mathbf{W}) = \sum_{t=1}^{T} \sum_{i=1}^{N} \log f_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log |\det \mathbf{W}|. \tag{5.25}$$

To simplify the notation, the sum over the sample index t is denoted by $\tilde{E}$ after dividing the likelihood by T to obtain

$$\frac{1}{T} \log L(\mathbf{W}) = \tilde{E} \left\{ \sum_{i=1}^{N} \log f_i(\mathbf{w}_i^T \mathbf{x}) \right\} + \log |\det \mathbf{W}|. \tag{5.26}$$

A problem of the maximum likelihood estimation, as well as other ICA algorithms, is that

the densities f_i of the independent components are not known. Since the likelihood is a function of these densities, the estimation is in general a nonparametric problem which is dealt in one of the following sections. But some restrictions can be made, to make little assumptions about the densities of the independent components to avoid nonparametric density estimation and estimate the independent components. By HYVÄRINEN et al., 2001*b*, it was shown that in maximum likelihood estimation, it is enough to use just two approximations of the density of an independent component. For each component, one just needs to determine which one of the two approximations is better. For the unknown densities f_i the following log densities g_i, where $g_i = (\log f_i)'$, may be proposed

$$\log g_i^+(s) = \alpha_1 - 2\log\cosh(s) \tag{5.27}$$

$$\log g_i^-(s) = \alpha_2 - [s^2/2 - \log\cosh(s)], \tag{5.28}$$

where α_1, α_2 are positive parameters that are fixed so as to make these two functions logarithms of probabilities. The motivation for these functions is that g_1^+ is a super-gaussian density, because the $\log\cosh$ is close to the absolute value that would give the Laplacian density. The density given by g_i^- is subgaussian, because it is like a gaussian log-density, $-s^2/2$ plus a constant, that has been flattened to the $\log\cosh$ function [HYVÄRINEN et al., 2001*b*].

An algorithm to obtain the maximum likelihood estimation is then given by gradient methods (BELL and SEJNOWSKI). The matrix gradient of the determinant of a matrix is given by

$$\frac{\partial \log|\det \mathbf{W}|}{\partial \mathbf{W}} = \frac{1}{|\det \mathbf{W}|}\frac{\partial|\det \mathbf{W}|}{\partial \mathbf{W}} = (\mathbf{W}^T)^{-1}, \tag{5.29}$$

see HYVÄRINEN et al., 2001*b*. $\mathbf{g}(\hat{\mathbf{s}}) = (g_i(\hat{s}_i), \ldots, g_N(\hat{s}_N))$ is a component-wise vector function with components g_i,

$$g_i = (\log f_i)' = \frac{f_i'}{f_i}. \tag{5.30}$$

The gradient of the log-likelihood in (5.25) is

$$\frac{1}{T}\frac{\partial \log L}{\partial \mathbf{W}} = (\mathbf{W}^T)^{-1} + \tilde{E}\{\mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T\}, \tag{5.31}$$

where the derivative with respect to a matrix is given by the deriving the function L for each element $w_{j,i}$ of $\mathbf{W}$

$$\frac{\partial L}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial L}{\partial w_{11}} & \cdots & \frac{\partial L}{\partial w_{1N}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial w_{M1}} & \cdots & \frac{\partial L}{\partial w_{MN}} \end{bmatrix}. \tag{5.32}$$

This gives the following iteration for the maximum likelihood estimation:

$$\Delta \mathbf{W} \propto (\mathbf{W}^T)^{-1} + \tilde{E}\{g(\mathbf{W}\mathbf{x})\mathbf{x}^T\}, \tag{5.33}$$

where $\Delta \mathbf{W} = \frac{1}{T}\frac{\partial \log L}{\partial \mathbf{W}}$ is the gradient.

For simplicity we will consider the 2-dimensional case for $N = 2$. The matrix $\mathbf{W}$ is a 2×2 matrix

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}.$$

Let $l = \frac{1}{T}\log L(\mathbf{W})$, then Equation (5.26) results in

$$l = \frac{1}{T}\sum_{t=1}^{T}\{\log f_1(w_{11}x_1 + w_{21}x_2) + \log f_2(w_{12}x_1 + w_{22}x_2)\} + \log|w_{11}w_{22} - w_{12}w_{21}|. \tag{5.34}$$

The partial derivatives are given by

$$\begin{aligned}
\frac{\partial l}{\partial w_{11}} &= \frac{1}{T}\sum_{t=1}^{T}\frac{f_1'(w_{11}x_1 + w_{21}x_2)x_1}{f_1(w_{11}x_1 + w_{21}x_2)} \pm \frac{w_{22}}{|w_{11}w_{22} - w_{12}w_{21}|} \\
\frac{\partial l}{\partial w_{12}} &= \frac{1}{T}\sum_{t=1}^{T}\frac{f_2'(w_{12}x_1 + w_{22}x_2)x_1}{f_2(w_{12}x_1 + w_{22}x_2)} \mp \frac{w_{21}}{|w_{11}w_{22} - w_{12}w_{21}|} \\
\frac{\partial l}{\partial w_{21}} &= \frac{1}{T}\sum_{t=1}^{T}\frac{f_1'(w_{11}x_1 + w_{21}x_2)x_2}{f_1(w_{11}x_2 + w_{21}x_2)} \mp \frac{w_{12}}{|w_{11}w_{22} - w_{12}w_{21}|} \\
\frac{\partial l}{\partial w_{22}} &= \frac{1}{T}\sum_{t=1}^{T}\frac{f_2'(w_{12}x_1 + w_{22}x_2)x_2}{f_2(w_{12}x_1 + w_{22}x_2)} \pm \frac{w_{11}}{|w_{11}w_{22} - w_{12}w_{21}|}.
\end{aligned}$$

Composing these four partial derivatives to one equation results in Equation (5.31).

After convergence of the algorithm (5.33) and selecting an appropriate function for g_i in advance, the independent components can be determined by Equation (5.4). This algorithm converges very slow due to the inversion of the matrix $\mathbf{W}$ which is needed in every step of the iteration [Bell and Sejnowski, 1995].

5.3.3 ICA by Minimization of Information

As we know already, the information is a natural measure of the dependence between random variables. This measure takes into account the whole dependence structure of the

variables. Therefore, the information is a criterion for finding an ICA representation.

The information is closely related to the likelihood function. Using Equations (2.35) and (A.11), the information for an invertible linear transformation $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$ is

$$I(f(\hat{s}_1, \ldots, \hat{s}_N)) = \sum_{i=1}^{N} H(f(\hat{s}_i)) - H(f(\hat{\mathbf{s}})) - \log|\det \mathbf{W}|. \tag{5.35}$$

To see the connection between these two functions, the log-likelihood (5.25) is considered. If the densities f_i were equal to the density functions of $(\mathbf{w}_i^T\mathbf{x})$, the first term would be equal to $-\sum_{i=1}^{N} H(f(\mathbf{w}_i^T\mathbf{x}))$. Thus the likelihood would be equal, up to an additive constant, to the negative of information as given in (5.35). Consequently the algorithms for minimizing information are the same as for the maximum likelihood estimation. Moreover, the minimization of information is equivalent to maximization of differential entropy, since information can be expressed in terms of differential entropy, see Equation (2.35) .

The iteration steps will not be further described in this section, since this algorithm is related to Infomax algorithm, see next section.

5.3.4 The Infomax Principle

An important class of algorithms consists of those based on maximization of network entropy, the Infomax (information maximization) algorithm as proposed by Bell and Sejnowski, 1995. They were first explaining the BSS problem from an information-theoretic viewpoint and applying them to separation and deconvolution of sources. This algorithm takes into account that the source signals should be independent of each other. If random variables are independent of each other then their information is minimized, i.e. the information will equal to zero, see Section 2.3.1. The Infomax principle is related to the maximum likelihood estimation in Section 5.3.3 since the information and differential entropy are related to each other in the way, that a minimization of information corresponds to a maximization of differential entropy, see also Section 2.3.1.

Assuming that $\mathbf{x}$ is the input, and the output of the neural network is

$$\hat{\mathbf{s}}_i = g_i(\mathbf{w}_i^T\mathbf{x}), \tag{5.36}$$

where the g_i are some nonlinear scalar functions then the output entropies have to be maximized

$$H(g(\hat{\mathbf{s}})) = H(g_1(\mathbf{w}_1^T\mathbf{x}, \ldots, g_N(\mathbf{w}_N^T\mathbf{x})). \tag{5.37}$$

The Infomax algorithm performs minimization of the information between the densities of the inputs $\mathbf{x}$ and the outputs $\hat{\mathbf{s}}$ which is equivalent to maximization of the output entropies. The information of two multivariate vectors is given in Appendix A.1. Using Equation (A.12) of transforming an entropy results in

$$H(g_1(\mathbf{w}_1^T\mathbf{x}),\ldots,g_N(\mathbf{w}_N^T\mathbf{x})) = H(\mathbf{x}) + E\{\log\left|\frac{\partial\mathbf{G}}{\partial\mathbf{W}}(\mathbf{x})\right|\}, \tag{5.38}$$

where $\mathbf{G}(\mathbf{x}) = (g_1(\mathbf{w}_1^T\mathbf{x}),\ldots,g_N(\mathbf{w}_N^T\mathbf{x}))$ denotes the function defined by the neural network. The derivative is then given by

$$E\{\log\left|\frac{\partial\mathbf{G}}{\partial\mathbf{W}}(\mathbf{x})\right|\} = \sum_{i=1}^{N} E\{\log g_i'(\mathbf{w}_i^T\mathbf{x}) + \log|\det\mathbf{W}|\}. \tag{5.39}$$

This output entropy is of the same form as the expectation of the likelihood in Equation (5.25). Under some conditions, the Infomax algorithm is equivalent to the maximum likelihood approach, see MacKay, 1996, Pham et al., 1992 and Lee et al., 1999*b*. This equivalence requires that the nonlinearities g_i used in the neuronal network (see Equation (5.36)) are chosen as the cumulative distribution functions corresponding to the densities f_i, i.e. $g_i'(.) = f_i(.)$. The density functions f_i of the independent components are replaced by the functions g_i', thus the output entropy is actually equal to the likelihood, meaning that the Infomax principle is equivalent to maximum likelihood estimation.

The connection between the likelihood function and information can further be explained by considering the expectation of the log-likelihood:

$$\frac{1}{T}E\{\log L\} = \sum_{i=1}^{N} E\{\log f_i(\mathbf{w}_i^T\mathbf{x})\} + \log|\det\mathbf{W}|. \tag{5.40}$$

Actually, if the f_i were equal to the actual distributions of $\mathbf{w}_i^T\mathbf{x}$, the first term would be equal to $-\sum_{i=1}^{N} H(f(\mathbf{w}_i^T\mathbf{x}))$. Consequently, the log likelihood would be equal, up to an additive constant, to the negative of the information. On the other hand since $H(f(\mathbf{s})) = \log|\det\mathbf{W}| - H(f(\mathbf{x}))$ [Papoulis, 1991], the likelihood is related to the information by

$$I(f(\mathbf{s})) = H(f(\mathbf{x})) - \log L, \tag{5.41}$$

thus the information is a constant, $H(f(\mathbf{x}))$, minus the log-likelihood. Regarding that large values in the log-likelihood correspond to small values in the information, because the information $I(f(\mathbf{x}))$ is equal to the negative of the entropy $H(f(\hat{\mathbf{s}}))$.

The Infomax algorithm is based on gradient ascent of the objective function. In Bell and Sejnowski, 1995 it was derived the following way:

$$\Delta \mathbf{W} \propto [\mathbf{W}]^{-1} - 2\tanh(\mathbf{W}\mathbf{x})\mathbf{x}^T \tag{5.42}$$

where the tanh function is applied separately on every component of the vector $\mathbf{W}\mathbf{x}$. The tanh-function is used here because it is the derivative of the log-density of the 'logistic' distribution Bell and Sejnowski, 1995. This function works for estimation of most super-gaussian (kurtosis ≥ 0) independent components.

The convergence of the Infomax ICA is very slow. Moreover, the original Infomax ICA with sigmoidal nonlinearities was only suitable for super-gaussian sources. Lee and colleagues realized that a key to generalizing the Infomax algorithm to arbitrary nongaussian sources was to estimate moments of the source signals and to switch the algorithm appropriately. He developed an efficient extended version of the Infomax ICA [Lee et al., 1999*a*] that is suitable for general nongaussian signal. Moreover, this algorithm shows superior convergence speed.

5.3.5 The FastICA Algorithm

The Fixed-Point algorithm or FastICA algorithm [Hyvärinen, 1999*c*] pursues the same goal as the Infomax algorithm [Bell and Sejnowski, 1995] using the concept of normalized differential entropy or negentropy, see Section 2.3.1. By expressing the information in terms of negentropy it is aimed to find an invertible transformation $\mathbf{W}$, which minimizes the information among the signals to obtain independent signals. This is equivalent to finding directions along maximal negentropy of the projected data.

The FastICA algorithm iteration finds a direction, i.e. a unit vector $\mathbf{w}$ such that the projection $\mathbf{w}^T\mathbf{x}$ maximizes nongaussianity. Nongaussianity is measured by the approximation of negentropy, see Appendix A.4. Note that, the variance of $\mathbf{w}^T\mathbf{x}$ is constrained to unity, for preprocessed data this is equivalent to constraining the norm of $\mathbf{w}$ to unity [Hyvärinen, 1999*b*].

The FastICA is based on a fixed-point iteration scheme for finding a maximum of the nongaussianity of $\mathbf{w}^T\mathbf{x}$. And it can be derived as an Newton iteration. In Newton iteration the new value $\mathbf{w}_l$ is computed from the old value $\mathbf{w}_{l-1}$ by iteration according to

$$\mathbf{w}_l = \mathbf{w}_{l-1} - \frac{F(\mathbf{w}_{l-1})}{F'(\mathbf{w}_{l-1})}. \tag{5.43}$$

Denote by g the derivative of some function, for example the derivatives

$$\begin{aligned} g_1(u) &= \tanh(u), \\ g_2(u) &= u\exp(-u^2/2). \end{aligned}$$

With NEWTON iteration it is aimed to find a local maximum, that $E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} = 0$. After choosing an initial (e.g. random) weight vector $\mathbf{w}$, the l-th iteration ($l = 1, 2, \ldots$) of the algorithm is defined as

$$\begin{aligned} \mathbf{w}_l &= \mathbf{w}_{l-1} = \frac{E\{\mathbf{x}g(\mathbf{w}_{l-1}^T\mathbf{x})\}}{E\{g'(\mathbf{w}_{l-1}^T\mathbf{x})\}} \\ E\{g'(\mathbf{w}_{l-1}^T\mathbf{x})\}\mathbf{w}_l &= E\{g'(\mathbf{w}_{l-1}^T\mathbf{x})\}\mathbf{w}_{l-1} - E\{\mathbf{x}g(\mathbf{w}_{l-1}^T\mathbf{x})\} \\ \mathbf{w}_l^* &= E\{\mathbf{x}g(\mathbf{w}_{l-1}^T\mathbf{x})\} - E\{g'(\mathbf{w}_{l-1}^T\mathbf{x})\}\mathbf{w}_{l-1} \\ \mathbf{w}_l &= \mathbf{w}_l^*/||\mathbf{w}_l^*||. \end{aligned} \tag{5.44}$$

The algorithm converged when the product of the old and new values of $\mathbf{w}$ is (almost) equal to 1. If the iteration l did not converges the $l+1$ iteration is performed.

This algorithm estimates just one of the independent components. In order to estimate more than one solution, and up to a maximum of M solutions, the algorithm must be run repeatedly. To prevent different vectors from converging to the same maxima, the outputs $\mathbf{w}_1^T\mathbf{x}, \ldots, \mathbf{w}_M^T\mathbf{x}$ are removed from the matrix after every iteration. A way of achieving decorrelation is a deflation scheme based on a GRAM-SCHMIDT-like decorrelation. This means to estimate the components one by one. Having estimated $l-1$ ($l-1 < M$) independent components, or $l-1$ vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}$ we run the algorithm for $\mathbf{w}_l$, and after every iteration step subtract from $\mathbf{w}_l$ the projections $(\mathbf{w}_l^T\mathbf{w}_{l-1})\mathbf{w}_{l-1}$, of the previously estimated $l-1$ vectors, and then renormalize $\mathbf{w}_l$:

$$\mathbf{w}_l^* = \mathbf{w}_l - \sum_{j=1}^{l-1}(\mathbf{w}_l^T\mathbf{w}_{l-1})\mathbf{w}_{l-1} \tag{5.45}$$

$$\mathbf{w}_l = \mathbf{w}_l^*/\sqrt{\mathbf{w}_l^{*T}\mathbf{w}_l^*}. \tag{5.46}$$

The FastICA algorithms has many advantages comparing to other methods for ICA. One important property is that the convergence is cubic (or at least quadratic), see HYVÄRINEN, 1999*a*. This is in contrast to ICA algorithm based on gradient descent methods, where the convergence is only linear.

5.3.6 Molgedey and Schuster Approach

The Molgedey and Schuster approach [Molgedey and Schuster, 1994] is based on the decorrelation of the variables utilizing that the independent sources have different autocorrelation functions. This approach does not directly fit to the other algorithms, but this algorithm should shortly be mentioned because of completeness. The problem of separating N linear superimposed uncorrelated sources or signals and determining their mixing coefficients is reduced to an eigenvalue problem which requires the simultaneous diagonalization of two symmetric matrices whose elements are measurable time delayed correlation functions.

Since the authors showed that the mixing matrix $\mathbf{A}$ is not necessarily symmetric, it is not sufficient to measure the symmetric correlation matrix $\mathbf{C}$ with $c_{ij} = E\{X_i(t) \cdot X_j(t)\}$, $i, j = 1, \ldots, N$. It is suggested that one should measure the time delayed correlation matrix $\bar{\mathbf{C}}$ with $\bar{c}_{ij} = E\{X_i(t) \cdot X_j(t+\tau)\}$, for $\tau = 1, \ldots, T-1$ additionally. This leads to $N(N+1)$ equations for a predefined τ:

$$c_{ij} = \sum_{k=1}^{N} a_{ik} a_{jk} \lambda_k, \quad \bar{c}_{ij} = \sum_{k=1}^{N} a_{ik} a_{jk} \bar{\lambda}_k \tag{5.47}$$

for the $N(N+1)$ unknowns $a_{i \neq j}$, λ_j, and $\bar{\lambda}_j$. Equation (5.47) shows that by construction the matrix $\mathbf{A}$ this diagonalizes $\mathbf{C}$ and $\bar{\mathbf{C}}$ simultaneously, i.e.

$$\mathbf{\Lambda} = \mathbf{A}^{-1}\mathbf{C}(\mathbf{A}^T)^{-1} \quad \text{and} \quad \bar{\mathbf{\Lambda}} = \mathbf{A}^{-1}\bar{\mathbf{C}}(\mathbf{A}^T)^{-1}. \tag{5.48}$$

But the elements of $\Lambda_{ij} = \lambda_i \delta_{ij}$ and $\bar{\Lambda}_{ij} = \bar{\lambda}_i \delta_{ij}$ are not simple the eigenvalues of the matrices $\mathbf{C}$ and $\bar{\mathbf{C}}$ because generally $\mathbf{A}$ is not an orthogonal matrix. Instead Equation (5.47) leads, after some steps, to the eigenvalue problem

$$\begin{aligned} \mathbf{C} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T \quad &\text{and} \quad \bar{\mathbf{C}} = \mathbf{A}\bar{\mathbf{\Lambda}}\mathbf{A}^T, \\ \mathbf{C}\bar{\mathbf{C}}^{-1} \quad &= \quad \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T(\mathbf{A}^T)^{-1}\bar{\mathbf{\Lambda}}^{-1}\mathbf{A}^{-1} \\ (\mathbf{C}\bar{\mathbf{C}}^{-1})\mathbf{A} \quad &= \quad \mathbf{A}(\mathbf{\Lambda}\bar{\mathbf{\Lambda}}^{-1}). \end{aligned} \tag{5.49}$$

Since $\mathbf{C}\bar{\mathbf{C}}^{-1}$ is usually not symmetric and the diagonal elements of $\mathbf{A}$ are normalized to unity, Equation (5.49) can be solved with standard techniques of numerical linear algebra, for further details see Molgedey and Schuster, 1994.

Using this method, the time-dependency information alone is sufficient to estimate independent components.

5.3.7 Nonparametric ICA

The algorithms for ICA described above are based on some weak assumptions on the source statistics, especially about the density functions of the signals which are completely unknown. But these algorithms may fail when the statistical model is inaccurate. Therefore, a nonparametric ICA algorithm is truly blind to the underlying distribution of the mixed signals [Boscolo et al., 2004, Samarov and Tsybakov, 2004]. As with other ICA algorithm, this algorithm performs a preprocessing and PCA in advance to restrict the computation to the case where the number of observations equals the number of source signals ($M = N$).

Using a nonparametric kernel density estimation the unknown probability density function of source signals $\mathbf{s}(t)$ and the unmixing matrix $\mathbf{W}$ are estimated.

The aim of ICA algorithm is the estimation of $\mathbf{W}$ and thus reconstructing the source signals $\mathbf{s}(t)$ as $\hat{\mathbf{s}}(t)$, see (5.4). The basic principle therefore is the minimization of the information I between the reconstructed signals, for $N = M$:

$$\mathbf{W}_{opt} = \min_{\mathbf{W}} I(\hat{s}_1, \ldots, \hat{s}_N) \tag{5.50}$$

To compute the information, the probability density functions (pdf) of the sources must be known, but the information is difficult to approximate and optimize on the basis of a finite sample. Equivalent to minimizing the information is the maximum likelihood principle when the source distributions are known, see Sections 5.3.2 and 5.3.3.

In nonparametric kernel density estimation the probability density function is directly estimated from the data using a kernel density estimation technique, this means direct evaluation of the function and its derivatives of the elements w_{ji}, $j, i = 1, \ldots, N$. If a sample data of size T is given, the marginal distributions of an arbitrary reconstructed signal are approximated with gaussian kernels as

$$f(\hat{s}_j) = \frac{1}{T\,h} \sum_{t=1}^{T} \phi\left(\frac{\hat{s}_j - \hat{S}_{jt}}{h}\right), \tag{5.51}$$

where h is the kernel bandwidth controlling the smoothness of the functional ($h = 1.06\,T^{-1/5}$ is supposed Silverman, 1986), ϕ is the density of standard gaussian distribution: $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$, and $\hat{S}_{jt}$ are the Kernel centroids $\hat{S}_{jt} = \mathbf{w}_j \mathbf{x}(t) = \sum_{j=1}^{N} w_{ji} X_{it}$ $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $\mathbf{w}_j$ is the j-th row of the matrix $\mathbf{W}$ and $\mathbf{x}(t)$ is t-th column of $\mathbf{X}$.

The gradient ∇ (the vector of first derivatives of the elements $\mathbf{w}_j$) is given by:

$$\nabla f(\hat{s}_j) = \frac{1}{T\,h^2} \sum_{t=1}^{T} \mathbf{x}(t)(\hat{s}_j - \mathbf{w}_j \mathbf{x}(t)) \phi \left(\frac{\hat{s}_j - \mathbf{w}_j \mathbf{x}(t)}{h} \right). \tag{5.52}$$

Using the maximum likelihood approach and a sample data $\mathbf{x}(k)$, $k = 1, \ldots, T$, the expectation in Equation (5.25) can be approximated by

$$\begin{aligned} \frac{1}{T} \log L(\mathbf{W}) &= \frac{1}{T} \sum_{j=1}^{N} \sum_{k=1}^{T} \log f(\mathbf{w}_j \mathbf{x}(k)) + \log |\det \mathbf{W}| & (5.53) \\ &\approx \frac{1}{T} \log L_0(\mathbf{W}) + \log |\det \mathbf{W}|, & (5.54) \end{aligned}$$

where $L_0(\mathbf{W})$ is a likelihood function obtained by replacing the marginal probability density functions f with their kernel density estimates

$$\begin{aligned} \log L_0(\mathbf{W}) &= \sum_{j=1}^{N} \sum_{k=1}^{T} \log \left[\frac{1}{T\,h} \sum_{t=1}^{T} \phi \left(\frac{\hat{s}_j - \hat{S}_{jt}}{h} \right) \right] & (5.55) \\ &\approx \sum_{j=1}^{N} \sum_{k=1}^{T} \log \left[\frac{1}{T\,h} \sum_{t=1}^{T} \phi \left(\frac{\mathbf{w}_j(\mathbf{x}(k) - \mathbf{x}(t))}{h} \right) \right], & (5.56) \end{aligned}$$

where $\hat{s}_j = \mathbf{w}_j \mathbf{x}(k)$, $k = 1, \ldots, T$ and $\hat{S}_{jt} = \mathbf{w}_j \mathbf{x}(t)$. The optimization problem with the likelihood function $L(\mathbf{W})$ can be posed as

$$\log L(\mathbf{W}) = \min_{\mathbf{W}} \sum_{j=1}^{N} \sum_{k=1}^{T} \log \left[\frac{1}{T\,h} \sum_{t=1}^{T} \phi \left(\frac{\mathbf{w}_j(\mathbf{x}(k) - \mathbf{x}(t))}{h} \right) \right] + \log |\det \mathbf{W}| \tag{5.57}$$

$$\text{with } ||\mathbf{w}_j|| = 1,\ j = 1, \ldots, N. \tag{5.58}$$

The matrix $\mathbf{W}$ can be initialized with random elements w_{ji} for $i, j = 1, \ldots, N$. The optimization algorithm can be performed using a NEWTON algorithm with an objective function $L(\mathbf{W})$ and the objective function's derivative $\nabla L(\mathbf{W})$ based on a fast fourier transformation. Further details can be taken from BOSCOLO et al., 2004.

Having estimated the elements of the matrix $\mathbf{W}$, the estimates $\hat{\mathbf{s}}(t)$ can be computed by 5.4.

5.3.8 Further ICA Algorithms

Besides these classical algorithms there exists a variety of modified algorithms. The problem of the classical algorithms is that the densities $f_i(.)$ of the source signals are unknown.

Therefore, the density functions are either nonparametrically estimated or assumed to have a certain density function $g_i(.)$. Amari et al., 1996, improved the Infomax ICA algorithm by using the natural gradient, which was also discovered by Cardoso, 1997. The original Infomax ICA algorithm with sigmodial nonlinearities was only suitable for super-gaussian sources. As mentioned already, Lee et al., 1999*a*, propose an efficient extended Infomax algorithm that is able to blindly separate mixed signals with super- and subgaussian source distributions (see description of kurtosis in Section 2.3.2). In addition this group developed an algorithm for more sources than mixtures using overcomplete representations [Lee et al., 1999*b*]. Cichocki, 2003, developed a blind source separation (BSS) algorithm with matrix constraints, i.e. with prior information about the mixing matrices. They hypothesize that the mixing or separating matrices have some special structure or some constraints are imposed for the matrices such as symmetries, orthogonality, nonnegativity, sparseness and specified invariant norm of the separating matrix. Furthermore, Calhoun et al., 2005, propose a semi-blind ICA of fMRI data by incorporating the experimental paradigm information into the spatial ICA.

The algorithms described above showed that some algorithms require selecting functions $g_i(.)$ according to the hypothetical (but unknown) probability density function of the sources to be estimated. Some methods are based on fourth order cross-cumulates in order to measure independency which leads to approximations of the information minimization. Other methods use parametric density estimation that alternates with a cost function optimization step in an iterative approximation framework. And some algorithms are based on a nonparametric density estimation of the signals. The performances of the different algorithms are tested in simulation studies in Chapter 6.

5.4 Performance of ICA Algorithms

The finite sample size induces statistical errors in the estimation of parameters. Moreover, as with any statistical method, it is necessary to analyze the performance of the estimated components. To evaluate the performance of the estimates of the different algorithms, an error index (EI) [Amari et al., 1996] can be used. The error index is computed using the matrix $\mathbf{D} = [d_{j_1,j_2}]_{j_1,j_2=1,\ldots,M}$ ($M \leq N$), where $\mathbf{D} = \mathbf{WA}$. It should be mentioned again, that $\mathbf{A}$ is the unknown mixing matrix and only known in case of simulations and $\mathbf{W}$ is the

estimated unmixing matrix. The error index EI is defined by

$$EI = \sum_{j_1=1}^{M} \left(\sum_{j_2=1}^{M} \frac{|d_{j_1 j_2}|}{\max\limits_k |d_{j_1 k}|} - 1 \right) + \sum_{j_2=1}^{M} \left(\sum_{j_1=1}^{M} \frac{|d_{j_1 j_2}|}{\max\limits_k |d_{k j_2}|} - 1 \right). \tag{5.59}$$

In the ideal case $\mathbf{D} \approx \mathbf{I}$, this also includes permutations of the matrix $\mathbf{D}$, see (5.7). And consequently $EI \approx 0$. In words, the smaller the error index, the better the estimation of independent components.

If the matrix $\mathbf{A}$ is not known, as in many applications like fMRI studies, other methods for testing the performance of the estimates are introduced. One of these methods might be bootstrapping.

Meinecke et al., 2002 propose classical analysis of statistical reliability as the bootstrapping to assess quality of the estimates. Consider a random variable X and regard x as a realization of it, $x = (x(1), \ldots, x(T))$. It is aimed to estimate a set of parameters $\theta = (\theta_1, \ldots, \theta_M)$ from the observed data. The estimated parameters are denoted by $\hat{\theta} = \hat{\theta}(x) = (\hat{\theta}_1(x), \ldots, \hat{\theta}_M(x))$ (i.e. the estimates of the unmixing matrix $\mathbf{W}$), where the estimator is a function of the given data set. An important quantity to assess stability of an estimate is the root-mean-squared error (RMSE) of the estimates θ_j, $j = 1, \ldots, M$ is defined by:

$$RMSE_j = \sqrt{E\{(\theta_j - \hat{\theta}_j(X))^2\}}. \tag{5.60}$$

Bootstrapping is a resampling method where the data sample x is randomly changed by simulating the sampling process. The algorithm (i.e. ICA algorithm) is run B times with the bootstrapped samples, where $M < N$ signals are chosen randomly [Efron and Tibshirani, 1993]. A scalar parameter θ_j is estimated with an estimator $\hat{\theta}_j(\mathbf{x})$. It is aimed to evaluate the RMSE of the estimator. Then, B new surrogate data sets, i.e. bootstrap samples, $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_T^{*b})$ are generated with $b = 1, \ldots, B$, by taking T *iid* random variables $x_1^{*b}, \ldots, x_T^{*b}$. Note that, some data points might occur several times, while others might not occur at all in a particular bootstrap sample. On each surrogate $\mathbf{x}^{*b}$, the estimator $\hat{\theta}_j^{*b} = \hat{\theta}_i(\mathbf{x}^{*b})$ is calculated, having B estimators $\hat{\theta}_j^{*b}, \ldots, \hat{\theta}_j^{*B}$. The bootstrap estimator of the RMSE is calculated as

$$\hat{RMSE}_j(B) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_j - \hat{\theta}_j^{*b})^2}, \tag{5.61}$$

where $\hat{\theta}_j$ is the estimate of a scalar parameter θ_j of a sample vector $\mathbf{x} = (x(1), \ldots, x(T))$ of B runs, B is the number of resampled data sets. The reliability can then be analyzed by

looking at the spread of the obtained estimates. It was shown that the bootstrap estimators are consistent Efron and Tibshirani, 1993, i.e. $\hat{\sigma}_j(B) \xrightarrow{P} \sigma_j$ (converges in probability) as $B \to \infty$.

In another way, Himberg et al., 2004, propose a method for validating the independent components of neuroimaging time series via clustering and visualization. Many ICA algorithms have stochastic elements, e.g. the gradient descent Infomax algorithm of the fixed-point iterative FastICA algorithm (see 5.3.5). Consequently, the algorithms give somewhat different results depending on the point where the algorithm started the calculation. Their method is based on estimating a large number of candidate independent components by running an ICA algorithm many times. Reliable components are then corresponding to clusters that are small and well separated from the rest of the estimates, contrary to unreliable components which correspond to points which do not belong to any cluster.

For the simulation studies in Chapter 6, we used the error index to analyze the performance of the estimated components.

5.5 ICA Applied to fMRI Data

ICA was originally developed to solve BSS problems like the cocktail-party problem. In cocktail-party problem it is aimed to isolate voices of mixtures of voices and other noises. The problem of isolating the electrical activity of single neurons in population recordings shares a number of similarities with the challenge of isolating voices at a cocktail party, which is therefore called the 'neural cocktail-party problem' [Brown et al., 2001]. ICA has been widely applied to further neuroscience data like EEG, MEG, or fMRI. In fMRI ICA shows good applicability to cognitive paradigms for which detailed *a priori* models of brain activity are not available. The relatively low image signal-to-noise ratio of the BOLD effect, head movements, and undesired physiological sources of variability of the subjects make detection of the small activation-related signal changes difficult. Therefore, ICA is a powerful method for recovering underlying signals, or independent components from linear mixtures in fMRI recordings. Overviews of ICA applied to fMRI data are McKeown et al., 2003, Calhoun et al., 2003. An introduction to ICA is given in Hyvärinen and Oja, 2000 or in Stone, 1999*b* discussing ICA for EEG, fMRI and optical imaging.

Nevertheless, with this work it is aimed to apply ICA to fMRI data as well, but it was of special interest if it is possible to detect learning related dynamic changes in fMRI time series especially over repeated sessions.

The first application of ICA to fMRI data was done by McKeown and Sejnowski, 1998. They used the Infomax principle to investigate task-related human brain activity in fMRI data. By determining the brain regions that contained significant amounts of specific temporally independent components, they were able to specify the spatial distribution of task-related, transiently task-related, and motion-related brain activations.

ICA can be applied to fMRI data in two different ways, i.e. as spatial ICA (sICA) or temporal ICA (tICA). This means that it is possible to either spatially localize [McKeown et al., 1998*b*] or temporally characterize [Biswal and Ulmer, 1999] the sources of BOLD activation. The algorithms introduced in Section 5.3 are referring to sICA. In sICA, statistical independence is assumed for the distribution in space of the extracted sources of signal change. This means the signal sources are independent in their spatial locations rather than in their time profile, which can exhibit high mutual correlations. In other words, sICA finds systematically nonoverlapping, temporally coherent brain regions without constraining the temporal design [Calhoun et al., 2003]. In tICA the sources are assumed to be independent as far as time is concerned. In general sICA and tICA are based on the same algorithms. In sICA it is assumed that the columns of the matrix $\mathbf{S}= [s_{j,t}]_{j=1,...,M,\ t=1,...,T}$ (see Equation (5.2)) are independent processes, whereas in tICA the rows of $\mathbf{S}= [s_{j,t}]_{j=1,...,M,\ t=1,...,T}$ are assumed to be independent. Whether sICA or tICA should be applied to the data is discussed in Calhoun et al., 2001, Stone, 1999*a*, Stone et al., 2002. So far the sICA dominated in the application of ICA to fMRI data sets. TICA was rather applied to EEG or MEG data, which have a high temporal resolution (in the milliseconds domain) [Makeig et al., 1997]. The electrical signals originating from the brain are quire weak at the scalp, in the microvolt range, and there are larger artificial components arising from eye movements and muscles. Makeig et al., 1997 applied the Infomax algorithm to EEG data showing that the algorithm can extract EEG activations and isolate artifacts. Jung et al., 2000 show that the extended Infomax algorithm is able to linearly decompose EEG artifacts such as line noise, eye blinks, and cardiac noise into independent components with sub- and supergaussian distributions. In Seifritz et al., 2002, they used an initial sICA to reduce the spatial dimensionality of the data by locating a region of interest in which they performed a tICA to study the structure of the nontrivial temporal response in the human auditory cortex in more detail.

For the ICA decomposition of out data in Chapter 7 we used a sICA to get activation clusters in the brain with associated time courses. The time courses were analyzed afterwards regarding dynamic changes in the signals.

Considering a typical 3D fMRI dataset where the number of time points T is much smaller than the number of voxels N, the spatial and temporal dimensions of the statistical samples suggest the use of the sICA for 3D-pattern generation, whereas tICA can reveal the presence of multiple dynamics in an anatomically or functionally selected region of interest (ROI) [Calhoun et al., 2001, Seifritz et al., 2002]. From the perspective of statistical power, sICA has the best potential for a robust representation of whole-brain fMRI data sets because of the sample size achievable. The statistical power of sICA can be as high as to enable useful sICA decomposition even using few points of a single slice fMRI time series [Esposito et al., 2003].

In most fMRI studies the cerebral cortex is the main target of analysis. Since only about 20 % of the voxels of a typical fMRI data set lie within the cortex, Formi sano et al., 2004 propose cortex-based ICA of fMRI time series. Through segmentation and reconstruction of the cortical surface of the brain a mask of the brain is created. Therefore, the spatial ICA decomposition is restricted to this subset of voxels in the mask, which results in noise reduction as well as a large dimension reduction in advance without loss of information. This cortex-based approach improves the separation of the independent components representing cortical activation because "uninformative" signals from, for example, the ventricles or near the eyes are excluded. Their inclusion in the data matrix leads to an increase of the complexity of the mixtures in terms of number of sources but does not improve the estimation of the cortical sources. Moreover, this reduction does not affect the maximal number of spatial components, since the maximal number of components is not affected by the number of voxels, but the number of components equals the number of time points, i.e functional images or scans.

After estimation of the independent components they should be projected back onto the original data set to illustrate the component maps. The projection of the j-th independent component onto the original data set is given by multiplying the j-th row of the estimated independent component matrix $\hat{\mathbf{S}}$ with the j-th column of the inverse unmixing matrix $\mathbf{W}^{-1}$. Consequently, brain activities of interest accounted by single or by multiple components can be obtained by projecting selected independent components back onto the activation map $\mathbf{X}_0 = \mathbf{W}^{-1}\hat{\mathbf{S}}_0$, where $\hat{\mathbf{S}}_0$ is the matrix of $\hat{\mathbf{S}}$ of activations with rows of irrelevant activation set to zero [Jung et al., 2001].

Thereby, each independent component map is described by a distribution of values for each voxel. These values represent the relative amount that a given voxel is modulated by the activation of that component [McKeown et al., 1998*b*]. There is one such a value for all

time points, because that value represent the average signal change between stimulation and resting condition for instance.

To find irrelevant component maps, the voxels (i.e. map values) are often scaled to z-scores. The z-scores are obtained by merely scaling the components to have zero mean and unit variance. So the z-scores are the number of standard deviations from the map mean, this is computed from the row of the estimated independent component matrix $\hat{\mathbf{S}}$. This is probably not strictly valid, as (by definition for the ICA) the components will not have a gaussian distribution. But once again, the z-score computed for each individual component merely represents how far the voxel intensities differ from the mean voxel intensity. To minimize the probability of false positives, only voxels with an absolute z-score ≥ 2 were considered as active voxels for that component [McKeown et al., 1998*b*]. In this case, the z-scores are used for descriptive purposes and have no definite statistical interpretation. Negative z-scores indicate voxels whose fMRI signals are modulated 'opposite' to the time course of activation for that component.

Estimating independent components reveals time courses and activation maps. One activation map may consist of several clusters which are temporally connected. For the evaluation of the connectivity, the mean inter-voxel correlation coefficient (MCC) can be calculated on a voxel by voxel basis,

$$MCC = \frac{2}{N_{map}(N_{map}-1)} \sum_{k=1}^{N_{map}-1} \sum_{l=k+1}^{N_{map}} cc_{kl}, \tag{5.62}$$

where cc_{kl} is the correlation coefficient between the k-th and the l-th time course of the detected component map voxels, N_{map} ($N_{map} \leq N$) refers to the number of detected voxels in the component map [Kiviniemi et al., 2003].

In many cases especially in fMRI data, the number of independent components might be very large, in spatial ICA this can be up to the number of time samples. The problem is that the estimated source signals are determined without any ordering. There are different proposals to select a "meaningful" subset from a large set of components. McKeown et al., 1998*b*, propose to order the independent components according to the contribution each component makes to the magnitude of the original data. The contribution δ_j ($j = 1, \ldots, M$), each component makes to the magnitude of the original data can be estimated by the RMSE of the data set reconstructed solely from this component,

$$\delta_j = \frac{1}{TM} \left(\sum_{t=1}^{T} \sum_{j=1}^{M} \mathbf{d}_j \right)^{\frac{1}{2}}, \tag{5.63}$$

where $\mathbf{d}_j$ is the j-th column of an N by M matrix $\mathbf{D}$ computed from the outer product of the two vectors: The j-th estimated component map and the j-th column of $\mathbf{W}^{-1}$, i.e. $d_{jt} = w_{ji}^{-1}\hat{s}_{jt}$. The outer product of two vectors is a vector again. Nevertheless, this contribution might be sometimes quite uninformative, another way might be to select the components on the basis of their time courses correlation with the stimulation protocol, if known [McKeown, 2000]. But, in this case components which are not necessarily correlated to the stimulation protocol are ignored. This may result that possibly interesting components, e.g. components reflecting learning related-process, not correlating to the stimulation protocol, are ignored.

In a work of Formisano et al., 2002 three measures are proposed to select the meaningful components. These are the kurtosis of the components' distribution of voxel values, the degree of spatial clustering of the voxels and the one-lag serial autocorrelation of its time course. Another way of ordering the components is proposed as topographical [Hyvärinen et al., 2001*a*] and frequency-based ordering [Moritz et al., 2003] of the components.

In Chapter 6 we used the proposed ICA algorithms from Section 5.3 for simulation studies. The quality of the ICA estimated was validated by error indices (Equation 5.59). In Chapter 7 we present an fMRI study. For decomposing different signals of fMRI data sets a cortex-based FastICA algorithm implemented in *BrainVoyager* was used. This FastICA algorithm dominated in the literature and additionally dominated in our simulation studies. Meaningful components were selected by crosscorrelating the time courses of independent components to the stimulation protocol and selecting meaningful regions in the brain.

6 Simulation Studies

This chapter describes the results of performed simulation studies. For this purpose the signal of the hypothetical hemodynamic response function (HRF) (see Figure 3.1) was modelled. Additionally, other signals that might contribute to an fMRI measurement are modelled like random noise, linear or exponential trend functions and sinus functions. These source signals are mixed linearly. The aim of these simulation studies was to test whether the source signals can be estimated on the data basis of the mixed signals using different ICA algorithms which were introduced in Section 5.3. The estimates of the different algorithms are presented. Moreover, time series analyses for the source signals as well as the mixed signals are performed to describe the signals according to their time series characteristics, i.e. stationarity and gaussianity characteristics, their autocorrelation and frequency domain.

In many fMRI studies, the task performance of the subjects is of special interest. It is assumed that the task performance is reflected in the neuronal response, i.e. the signal of HRF changes. Therefore, the HRF was varied in different parameters like the signal amplitude, an amplitude increase or decrease within or between stimulation blocks as well as temporal shifts between two HRFs. These varied HRF signals were once again mixed linearly with other contributing signals. The decomposition into the source signals was performed to draw conclusions at which parameters the mixed signals can still be separated into the source signals.

Another important point is the fact, that ICA is a dimension reduction method but the determination of the number of independent components is mostly subjective. Therefore, simulation studies are performed to determine the number of independent components objectively. The estimates of ICA were also investigated under the aspect of over- or underestimating the number of independent components.

Finally the results of ICA decomposition are compared with the classical method for analyzing fMRI data, the general linear model (GLM) (see Section 3.3).

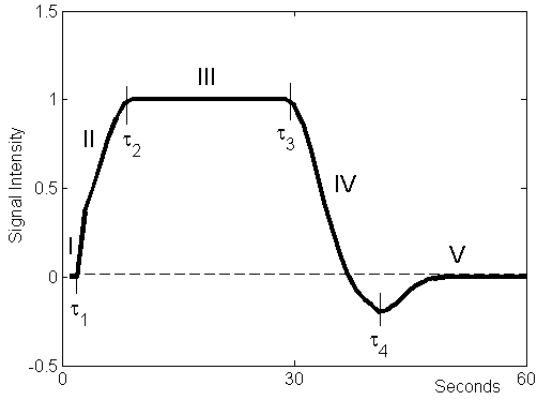

Figure 6.1: Hemodynamic response model

6.1 Modelling the Signals

6.1.1 Modelling the Hemodynamic Response Function

Using a block-design experiment, the time course of the HRF was modelled according to GÖSSL et al., 2001. The signal increase and decrease are described by two independent functions to achieve a greater flexibility. Truncated gaussian functions are used to describe the signals' increase and decrease, parameterized by β_1 and β_2. They further introduced a lag parameter β_3 determining the start of the signal increase. The time during which the signal remains elevated is modelled by a parameter β_4 (fixed according to the stimulus time course). A final parameter β_5 is introduced to account for the poststimulus undershoot modelling the amplitude of the signal drop below the baseline. The slow decay to the baseline of this undershoot is achieved by a third gaussian function with fixed parameters. Parameters for the time delays can be taken from JEZZARD, 2001. The response function for one stimulation block can consequently be divided into five intervals which are displayed in Figure 6.1:

I.	Baseline	for $\tau < \tau_1$
II.	Signal increase	for $\tau \in [\tau_1, \tau_2[$
III.	Plateau	for $\tau \in [\tau_2, \tau_3[$
IV.	Signal decrease	for $\tau \in [\tau_3, \tau_4[$
V.	Maximum undershoot until return to baseline	for $\tau \geq \tau_4$.

One phase in the fMRI experiment is composed of a period of stimulation followed by a period of rest. Thus, the function $x_i(\tau)$, where $\tau \in T$ is the time parameter for one phase, models the hemodynamic response for a voxel $\mathbf{v}^i, i = 1, \ldots, N$, for one phase as follows [Gössl et al., 2001]:

$$x_i(\tau) = \begin{cases} 0 & \tau < \tau_1 \quad (I) \\ \exp\left(-\left(\frac{\tau-\tau_2}{\beta_1}\right)^2\right) & \tau \in [\tau_1, \tau_2[\quad (II) \\ 1 & \tau \in [\tau_2, \tau_3[\quad (III) \\ (1+\beta_5)\exp\left(-\left(\frac{\tau-\tau_3}{\beta_2}\right)^2\right) - \beta_5 & \tau \in [\tau_3, \tau_4[\quad (IV) \\ -\beta_5 \exp\left(-\left(\frac{\tau-\tau_4}{4}\right)^2\right) & \tau \geq \tau_4 \quad (V). \end{cases} \tag{6.1}$$

The time points are related to the model parameters β as follows:

$$\begin{aligned} \tau_1 &= \beta_3 \\ \tau_2 &= \beta_1 + \beta_3 \\ \tau_3 &= \beta_1 + \beta_3 + \beta_4 \\ \tau_4 &= \beta_1 + \beta_2 + \beta_3 + \beta_4, \end{aligned} \tag{6.2}$$

such that the β's depend on τ as follows: $\beta_1 = \tau_2 - \tau_1$, $\beta_2 = \tau_4 - \tau_3$, $\beta_3 = \tau_1$, and $\beta_4 = \tau_3 - \tau_2$, and β_5 can be freely chosen. Figure 6.1 displays the hemodynamic response model for a stimulation block with 30 seconds and a resting block with 30 seconds too. Consequently the following parameters were chosen: $\tau_1 = 2$, $\tau_2 = 8$, $\tau_3 = 30$, $\tau_4 = 40$ given in seconds, and $\beta_5 = 0.25$. Whereby the transition of phase (I) $\rightarrow$ (II) and phase (IV) $\rightarrow$ (V) are adapted to give a smooth curve. These time parameters τ's and the parameters β's were so far chosen for all the following simulations. Further parameters, that were varied are introduced in the following.

A block-design experiment consists of phases that are repeated several times. The total length of an fMRI experiment can be defined by different parameters. First of all the duration of the stimulation block given in seconds is defined by κ_s ($\kappa_s > 0, \kappa_s \in \mathbf{N}$). A second parameter κ_r ($\kappa_r > 0, \kappa_r \in \mathbf{N}$) is the duration of the resting block where the signal returns to baseline. In the following it is assumed that the duration of stimulation equals the duration of resting ($\kappa_s = \kappa_r$). Another parameter is the number of phases given by κ_p ($\kappa_p > 0, \kappa_p \in \mathbf{N}$). And finally the number of fMRI images recorded in a phase is defined by κ_b ($\kappa_b > 0, \kappa_b \in \mathbf{N}$). Based on these given parameters, the total length of the fMRI

experiment given in images κ_B ($\kappa_B > 0, \kappa_B \in \mathbf{N}$) is determined by

$$\kappa_B = \kappa_p \cdot \kappa_b. \tag{6.3}$$

The length of one image κ_{TR}, i.e. the repetition time (TR), is computed by

$$\kappa_{TR} = (\kappa_s + \kappa_r)/\kappa_b, \tag{6.4}$$

$$\kappa_{TR} = (2 \cdot \kappa_s)/\kappa_b \quad \text{for } \kappa_s = \kappa_r. \tag{6.5}$$

The total length given in seconds κ_T ($\kappa_T > 0, \kappa_T \in \mathbf{N}$) is determined by

$$\kappa_T = \kappa_B \cdot \kappa_{TR}. \tag{6.6}$$

Figure 6.2 shows the hypothetical HRF with the following parameters: β's and τ's are chosen as in Figure 6.1, $\kappa_s = \kappa_r = 30s$, $\kappa_p = 15$, and $\kappa_b = 20$. These parameters where chosen because they correspond to the parameters used in the auditory fMRI experiment (see Chapter 7).

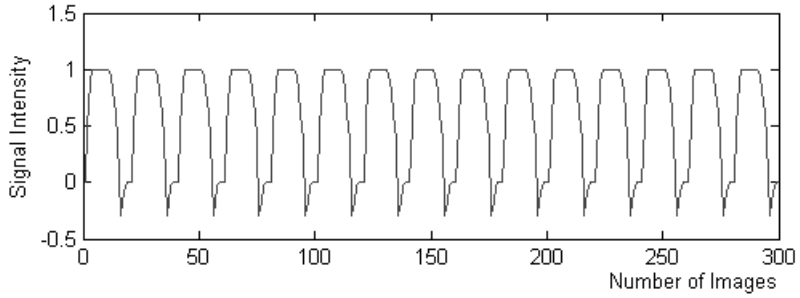

Figure 6.2: Time course of the hypothetical hemodynamic response function.

6.1.2 Variations in the Hemodynamic Response Function

Since it is assumed that the task performance of the subjects affects the neuronal response, i.e. the HRF, different models of changes in the HRF are introduced.

A first parameter is the signal amplitude κ_a ($\kappa_a \in \mathbf{R}$). This parameter describes the percentage change between stimulation and resting condition. Note that it is assumed that the resting condition itself contains some basic noise, even when the subject is asked to do and think nothing, some noise is measurable. This noise is set to 100%, and the percentage change between stimulation and resting condition ist computed in relation to this 100%.

The value for κ_a can be found in Equation (6.1) for $\tau \in [\tau_2, \tau_3[$. One the one side, the signal amplitude depends on the magnetic field strength of the MR tomograph (usually 2-3 % at 1,5 Tesla, 4-6 % at 3 Tesla and 8-12 % at 7 Tesla). One the other side, the signal amplitude also depends on the presented stimuli and the task. Moreover, the signal amplitude may depend on the region in the human cortex so that two different regions A and B in the cortex may show different signal changes, 2 % in region A and 4% in region B for instance.

It is also possible that an fMRI experiment consists of several stimulation conditions that might show different signal amplitudes. Figure 6.3 shows an fMRI design with two alternating stimulation conditions assuming that condition I has a signal amplitude of $\kappa_{a1} = 1$ and condition II has a signal amplitude of $\kappa_{a2} = 0.3$.

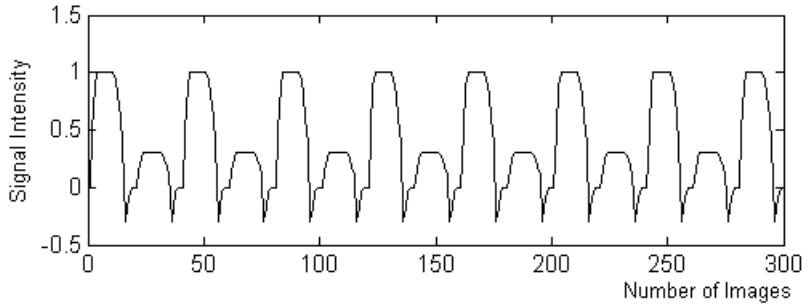

Figure 6.3: Time course of the hypothetical hemodynamic response function with two alternating signal amplitudes.

It might also be possible that the signal amplitude varies between different stimulation blocks, i.e. the signal amplitude increases or decreases by time. This signal increase or decrease may be described by a parameter κ_m ($\kappa_m \in \mathbf{R}$). Therefore, the time series $x(\tau)$ of Equation (6.1) was multiplied with a increasing linear function defined by an beginning and ending amplitude. Figure 6.4 shows the HRF with signal increase from 2% to 4% at the end.

The signal amplitude increase or decrease might also be possible within on stimulation block which is demonstrated in Figure 6.5 where the signal is not sustained over the stimulation block anymore. The signal increase or decrease might be described by the parameter κ_n ($\kappa_n \in \mathbf{R}$). Therefore, Equation (6.1) is modified for $\tau = [\tau_2, \tau_4[$ as $x(\tau) = 1 - n\tau$, where n is a parameter for signal increase or decrease. In Figure 6.5 κ_n is chosen to be 1/25.

Considering that the changes in the HRF my be dynamic with the time it might be imag-

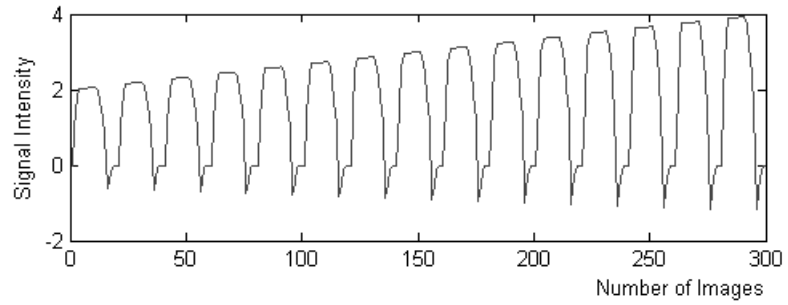

Figure 6.4: Time course of the hypothetical hemodynamic response function with signal amplitude increase over the session.

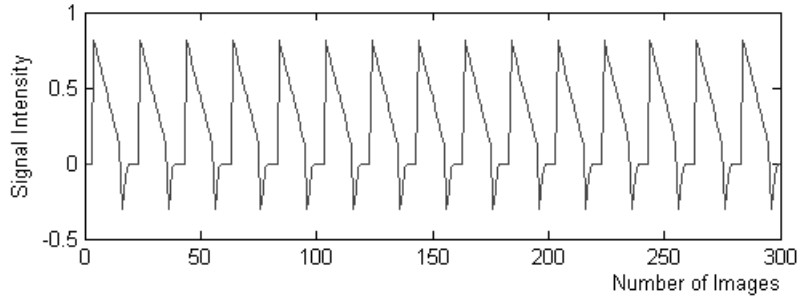

Figure 6.5: Time course of the hypothetical hemodynamic response function with signal amplitude decrease within one stimulation block.

inable that the parameter κ_n is not static over the duration of the experiment but rather dynamic over the time as displayed in Figure 6.6. The parameter κ_n for the last stimulation block was chosen to be 1/50.

Regarding the dynamic of changes in the HRF, a further case should be considered, namely that the HRF signal is affected by an additional increasing noise as displayed in Figure 6.7. Thereby the HRF signal was superimposed by an additional noise, increasing over time. Such signal were observed in real fMRI measurements. This might be interpreted, that the attention of the subjects changes from beginning to the end of the session.

A final parameter modifying the HRF might be introduced as κ_c ($\kappa_c \in \mathbf{N}$) given in number of images, which defines the temporal shift between two HRFs as illustrated in Figure 6.8, where the second HRF (dashed line) is shifted by 3 images compared with the first HRF (solid line). Such appearances might happen at functionally connected regions where the information is transmitted time-delayed to another region, e.g. from auditory regions to

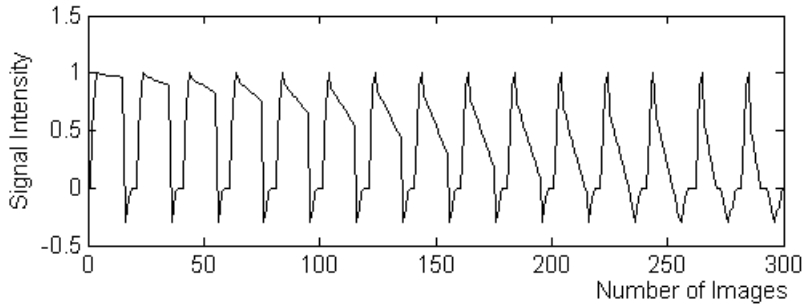

Figure 6.6: Time course of the hypothetical hemodynamic response function with a dynamic signal amplitude decrease.

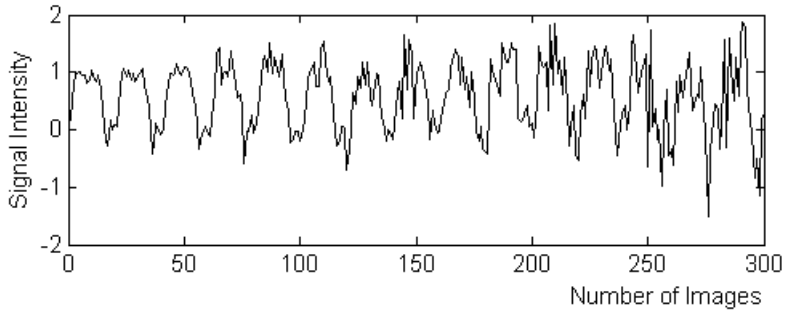

Figure 6.7: Time course of the hypothetical hemodynamic response function with increasing noise.

frontal regions.

6.1.3 Further Contributing Signals

It is supposed that the measured time series of each voxel consisted of a linear mixture of different signals. Besides the HRF signals there are linear or exponential trends maybe representing motion artifacts. Moreover, sinus functions are modelled mimicking the neuronal processing of cardiac or respiratory pulsations, and background noise is modelled, which can be modelled as a gaussian white noise process. These contributing signals are computed in the following way. The noise is modelled as

$$s_{noise}(t) = NV(0, \sigma), \tag{6.7}$$

where $NV(0, \sigma)$ is a realization of a gaussian distributed random variable with mean zero and variance σ with independent realizations for different t. The parameter σ ($\sigma \in \mathbf{R}$) is

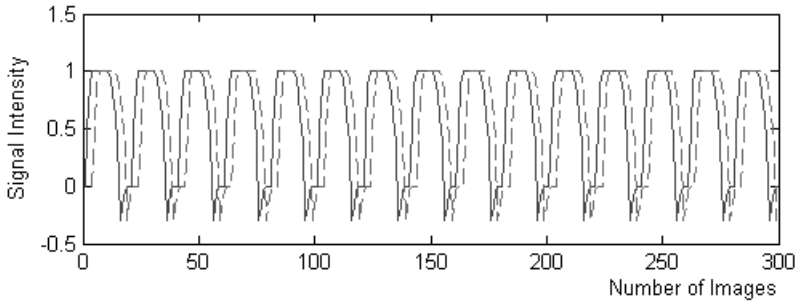

Figure 6.8: Time courses of two shifted hypothetical hemodynamic response functions.

the noise parameter that might be varied.

The linear trend is given by

$$s_{linear}(t) = m \cdot t. \tag{6.8}$$

In this case m ($m \in \mathbf{R}$) is the parameter of signal increase or decrease during the experiment. An exponential trend might be given by

$$s_{exp}(t) = 1 - e^{-a \cdot t}, \tag{6.9}$$

where $a \in \mathbf{R}$.

A sinus function is modelled as

$$s_{sin}(t) = \sin(P \cdot t), \tag{6.10}$$

where P ($P \in \mathbf{R}$) models the period of the sinus function.

In simulation studies variations of these function were used to be mixed with each other 500 times with different mixing matrices $\mathbf{A}$. Thereby also for every simulation the noise signal was modelled randomly. Consequently, for every simulation a new mixing matrix $\mathbf{A}$ with the contribution of the several signals and a noise signal were modelled randomly. Then after estimating the independent components, these estimates were validated by error indices.

6.2 Performing the Programming

For comparing ICA algorithms by simulation studies program codes were written and a graphical user interface (GUI) in *MATLAB* (The MathWorks Inc., Massachusetts, USA)

was designed (see Figure 6.9). The user can first create the hypothetical HRF by defining the length of the stimulation block κ_s and the length of the resting block κ_r, both given in seconds. Additionally, the user can constitute the number of phases κ_p, where one phase is composed of a stimulation block and a following resting block, and the number of images κ_b recorded in a phase is also given. Consequently, the total length of the fMRI signal κ_B follows from the number of phases multiplied by the number of images in a phase. In a second step the contributing signals can be chosen. Some of these signals might be white noise with a noise parameter σ mimicking the neuronal processing of the scanner noise or artifacts, a linear trend with a signal increase or decrease parameter, or an exponential trend. Additionally, different sinus signals can be chosen mimicking the cardiac or respiratory rates of the subjects.

These source signals are then linearly mixed by a mixing matrix $\mathbf{A}$ chosen to have either uniformly distributed or gaussian distributed elements, where the number of mixed signals, i.e. the number of observed signals, is given.

Many ICA algorithms are offered as MATLAB-files for downloading from the corresponding web pages. The following ICA algorithms from Section 5.3 were applied to the simulated data

- Bell and Sejnowski Infomax,
  http://hendrix.imm.dtu.dk/software/lyngby
- the Hyvärinen FastICA,
  http://www.cis.hut.fi/projects/ica/fastica
- Maximum Likelihood ICA,
  http://mole.imm.dtu.dk/toolbox/ica/index.html
- and a nonparametric ICA by Boscolo,
  http://www.ee.ucla.edu/∼riccardo/ICA/npica.html
- the Molgedey and Schuster approach
  http://hendrix.imm.dtu.dk/software/lyngby.

Additionally a PCA is created to be compared with the ICA algorithms. The Jutten-Hérault algorithm, which has not much practical relevance in fMRI data, was omitted, because it was only applied to speech signals and a *MATLAB* code was not available.

After estimating the independent components, these signals were resorted by maximum

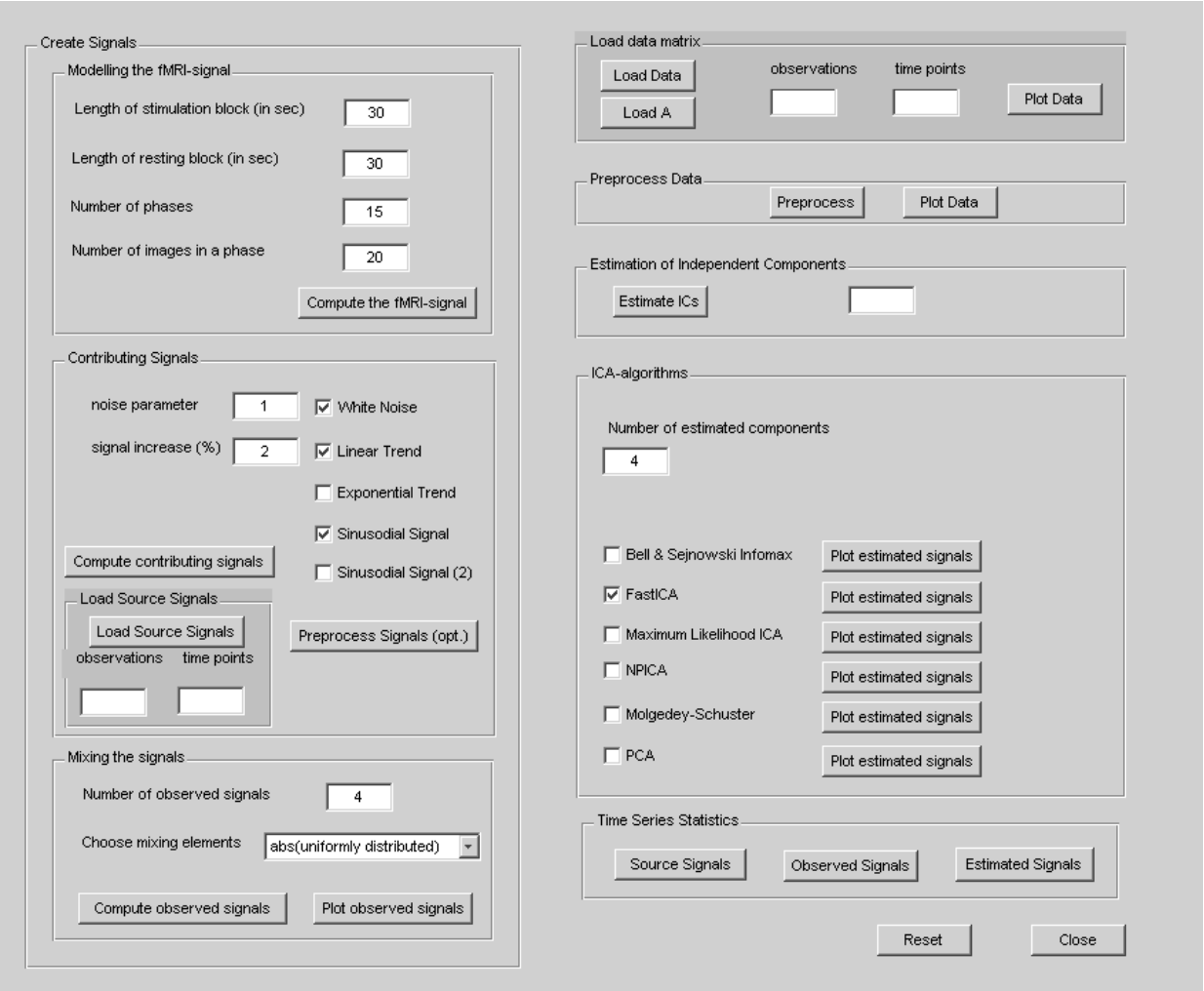

Figure 6.9: MATLAB graphical user interface for simulation studies

correlation according to the source signals. The mapping of estimated components to source signals was unique. Of course the better the estimation was, the better was the correlation to the source signals. Algorithms that could not estimate the independent components that good showed difficulties in correlating the independent components to the source signals and the correlation coefficient were much smaller. If it was necessary, the sign of the signals was changed.

6.3 Illustrative Results of ICA Decomposition

In this section an ICA decomposition of exemplary mixed signals is presented. Therefore all time parameters of the HRF signal and parameters of the contributing signals like σ, m, and P are fixed. Variations of these parameters of the signals will be performed later.

The source signals are the HRF (see figure 6.2), a realization of a white noise process with $\sigma = 0.5$ for every simulation, a linear trend with $m = -2$, and a sinus function with $P = 7$. The source signals are displayed in Figure 6.10 (left side). Theses signals are mixed linearly by a mixing matrix $\mathbf{A}$. The exemplary matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} 0.7036 & 0.3654 & 0.6739 & 0.3603 \\ 0.4850 & 0.1400 & 0.9994 & 0.5485 \\ 0.1146 & 0.5668 & 0.9616 & 0.2618 \\ 0.6649 & 0.8230 & 0.0589 & 0.5973 \end{pmatrix}. \tag{6.11}$$

Figure 6.10 (right side) shows the realization of these exemplary mixed observed signals.

The ICA decomposition was performed with the algorithms mentioned above. The Bell & Sejnowski Infomax algorithm was not able to detect the source signals from the mixed signals (Figure 6.11). The estimated signals are still mixed with each other. The Hyvärinen FastICA algorithm did a good performance in decomposing the mixed signal into the source signals (Figure 6.12). The Maximum Likelihood estimation was also not able to perform the decomposition (Figure 6.13). The nonparametric ICA decomposition estimated the

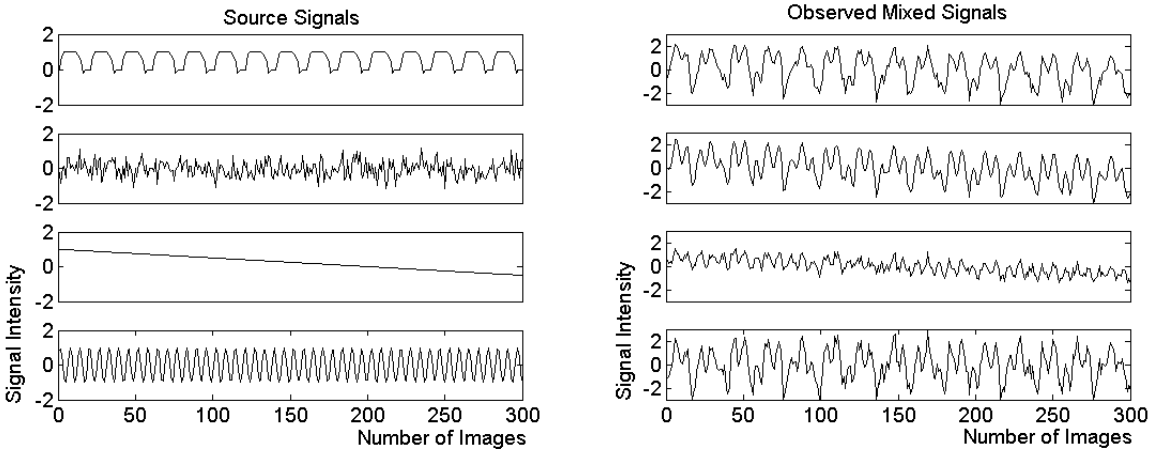

Figure 6.10: Time courses of four source signals and four mixed signals.

source signals quite well (Figure 6.14). The MOLGEDEY and SCHUSTER algorithm did also a good decomposition of the source signals (Figure 6.15). And finally PCA cannot estimate the source signals (Figure 6.16).

As could be seen in Figures 6.11 - 6.15 some algorithms are able to detect the source signals, and others not. But how good are these estimates? In the illustrative result the error indices EI (see Equation 5.59) are displayed in Table 6.1.

Table 6.1: Error indices for ICA estimates (see Figures 6.11 - 6.15)

ICA algorithm	**error index** EI
BELL and SEJNOWSKI	7.7921
FastICA	0.1943
Maximum Likelihood ICA	5.9076
Nonparametric ICA	0.0882
MOLGEDEY and SCHUSTER	0.5297

As seen from the indices, the algorithms, which, by the optical impression of Figures 6.11 - 6.15 were able to detect the source signals, had small error indices. Algorithms, which could not detect the source signals had large error indices. This illustrative example shows that the ICA decomposition worked well for some algorithms. This should be further investigated for several simulation runs.

Figure 6.17 shows the median, 0.25-quantile and 0.75-quantile of error indices for 1000 simulations. For each simulation run the source signals from the example above were taken, whereas a new noise signals was modelled for every simulation run, and the elements of the mixing matrix $\mathbf{A}$ were randomly created as $[0, 1]$-distributed random variables for every simulation run as well . The FastICA algorithm showed very good results the same like the nonparametric ICA. The Infomax ICA and maximum likelihood ICA showed worse results compared with the other algorithms.

In the literature, ESPOSITO et al., 2002, compared the results of Infomax and FastICA algorithm. They performed ICA of fMRI time series and simulated signals. As result, the FastICA algorithm outperformed the Infomax in terms of spatial and temporal accuracy. Spatial accuracy was assessed by receiver operating characteristics (ROC), where it was possible to separate with infinite precision false and true activation areas in simulated

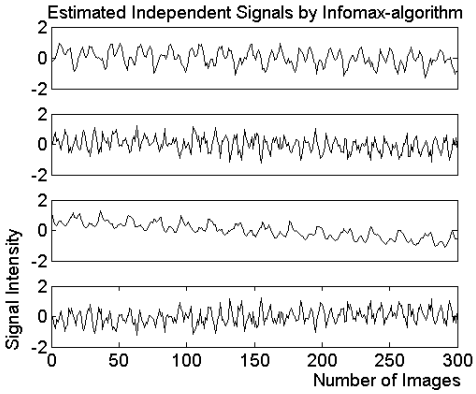

Figure 6.11: Bell and Sejnowski Infomax algorithm

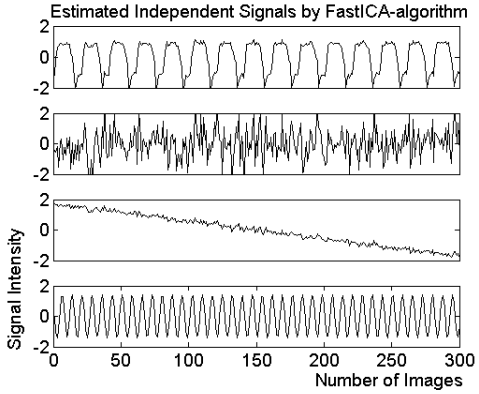

Figure 6.12: Hyvärinen FastICA algorithm

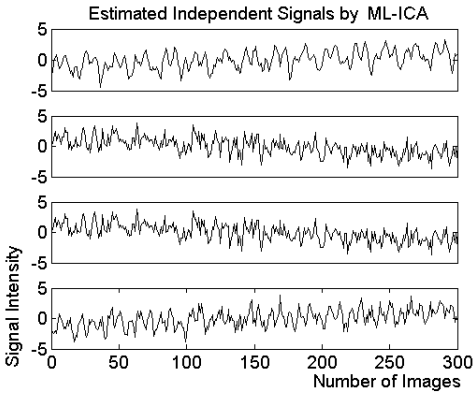

Figure 6.13: Maximum Likelihood estimation

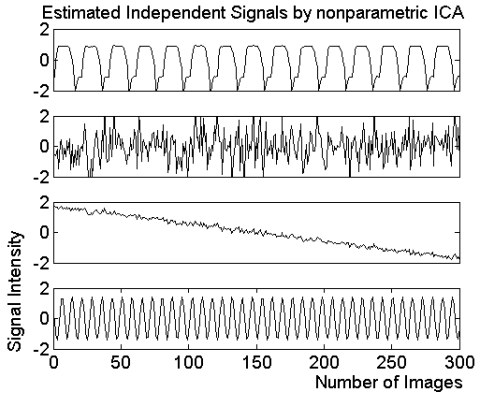

Figure 6.14: Nonparametric ICA estimation

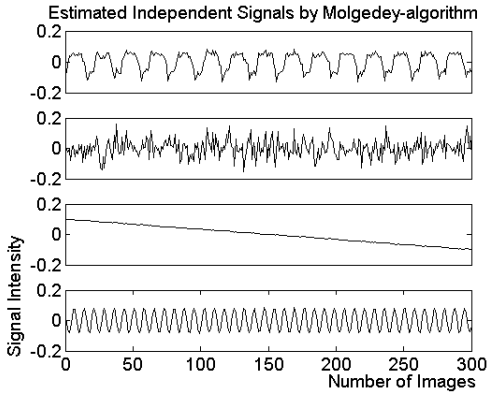

Figure 6.15: Molgedey and Schuster algorithm

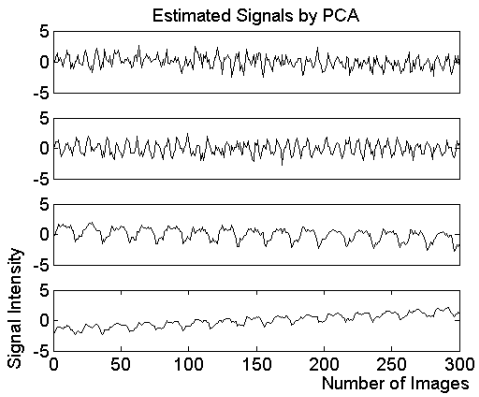

Figure 6.16: Principal component analysis

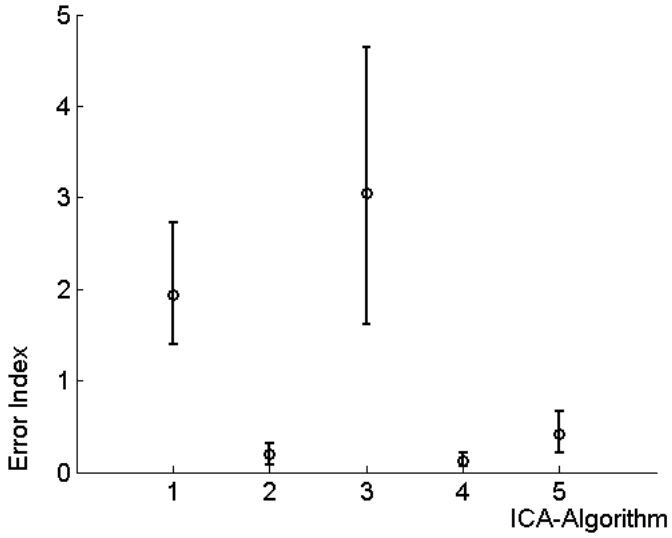

Figure 6.17: Error indices for 1000 simulations (ICA algorithms 1: Infomax ICA, 2: FastICA, 3: Maximum Likelihood ICA, 4: nonparametric ICA, 5: Molgedey and Schuster ICA)

data. Thereby for ICA estimates, an ROC curve has been fitted after the determination of the corresponding false positive fraction and the false negative fraction at varying thresholds for the selected ICA z-map. Temporal accuracy was assessed by correlation analysis. Conversely, the Infomax sICA was superior in terms of reducing the noise in the components. Additionally, the Infomax generated more structured components in terms of degree of clustering, i.e. more connecting voxels.

In the following the parameters of the signals are systematically varied to test the quality of the ICA algorithm.

6.4 Simulation Studies with Variations in the HRF

In this section, simulation studies with varied HRFs (see Section 6.1.2) will be presented. Therefore, different parameters like signal amplitude κ_a, trends between stimulation blocks κ_m, or the noise parameter σ of the white noise process are varied. For each parameter dimension the ICA algorithms run 500 times, by varying the mixing matrix $\mathbf{A}$ and the noise signal for every simulation run. The error index EI (Equation 5.59) was computed for every simulation.

It should be considered that this are only examples with a small number of observed signals and independent components, ranging from 2 to 8 signals, that would not represent a whole fMRI dataset. Nevertheless, with these examples, the performance of the ICA algorithms by varying different parameters of the source signals will be tested. Thereby, it is also investigated what changes of the estimates if the independence of the source signals is violated.

The presentation of results is restricted to the FastICA algorithm (see Section 5.3.5), since it was seen that this algorithm outperforms the other algorithms, see Figure 6.17. Although, the nonparametric ICA showed good results, the FastICA has the additional advantage that it can estimate less independent signals than given mixtures, which is not possible with nonparametric ICA, see Section 5.3.7. Moreover the FastICA algorithm is an algorithm dominating in literature and practical use for different kinds of data, and especially for fMRI data. This algorithm was also used for the analysis of fMRI data sets in Chapter 7. These following simulation studies had also been performed with the other ICA algorithm (see Section 5.3) as presented in Figure 6.17, but the FastICA algorithm once again outperformed the other algorithms, so the presentation of results and figures is restricted to this algorithm, even sometimes it is compared to other algorithm, like in Section 6.5 where the results are compared to Molgedey and Schuster algorithm.

Varying the Signal Amplitude and the Noise

First of all consider the simplest case where two signals are given: an HRF signal an a noise signal. The amplitude κ_a of the HRF signal and the noise parameter σ of the noise signal were varied in the set: $\kappa_a = [0.5, 1, 3, 5]$ and $\sigma = [0.5, 1, 3, 5]$ ($\kappa_s = 30, \kappa_p = 15, \kappa_b = 20$). The two signals were mixed linearly and decomposed into two signals and the error index EI of the estimates was determined. Figure 6.18 displays the median, 0.25-quantile and 0.75-quantile of error indices of 500 simulations for the FastICA algorithm. In each subplot the results for $\kappa_a = [0.5, 1, 3, 5]$ are displayed and each subplot shows the results for a defined σ. It is obviously from Figure 6.18 that for small signal amplitudes and large noise parameters the estimation of independent components is declined. A smaller noise parameter does not influence the estimation. Consequently, the goodness of estimation strongly depends on the ratio of κ_a to σ. It should be mentioned at this point that in the simulation studies all signals, in this particular case the HRF signal and noise signal, are equivalent signals having the same chance to contribute to the mixture and consequently to the error index.

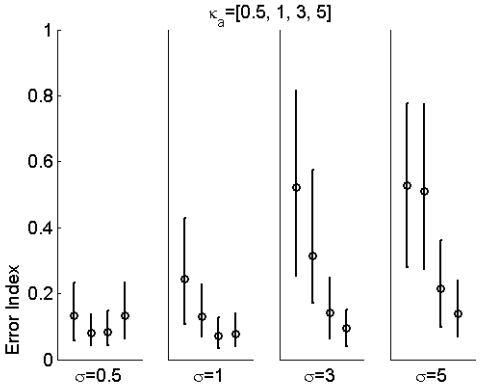

Figure 6.18: Error Indices of 500 simulations by varying the signal amplitude κ_a and the noise σ, displaying median, 0.25-quantile and 0.75-quantile. Each subplot displays for a specified value of σ the results for $\kappa_a = [0.5, 1, 3, 5]$ (results of FastICA algorithm).

In a further simulation four source signals were mixed linearly. An HRF signal with two different alternating signal amplitudes, mimicking different stimulation conditions (see Figure 6.3) were created. The parameters $\kappa_{a1} = 1$ was chosen and κ_{a2} was varied as $\kappa_{a2} = [0.3, 0.7, 2, 3, 5]$ ($\kappa_s = 30, \kappa_p = 15, \kappa_b = 20$). Furthermore, a noise signal was modelled with varied noise parameter $\sigma = [0.5, 1, 3, 5]$, a sinus function with period $P = 7$ and a trend function with $m = -2$. Figure 6.19 displays the median, 0.25-quantile and 0.75-quantile of error indices of 500 simulations for the Fast ICA algorithm. Each subplot displays the results for $\kappa_{a1} = 1$ and $\kappa_{a2} = [0.3, 0.7, 2, 3, 5]$ for a given noise parameter σ. It can be seen that the results are not affected by the variation of signal amplitude κ_{a2}. The performance is only affected by increasing noise.

Varying the Trend between Stimulation Blocks and the Noise

To investigate if a trend may affect the goodness of the estimates two source signals were modelled and mixed with each other. An HRF signal with increase κ_m ($\kappa_m = [0.5, 1, 3, 5]$) ($\kappa_s = 30, \kappa_p = 15, \kappa_b = 20$) and a noise signal with $\sigma = [1, 5]$. Figure 6.20 displays the median, 0.25-quantile and 0.75-quantile of error indices of 500 simulations for the Fast ICA algorithm. Each subplot displays the results for $\kappa_m = [0.5, 1, 3, 5]$. As it can be seen from Figure 6.20, a linear trend between phases does not essentially influence the estimation, only the noise parameter σ influences the quality of estimation.

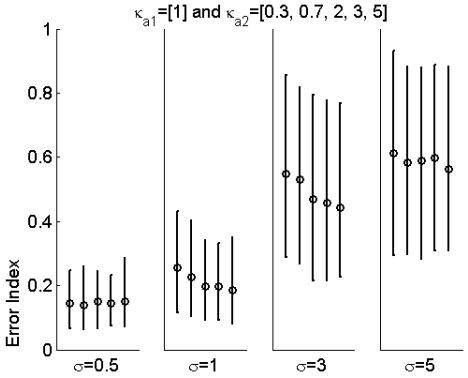

Figure 6.19: Error Indices of 500 simulations by varying the signal amplitude κ_{a2} and the noise σ, displaying median, 0.25-quantile and 0.75-quantile. Each subplot displays for a specified value of σ the results for $\kappa_{a1} = 1$ and $\kappa_{a2} = [0.3, 0.7, 2, 3, 5]$ (results of FastICA algorithm).

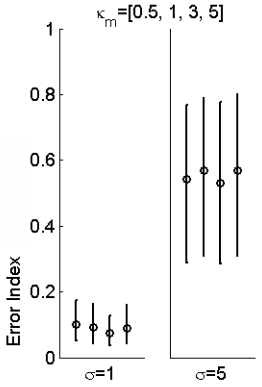

Figure 6.20: Error Indices of 500 simulations by varying the trend between phases κ_m and the noise σ, displaying median, 0.25-quantile and 0.75-quantile. Each subplot displays for a specified value of σ the results for $\kappa_m = [0.5, 1, 3, 5]$ (results of FastICA algorithm).

Varying the Trend within Stimulation Blocks and the Noise

The trend within one stimulations block $\kappa_n = [0.5, 1, 3, 5]$ and $\sigma = [1, 5]$ was varied. ($\kappa_s = 30, \kappa_p = 15, \kappa_b = 20, \kappa_a = 1$). The results were the same, the estimation was not corrupted by a trend or signal amplitude, it was only corrupted by the noise parameter. Therefore, a figure is not displayed for this example.

A further simulation was performed with five source signals. Four source signals were taken

from Figure 6.10 (left side), where the noise was varied as $\sigma = [0.5, 1, 3, 5]$. The fifth source signal is shown in Figure 6.5 (trend within one stimulation block). This signals were mixed by a 5×5 mixing matrix and five independent components were estimated. And exemplary results is shown in Figure 6.21. The left side displays the mixed signals and the right side displays the estimated independent signals. As it could be seen the algorithm lacks in estimating the two hemodynamic response signals exactly. The median of 500 error indices are the following: 10.0700 (for $\sigma = 0.5$), 8.6045 (for $\sigma = 1$), 6.0321 (for $\sigma = 3$), and 7.1631 (for $\sigma = 5$). These medians are much larger compared with other estimation with only one HRF. This might be explained that the two HRF signals have the same frequency of their stimulation and resting blocks, (see exemplary signals in Figure 6.21). Therefore, these two signals are not independent from each other. The independence assumption of source signals is violated resulting in worse estimation of source signals.

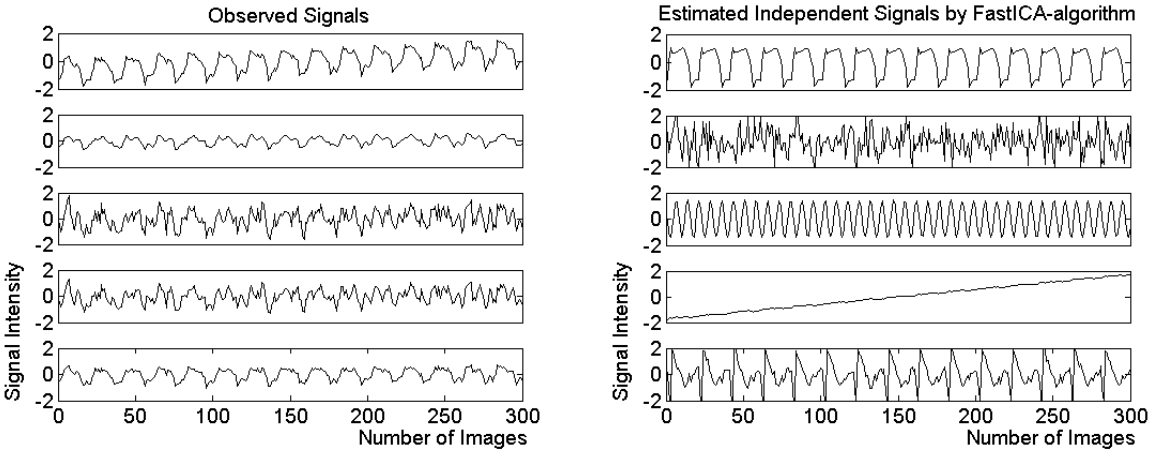

Figure 6.21: Time courses of mixed signals (left side) and estimated independent signals (right side).

In a final simulation of this subsection, four source signal were mixed linearly. An HRF signal with dynamic signal amplitude decrease as modelled in Figure 6.6 is the first signal. The noise signal is the second one. It varied with a noise parameter $\sigma = [0.5, 1, 3, 5]$. Furthermore, a sinus function ($P = 7$) and a trend function ($m = 2$) contributed to the mixture. The error indices for 500 simulations are displayed in Figure 6.22 (left side). On the right side of Figure 6.22 an exemplary ICA decomposition is shown. The error index of this example was $EI = 1.3668$. The estimation of the independent components was not optimal as indicated by the larger error indices. It can be seen that the trend function is still affected by the HRF signal.

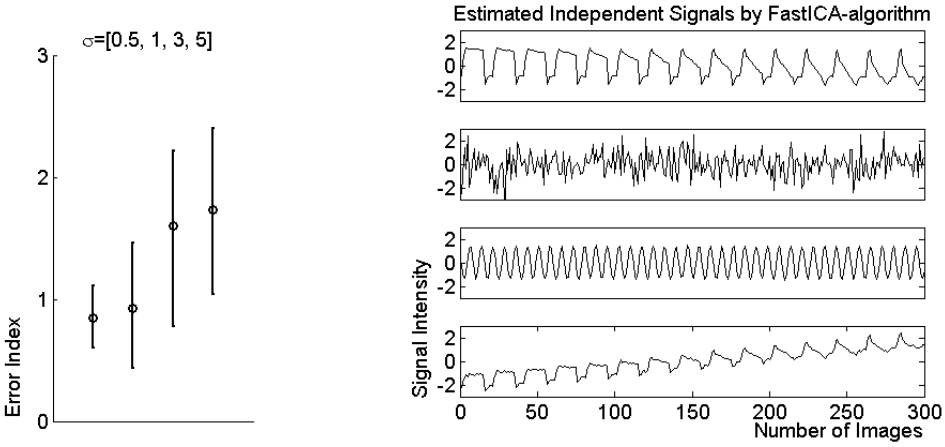

Figure 6.22: Error Indices of 500 simulations by varying dynamically the signal decrease within blocks and the noise σ, displaying median, 0.25-quantile and 0.75-quantile (results of FastICA algorithm)(left side) and exemplary decomposition (right side)

Varying the Noise within the HRF Signal

In this subsection it should be investigated if an HRF Signal with increasing noise (see Figure 6.7) and an additional noise with varying intensity σ can be detected of a mixture of four signal. Additional source signals are the sinus function and the trend function. The error indices for 500 simulations are shown in Figure 6.23 (left side). The right side of Figure 6.23 shows an exemplary ICA decomposition. The additional increasing noise in the HRF signal does not affect the goodness of the estimated that much comparing to the previous results.

Varying the Shift Parameter

In the following five source signals are mixed: an HRF, a temporal shifted HRF ($\kappa_c = [1 : 10]$), random noise with $\sigma = 1$, a linear trend and a sinus function ($\kappa_s = 30, \kappa_p = 15, \kappa_b = 20$). Figure 6.24 displays the median, 0.25-quantile and 0.75-quantile of error indices of 500 simulations for the Fast ICA algorithm. The error index was the smallest at $\kappa_c = 5$ and very high for $\kappa_c = 1$ or $\kappa_c = 2$. In these cases the shift of the two signals was to small and two identical HRFs were estimated instead of two shifted HRFs. The same thing happened for large κ_c, e.g. $\kappa_c = 10$ which would mean that the stimulation is shifted in the resting condition. There is the additional problem of ICA that the method is not able to estimate the signs and consequently for large κ_c identical HRFs are estimated. And explanation

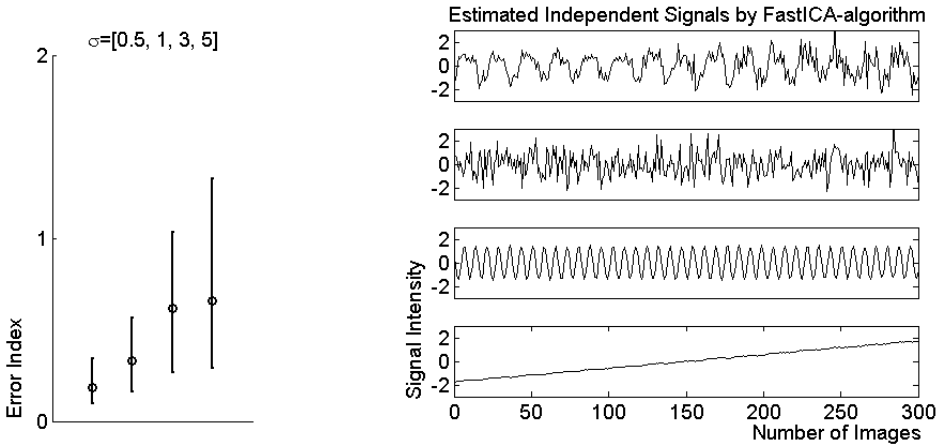

Figure 6.23: Error Indices of 500 simulations for HRF signal with increasing noise and by varying additional the noise σ of the noise signal, displaying median, 0.25-quantile and 0.75-quantile (results of FastICA algorithm)(left side) and exemplary decomposition (right side)

therefore is again that the assumption of independent source signals is violated. For a small temporal shift, the two HRF signals are thought to be very similar and therefore not independent from each other. For large temporal shifts, the second HRF signal is the reciprocal of the first HRF signal. And therefore they are also not independent from each other explaining the large error indices.

Varying the Number of Mixtures

In the following four source signals (see Figure 6.10 (left side)) are mixed linearly. The number of mixtures was varied as $N = [4, 10, 20, 50]$ and the noise was varied as $\sigma = [0.5, 1, 3, 5]$. Always four independent components were estimated. Figure 6.25 displays the median, 0.25-quantile and 0.75-quantile of error indices of 500 simulations for the Fast ICA algorithm. Each subplot displays the results for $N = [4, 10, 20, 50]$. The error index does not depend essentially on the number of mixtures N, it only depends on the σ.

Varying the Number of Stimulation Blocks

In the following four source signals (see Figure 6.10 (left side)) are mixed linearly by a 4 by 4 matrix $\mathbf{A}$ resulting in four observed signals. Thereby the number of stimulation phases was varied as $\kappa_p = [3, 5, 10, 15, 20, 50]$ to obtain different lengths of experiments, i.e. time series, and the noise was varied as $\sigma = [0.5, 1, 3, 5]$. Always four independent components

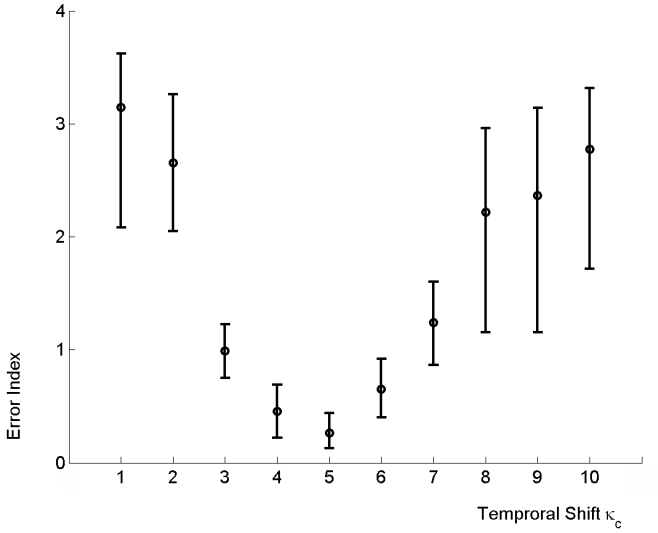

Figure 6.24: Error Indices of 500 simulations by varying the temporal shift κ_c, displaying median, 0.25-quantile and 0.75-quantile (results of FastICA algorithm).

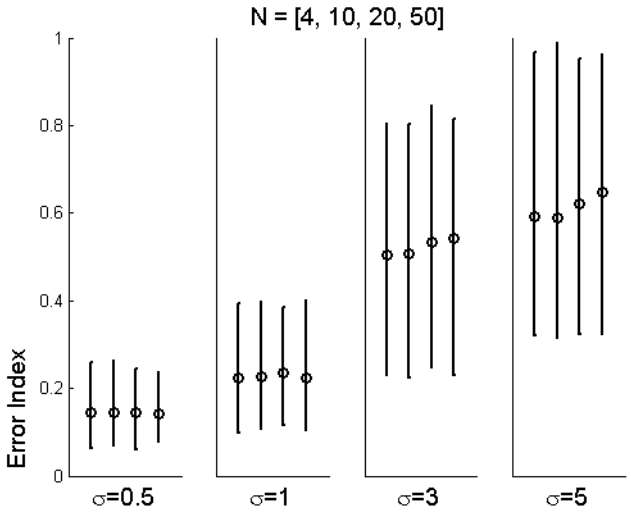

Figure 6.25: Error Indices of 500 simulations by varying the number of observations N and noise σ, displaying median, 0.25-quantile and 0.75-quantile. Each subplot displays for a specified value of σ the results for $N = [4, 10, 20, 50]$ (results of FastICA algorithm).

were estimated. Figure 6.26 displays the median, 0.25-quantile and 0.75-quantile of error indices of 500 simulations for the Fast ICA algorithm. Each subplot displays the results

for $\kappa_p = [3, 5, 10, 15, 20, 50]$. The error index depends on the number of phases κ_p, in the way that the longer the time series is, the smaller is the error index. This indicates that a sufficiently number of time points t is necessary for the estimation of independent components. The index also depends on the σ as was found out in previous simulations.

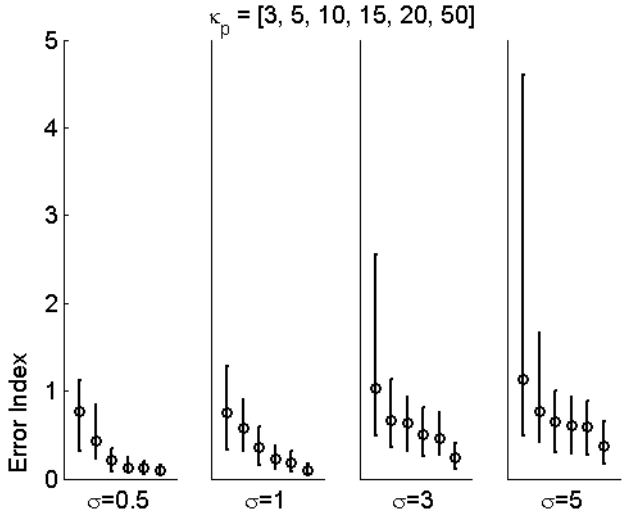

Figure 6.26: Error Indices of 500 simulations by varying the number of phases κ_p and noise σ, displaying median, 0.25-quantile and 0.75-quantile. Each subplot displays the results for $\kappa_p = [3, 5, 10, 15, 20, 50]$ (results of FastICA algorithm).

So far the results of simulation studies can be summarized that the FastICA algorithm outperformed the other ICA algorithms, see Figure 6.17. This was also observed by other simulation studies, but the results were not presented. This algorithm shows good results with varied HRF (amplitude, signal increase or decrease, number of mixtures). But it lacks at estimating two different HRFs from mixtures (temporal shift, and mixture of HRF and HRF with signal decrease) which is explained that the assumption of independent source signals is violated. Considering real fMRI data this would mean that it is difficult to estimate two HRF functions maybe originating from different regions that are time delayed by only some seconds (see Figure 6.24). For smaller delays the ICA algorithm would estimate them as two identical signals and consequently the two signals would be summarized in one independent component. Furthermore, it would be difficult to estimate two or more HRF signals that look similar but one signal is affected by a trend within the stimulation blocks, the other not, see Figure 6.21. Dynamic changes over the time in the signal are a smaller problem (see Figures 6.22 and 6.23). These changes within a signal can

be well estimated with ICA. This result motivated for the fMRI study in Chapter 7 assuming that there are dynamic changes in the fMRI signals due to learning related processes within and between repeated fMRI sessions of the subjects which should be detected with ICA.

6.5 Over- and Underestimation of the Number of Independent Components

A problem in ICA application is that it is not clear how many source signals are hidden in a mixture of signals. Underestimation of the dimensionality will discard valuable information and result in suboptimal signal extraction. Overestimation, however, results in a large number of spurious components due to underconstrained estimation and a factorization that will overfit the data, harming later inference and dramatically increasing computational costs [Beckmann and Smith, 2004]. Referring to fMRI data this possibly might mean that setting the number of independent components too high may result in functionally connected regions split into separate components. One way of determining the number of components may be done by singular value decomposition (SVD) of the matrix $\mathbf{X}\mathbf{X}^T$. The dimensionality is then chosen by selecting those eigenvalues exceeding some predefined threshold (e.g. Kaiser-Guttman criteria) or by scree plots. Another way of determining the number is done by the Bayesian information criterion (BIC) as proposed in [Hu et al., 2005]. They compute the BIC as

$$BIC = 2\ln(f(x|\theta_0)) + k\ln(T), \tag{6.12}$$

where x denotes the observed data, θ are the model parameters, which might be the number of unknown independent components. The parameters θ_0 are the parameters that maximize the posterior probability density function $f(x|\theta_0)$, k is the number of parameters and T is the sample size. The model with k parameters, that minimizes the criterion is chosen, k corresponds to the number of independent components, constrained by PCA dimension reduction.

Some simulation studies concerning the under- and overestimation of the number of source signals are performed. Therefore four source signals $\mathbf{s}(t)$ (see Figure 6.10 (left side)) are multiplied by an 8×4 mixing matrix to obtain eight mixed signals $\mathbf{x}(t)$. In a first step just two independent components are estimated from the mixture of the eight signals. The FastICA algorithm estimated the HRF signal and the sinus signal as two components of the mixture (see Figure 6.27, left side), whereas the Molgedey and Schuster algorithm,

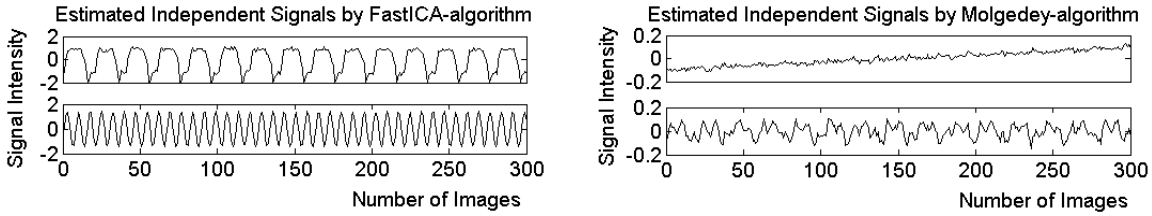

Figure 6.27: Underestimating the number of independent components. Estimating two independent components from a mixture of eight signals from four source signals. (FastICA algorithm - left side, MOLGEDEY and SCHUSTER algorithm - right side)

for instance, estimated two signals which are still mixed (see Figure 6.27, right side). The result of MOLGEDEY and SCHUSTER algorithm was also observed by the other ICA algorithms. These results showed clearly that the performance of ICA algorithms is degraded with a decreasing number of estimated components, since the estimated signal is disrupted by some noise (except FastICA algorithm were a selection of some source signals is made).

A very interesting finding for the FastICA algorithm occurred by underestimating the number of independent components. In the exemplary result in Figure 6.27 (left side), the HRF signal and the sinus signal were estimated as independent components. It should now be investigated, which signals are estimated by estimating only one, two or three independent components from a mixture of four signals (four source signals). The results of simulation studies are summarized in Table 6.2. The ICA decomposition for each predefined number of independent components (1, 2, and 3) run 500 times and the number of estimated source signals was counted. The noise signal is detected least of all source signals as it could be seen in Table 6.2, possible because the noise signal was the at least structured signal. The signal detected mostly was the sinus signal, then the HRF signal and trend signal, but the differences are not large. Moreover, the simulation run for different noise parameters $\sigma = [0.5, 1, 3, 5]$. As a result, an increasing σ does not influence the estimation of source signals, possible because the noise signal is estimated least.

On the other hand an overestimation of the number of independent components is a smaller problem. After the FastICA algorithm estimated the four hidden signals correctly the algorithm was truncated (see Figure 6.28, left side). Instead of estimating the dependencies between the components, the FastICA algorithm is based on computing the deviations between the single components and a gaussian distribution using the kurtosis. This approach allows a component-wise estimation, so that in case of overestimation not all components

Table 6.2: Underestimating the number of independent components (H = hypothetical HRF signal, N = noise signal, S = sinus signal, T = trend signal) Each cell contains the counts of that signal in 500 simulations

Estimating 1 independent component				
σ	H	N	S	T
0.5	168	7	173	152
1	154	7	195	144
3	172	2	179	147
5	163	2	178	157
Estimating 2 independent components				
σ	H	N	S	T
0.5	320	19	351	310
1	326	13	340	321
3	335	17	339	309
5	322	11	364	303
Estimating 3 independent components				
σ	H	N	S	T
0.5	486	47	489	478
1	485	56	487	472
3	474	61	485	480
5	479	52	484	485

must be estimated. The Molgedey and Schuster algorithm estimated six independent components from eight mixed signals. This algorithm estimated the three 'real' signals (HRF signal, trend signal, sinus signal) and probably decomposes the noise signal in further components, whereas two estimated components represent not much signal contribution, see Figure 6.28 (right side).

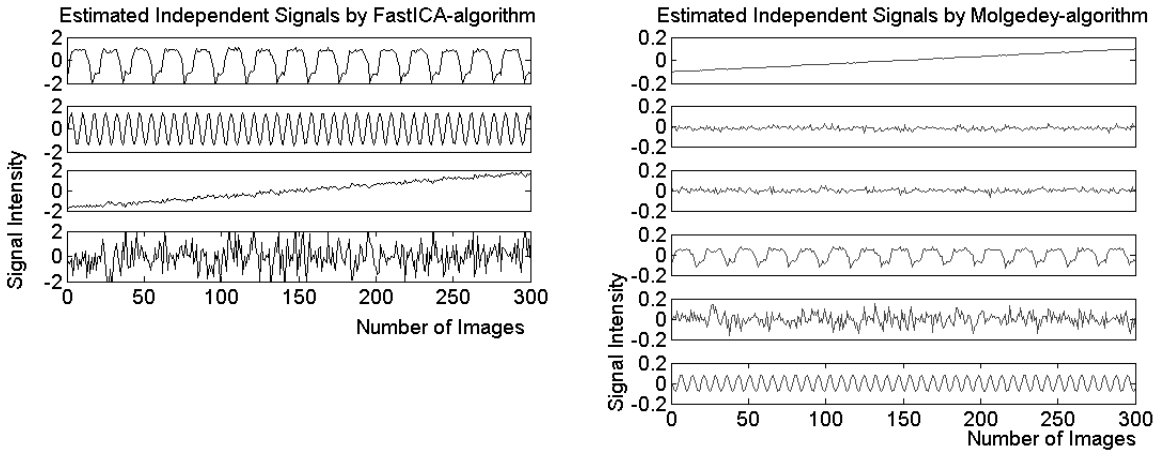

Figure 6.28: Overestimating the number of independent components. Estimating six independent components from a mixture of eight signals from four source signals. (FastICA algorithm - left side, MOLGEDEY and SCHUSTER algorithm - right side)

6.6 Comparing the results of GLM analysis of mixed signals with and without included ICA

The general linear model (GLM) is a multivariate linear regression method (see Section 4.1). In fMRI data the GLM tests the time series $x_i(t)$ of a voxel $\mathbf{v}^i$, $i = 1, \ldots, N$ against the hypothetical reference function $r(t)$ (hypothetical HRF or boxcar-function - a vector coded with zeros and ones, zeros for resting blocks and ones for stimulation blocks).

In this section, the results of GLM analysis will be compared for two cases in a simulation study: In the first case the GLM is directly applied to the observed mixed signals $x_i(t)$. In the second case an ICA decomposition was first applied to the observed mixed signals. After determining the signal representing the HRF signal (by cross correlating that signal to $r(t)$), the GLM was applied to that estimated HRF signal. This will be explained in detail later.

The GLM of a time series $x_i(t)$ is defined as

$$x_i(t) = \gamma_0 + \gamma_1 \cdot r(t) + \varepsilon, \tag{6.13}$$

where γ_0 and γ_1 ($\gamma_0, \gamma_1 \in \mathbf{R}$) are the estimated coefficients and $\varepsilon \sim \Phi(0, \sigma^2)$ is residual noise. Only the parameter γ_1 is of further interest, since this parameter describes the relation of the observed time series $x_i(t)$ and the hypothetical reference function $r(t)$.

In a simulation study, source signals are mixed with a random mixing matrix $\mathbf{A}$ to obtain the observed signals $x_i(t)$ which are then used in the GLM with or without included ICA. The procedure is repeated 500 times with new random mixing matrices and new generated noise signals for each simulation run. It is the aim of the study to assess the percentage of simulation runs in which a significant effect in the test of the null hypothesis $H_0 : \gamma_1 = 0$ can be detected (i.e. a p-value smaller than $\alpha = 0.05$ is obtained).

As already mentioned, the analysis is performed for two cases:
• In the first case, the GLM analysis is directly performed on the observed mixed signals $x_i(t)$. In the first subcase two signals $x_1(t)$ and $x_2(t)$ are observed, mixed from two source signals, an HRF signal and a noise signal with the parameter σ varying in the set $[0.5, 1, 3, 5]$. In a second subcase four signals are observed, mixed from four source signals (HRF signal, noise signal with the parameter σ varying in the set $[0.5, 1, 3, 5]$, trend function, and sinus function, as in Figure 6.10 (left side)). For each subcase and each observed signal $x_i(t)$, the GLM analysis was performed with two reference functions $r(t)$: a boxcar-function (BOX), and the hypothetical hemodynamic response function (HRF) (see Equation (6.1)).
The percentage of simulation runs where the null hypothesis $H_0 : \gamma_1 = 0$ has been rejected is given in Tables 6.3 (for two source signals) and 6.4 (for four source signals).
The percentage of significant results (empirical power) behaves as expected: The results for the two or four observed signal are equal expect from random fluctuations, the power decreases with increasing noise and the power is larger when the observed signals are compared to the HRF and not to the raw boxcar-function. Furthermore, the power is larger when only the HRF is mixed with noise and no additional signals which are not included in the GLM.

Table 6.3: Percentage of simulation runs with significant test results for the parameter γ_1 for two source signals (HRF signal and noise), separately for the two observed variables, for two different reference functions and for different values of σ

r(t)	observed signal	σ 0.5	1	3	5
BOX	x_1	86.8	75.0	47.8	31.8
BOX	x_2	87.8	71.0	43.2	34.4
HRF	x_1	92.2	84.0	65.2	44.6
HRF	x_2	93.8	84.6	63.4	48.6

Table 6.4: Percentage of simulation runs with significant test results for the parameter γ_1 for four source signals (HRF signal, noise, trend signal, and sinus signal), separately for the four observed variables, for two different reference functions and for different values of σ

r(t)	observed signal	σ 0.5	1	3	5
BOX	x_1	63.8	59.0	39.0	25.2
BOX	x_2	65.4	57.2	38.2	36.6
BOX	x_3	64.0	59.4	36.6	25.2
BOX	x_4	60.8	59.8	35.4	29.8
HRF	x_1	79.4	77.4	59.2	46.0
HRF	x_2	83.2	74.8	60.4	45.6
HRF	x_3	82.6	78.6	58.0	46.6
HRF	x_4	79.4	80.0	56.0	44.0

• In the second case, a FastICA algorithm decomposed the two or four observed signals into independent source signals. Then the signal reflecting the neuronal response was chosen by cross correlating that signal to $r(t)$. The GLM analysis was finally performed with the selected signal instead of the observed signals. Again, both reference functions (BOX and HRF) were used in parallel. With this procedure, all simulation runs gave significant results (100% empirical power, not displayed in a table). Thus the inclusion of the ICA drastically improved the results of the GLM by combining the two observed variables and selecting the best adapted source signal. Of course, in a strong sense, the rules of the inference statistic are harmed here by choosing the signal with the best fit to the hypothetical HRF.

Note that only one reference signal $r(t)$ was given representing either the hypothetical response function (HRF) or the experimental design (BOX). If explicit information of contributing signals beyond the hypothetical response function is available, this can also be included in GLM as a vector of response functions (without the inclusion of an ICA), possibly improving the results of the GLM test. But in applications the form of the possibly disturbing signals (like noise, trend and sinus signals) is often not known. The ICA works without that knowledge and thus can improve the GLM also in that such cases.

6.7 Illustrative Results of Time Series Decomposition

The chapter of simulation studies should be closed with the time series decomposition as described in Section 4.3 of the of the observed signals from Figure 6.10 (right side). From the literature it is known that, every time series can be decomposed into a trend component, a seasonal component, and an irregular remaining component [Schlittgen and Streitberg, 1995]. After removing the trend and the seasonal component from the time series, the remaining part should be stationary to apply time series characteristics such as autocorrelation functions and frequency analysis to the time series. The time series decomposition was performed with the software package *R* (*http://www.cran.r-project.org*). For the decomposition a period for the seasonal component must be given. As we know that the length of a phase is given with 20 images (time points), the period is given by $P = 20$. Note that, if there are any information of periods of further seasonal components, like the sinus function mimicking the heart beat in our example (P=7), also such a period might be chosen. But only the period of one seasonal component can be given at a time with this software package.

The results of the time series decomposition of the four mixed signals are shown in Figure 6.29. As it could be seen from this figure, the HRF signal was estimated as a seasonal component and the trend component could also be estimated. These estimated components give information about the weights of the mixing matrix. See therefore the HRF signal of the third decomposition. The contribution of the HRF signal to the mixture was only $a_{3,1} = 0.1146$ (see mixing matrix $\mathbf{A}$ in Equation (6.11)) which is reflected in the small amplitude for the seasonal component. This is an advantage of time series decomposition compared with ICA, since in ICA it is not possible to estimate the scaling factors (see Section 5.1). A second advantage of time series decomposition compared with ICA is directly obvious from the estimates. The estimated components not just inform about the weights of the components, they also give information about the sign of the hidden components which is in general also not possible with ICA (see Section 5.1 again).

Regarding the residual components some drawbacks of time series decomposition are apparently. The residual component is not only white noise, this component is affected by a sinus function. This is obviously since the sinus functions was modelled as a source signal but so far this signal was not estimated by the seasonal or trend component. Consequently, a drawback of time series decomposition is that only a limited number of components can be detected. If there are several seasonal components as in this case an HRF signal with period of $P = 20$ and a sinus signal with a period of $P = 7$, the time series decomposition

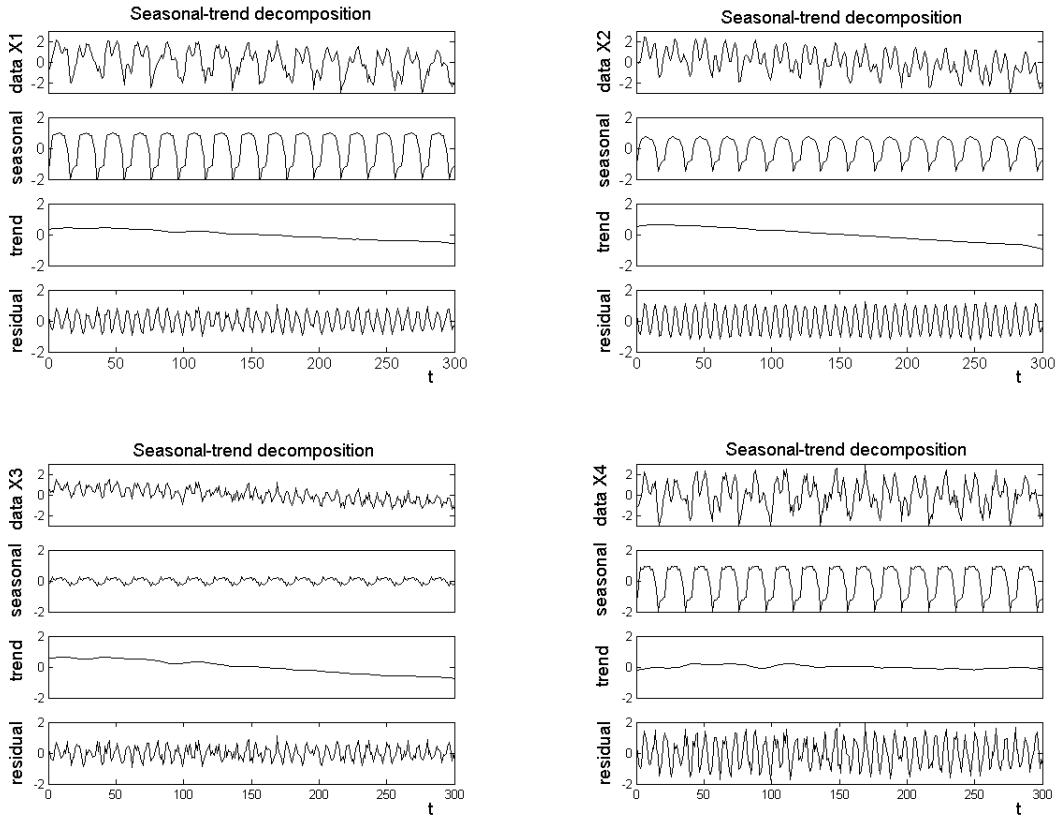

Figure 6.29: Time series decomposition of observed signals

cannot estimate both seasonal components in one decomposition. If this residual signal would be once again decomposed into a trend, further residual and seasonal component with $P = 7$, than also the sinus function is detected. A frequency analysis should help to determine the hidden frequencies.

Nevertheless, it might be possible to combine both algorithms, namely ICA and time series decomposition. It is conceivably to use the ICA in a first step to get independent components like the HRF signal, i.e. the periodic signal, the trend signal and noise signal. The time series decomposition is then performed as a postprocessing step to determine the sign and amplitude of the HRF signal for instance. Although not all signals, particulary unexpected ones, can be detected by time series decomposition, the amplitude and sign of some relevant independent components can be estimated.

Concluding this section some time series statistics of chosen source signals should be considered, namely the HRF signal and the noise signal. These two signals were chosen because of the complementarity of a structured HRF signal and an unstructured noise signal. The

time series analyses are performed with *SPSS* (*http://www.spss.com*) for test for stationarity and autocorrelation function and with *MATLAB* for test for gaussian distribution, frequency analysis, and empirical probability density function.

Test for White-Noise Process

The Wald-Wolfowitz runs test was used to investigate if the signals are stationary or if the observations occur with some structure. The test revealed that the HRF signals is structured (median = 2.5096, number of sequences = 31, z-statistic = -13.880, p-value $< 10^{-3}$) but the noise signal is unstructured (median = 0.0036, number of sequences = 167, z-statistic = 1.851, p-value = 0.064). That means the HRF function is not a stationary signal and preprocessing steps like removing a seasonal or trend component are necessary.

Autocorrelation Function

Figure (6.30) displays the autocorrelation functions (ACF) for different lags $\tau = 1, \ldots, 30$ of the HRF signal (left side) and noise signal (right side). What can be clearly seen from that figure is that the HRF signal is a structured signal (left side), where the value of the actual time point depends on the value of the previous ones. Whereas the noise signal is an unstructured signal indicating by low values of autocorrelation functions.

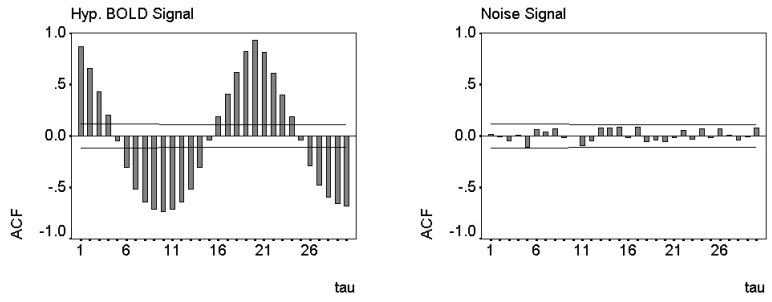

Figure 6.30: Autocorrelation functions of HRF signal (left side) and noise signal (right side)

Test for Gaussian Distribution

Since it is assumed, that the "real" source signals (except the noise signal) have nongaussian densities, the normality of the signals was analyzed with Kolmogorov-Smirnov test. The test revealed that the HRF signal is nongaussian distributed (KS-statistic = 0.5829, p-value $< 10^{-3}$) but the noise signal is gaussian distributed as expected (KS-statistic = 0.0491, p-value = 0.4561). The critical value for the hypothesis of the Kolmogorov-Smirnov test is 0.077.

Frequency Analysis

Figure (6.31) displays the frequency spectra of the HRF signal (left side) and noise signal (right side). The frequency spectra of the HRF signal is characterized by a peak at $f = 0.05$ Hz corresponding to a period of $P = 20$ time points of the experimental design. In the frequency spectra of the noise signal no periods can be detected.

In the frequency spectra of the observed mixed signals of Figure 6.10 (right side) (the frequency spectra are not displayed), two peaks are clearly found. On peak at $f = 0.05$ Hz, the HRF signal, and a second peak at about $f = 0.15$ Hz, representing the frequency of the sinus signal with $P = 7$.

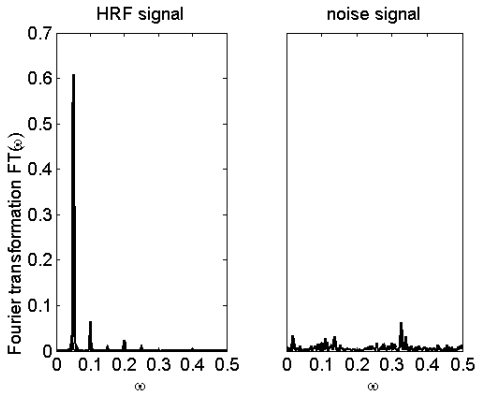

Figure 6.31: Fast Fourier Transformation of HRF signal and noise signal

Empirical Probability Density Function

Figure 6.32 displays the empirical probability density function of the HRF signal (left side) and noise signal (right side). These empirical probability densities are performed through

kernel probability density estimations. The estimated pdf of the noise signal is quite similar to a gaussian pdf, whereas the pdf of the HRF signal is not. Note that, the densities of the signals, e.g. HRF signal can only be regarded as pseudo-densities because they are not densities in the intrinsic sense, see therefore Section 5.1, where these estimates are more deterministic than probalistic measures.

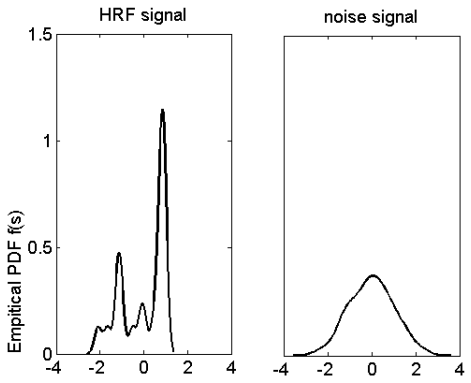

Figure 6.32: Graph of estimated probability density functions of HRF signal and noise signal

7 An Auditory Working Memory fMRI Study and ICA-Results

In this section an auditory fMRI study investigating a working memory (WM) task should be introduced and the results of ICA will be presented. The basic concept of WM refers to "a brain system that provides temporary storage and manipulation of the information necessary for cognitive tasks" [Baddeley, 1992]. In a WM task the arriving information must be maintained, recalled and, compared with test items according to previously instructed rules. In this study, a so called one-back task served a WM task. In this task the WM content had to be continuously reorganized and updated.

7.1 Material and Method

Acoustic Stimulation

For the experiment frequency modulated (FM) tones were used as acoustic stimuli. In principle, FM tones are sinus functions with the following parameters: the sampling frequency (F_s), given as $F_s = 44100\ Hz$, a start frequency f_1 and an end frequency f_2, each given in Hz. For these FM tones $f_2 = 2 \cdot f_1$ for rising tons and $f_1 = 2 \cdot f_2$ for falling tons . Additionally a duration T in sec is needed. All FM tones were computed using the software $MATLAB$, where the time scale vector $t = 1, \ldots, T$ in steps of $1/F_s$ and F_s is the sampling frequency (1 sec corresponds to 44100 sampling points). The FM tone is then defined as

$$FM(t) = \sin(2 \cdot \pi(a/2 \cdot t + f_1) \cdot t), \tag{7.1}$$

where $a = (f_2 - f_1)/T$. Additionally a ramp of 10 $msec$ was added to the beginning and the end of the tone. To demonstrate this, Figure 7.1 shows an 100 $msec$ FM tone with increasing frequency from 0.5 kHz to 1 kHz, but in the fMRI study longer FM tones (300 $msec$ - 600 $msec$) are used. The FM tones were arranged in stimulation blocks of 30 seconds. Each block consisted of 30 randomized FM tones, 15 rising tones and 15 falling

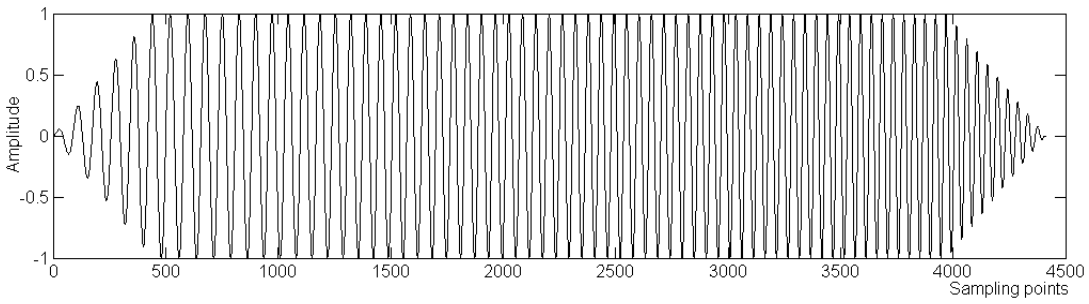

Figure 7.1: Plot of a frequency modulated tone of 100 msec with increasing frequency from 0.5 kHz to 1 kHz.

tones. The frequency range of the FM tones varied between 0.5 - 2 kHz in steps of 0.1 kHz. Six different durations were used; 300 $msec$, 350 $msec$, 400 $msec$, 550 $msec$, 600 $msec$, and 650 $msec$. Each duration was created five times in a block by randomly order. A gap of 525 $msec$ between two tones was created. The FM tones were presented at five different sound levels covering a range of 24 dB in steps of 6 dB all at a comfortable loudness. Each sound level was presented five times in a block in randomly order. One experimental session consisted of 15 alternating stimulus and resting blocks (see the experimental paradigm in Figure 7.2, gray indicates stimulus blocks and white indicates resting blocks).

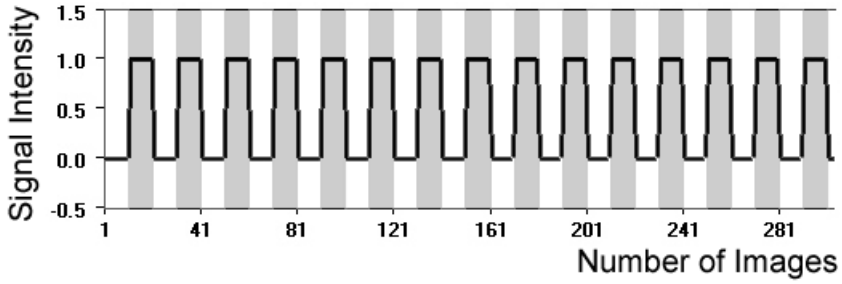

Figure 7.2: Experimental paradigm of fMRI experiment.

Task procedure

The task of the subjects during the session was a one-back working memory task. The subjects continuously had to compare the actual tone with the tone presented one back in the sequence and had to indicate whether the two tones matched in direction (rising or falling) by button pressing. Figure 7.3 displays the FM tones of one exemplary stimulations

block. Each block consisted of 30 FM tones with 12 targets in randomized order. Targets are indicated as stars in Figure 7.3.

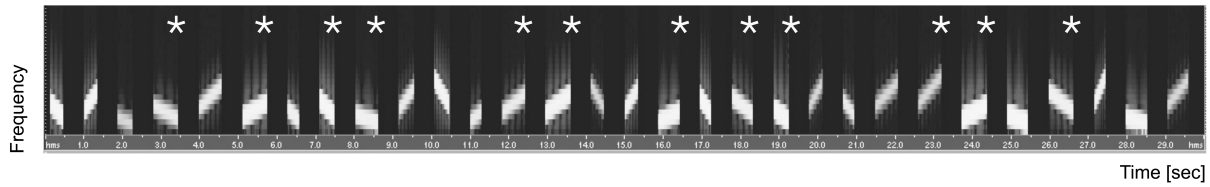

Figure 7.3: Targets in experiment for one exemplary stimulation block

Subjects

In the study 3 females and 3 males (22-27 years old, mean age 24) were scanned. Each subject repeated the measurement five times over a period of five weeks, resulting in 30 overall sessions. All subjects were naive according to FM tones, i.e. they never performed a discrimination task before this study but they were familiar to the fMRI procedure. All subjects were right-handers with normal hearing and gave written consent to the study, which was approved by the ethical committee of the University of Magdeburg.

Scanning procedure

The fMRI study was carried out on a 3 Tesla scanner (Siemens Trio, Erlangen, Germany) equipped with an eight channel head coil. For every subject a 3D anatomical data set of the subjects brain was obtained (192 slices of 1 mm each) with very high resolution. Functional images were acquired by echo planar imaging (EPI) sequence. The whole fMRI data set consisted of 310 volumes, i.e time points. Thereby, an fMRI image was recorded every 3 seconds. This is called the repetition time (TR) = 3000 $msec$. Further scanning parameters are the echo time (TE) = 30 $msec$; flip angle = 80°; field of view (FOV) = 192 mm; voxel size 64 $\times$ 64). 40 slices of 3 mm each (0.3 mm gap) covering the whole brain were recorded. Before the functional measurement an inversion-recovery EPI (IR-EPI) was recorded. This are images with the same geometry as the functional images. The functional data is projected on the IR-EPI images and afterwards this data is mapped to the 3D anatomical data set.

Behavioral Data

During the fMRI measurement the individual responses of the subjects are recorded additionally. In detail, this means the correct (H) and false (F) responses, missing (M) and correct rejections (R) according to the stimuli are recorded. With these data, the sensitivity index d' (signal detection theory [Green and Swets, 1966]) can be computed to draw conclusions about the subjects' task performances. The sensitivity index is defined by

$$d' = \Phi^{-1}(h) - \Phi^{-1}(f), \tag{7.2}$$

where h is the percentage rate of correct responses ($h = H/(H+M)$) and f is the percentage rate of false responses ($f = F/(F+R)$), respectively, and Φ^{-1} is the inverse of the standard gaussian distribution. This measure is mostly used for describing subjects performances in biological trials because it also considers the false responses. The sensitivity index is a measure that displays the percentage rate of correct and false responses to the gaussian distribution. The range of the sensitivity index is given about 5 for good task performances and -5 for bad task performances (considering that $\Phi^{-1}(1)$ and $\Phi^{-1}(0)$ are ∞ and $-\infty$, respectively). Furthermore, the response times in seconds to correct responses were recorded to obtain an average response time.

7.2 Data Analysis

In a first step, the data analysis was performed using the brain imaging analyzing and visualization tool *BrainVoyager* (University of Maastricht, Netherlands). In a second step, the relevant time courses were exported to *MATLAB* (MathWorks, Inc., Massachusetts, USA) for further computations and analysis.

Preprocessing

Before analyzing the fMRI data some preprocessing and normalization steps are necessary. In a first step a 3D motion correction is performed. Even though subjects are told to move as little as possible inside the scanner, some head movements are unavoidable, with the result that the same voxels do not represent the same location in the brain throughout the time. Therefore, one brain volume is taken as the reference volume and all the other volumes are repositioned for translation and rotation in all three dimensions until all volumes are the same position as the reference volume. Another important step is the spatial normalization. During an fMRI study, data are usually collected from several subjects.

But each individual brain differs in orientation, size and shape relatively to other brains. To compare activations of different subjects the individual brains should be matched to a standard brain. One method for spatial normalization is the Talairach transformation [Talairach and Tournoux, 1988], matching all brains according to eight points of the cerebral cortex: These eight points are the Anterior Commissure, Posterior Commissure, the most anterior and posterior point of the cortex, the most superior and inferior point and the most left and right point of the cortex. With these point the brain is subdivided into cuboids. All brains are then zoomed or rotated to match these cuboids.

ICA analysis

The ICA implemented in the software package *BrainVoyager* is a cortex-based ICA [Formisano et al., 2004] with FastICA algorithm proposed by Hyvärinen. For each subjects and each session an ICA with 30 independent components was performed. The number of 30 components was chosen because still most of the variance is explained with 30 components. An PCA for each subject and each session was performed in advance revealing that 10 - 30 components are needed to decompose the data set. In order to assume the same number of independent components for each subject and each session, 30 independent components were chosen as a fixed number to cover all important and relevant components. ICA was also performed with more or less than 30 components to demonstrate these results. On the one side, estimating only a few components (< 5 components) results in noisy components, where none of the time courses was correlated to the stimulation protocol, i.e. the box-car function. In general, the FastICA algorithm finds successive relevant components, see Section 6.5, but for this real data sets the FastICA algorithm can not judge if noise is an essential component. On the other side, estimating too much components (> 50 components) results in overlapping components where the activation clusters are not disjunct anymore. For the independent components activation map and associated averaged time courses are produced. The time series were normalized in $[0, 1]$. To select meaningful components, in a first step, the time courses of the components were sorted according to their correlation coefficients with the time course of the experimental paradigm. So far, the experimental paradigm was not part of the ICA estimation. The activation clusters were considered of those components which had the highest correlation ($|\rho| \geq 0.7$). Only voxels of an activation map with a $|z|$-score ≥ 2 (see McKeown et al., 1998*b*) of the time and at least 50 connected voxels were considered as activated voxels. Consider that an independent component might consists of several connected activation clusters. The most interesting component was the one with activation clusters in the auditory cortex (AC). Since it is an

auditory fMRI study, activation in auditory areas was expected. Therewith, the selected component was clearly defined by the time course (i.e. correlation to stimulation protocol) and activation clusters (i.e. activation in AC).

Definition of Volumes of Interest

The neuronal responses were analyzed within defined volumes of interest (VOI). The definition of VOIs was based on BRODMANN areas (BAs) [BRODMANN, 1909]. Figure 7.4 shows the BAs of the lateral view of the brain and Table 7.1 summarizes the names, locations and assumed involvement of BAs which were mostly relevant for our fMRI study, i.e. BAs in AC and frontal cortex, even there exists much more BAs. These information are taken from *http://www.fmri-easy.de/start1.htm* and *http:medical-dictionary.thefreedictionary.com*. The BAs were defined for each subject, each hemisphere, and each session based on the anatomy of each individual brain. But only the activated voxels within a BA were considered.

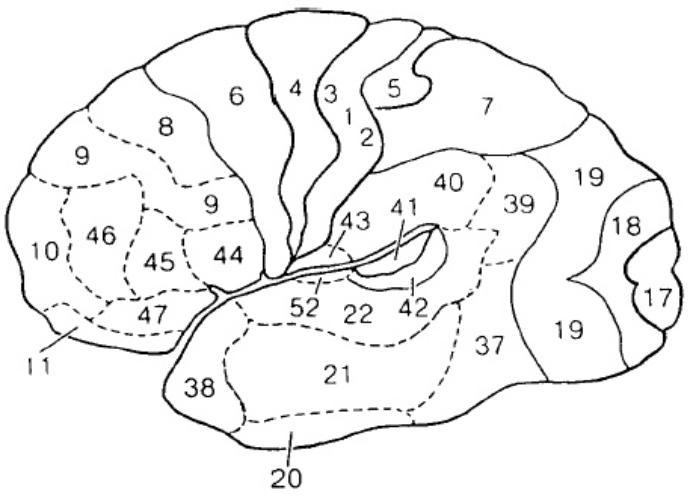

Figure 7.4: Brodmann areas

Time course analysis

In a following step, the changes in the time course of these VOIs over the five repeated sessions were investigated. First, the signal at each time point t was converted to a percent change in signal relative to baseline. The baseline was defined as the average signal from $t_b = -3$ to -1 ($t_b \in T$) of every stimulation block with $t_b = 0$ corresponding to the onset of stimulation. In a second step the time course signal was averaged over all 15 stimulation

Table 7.1: Brodmann areas: Their location and involvement

BA	**Name (Location)**	**Involvement**
BA06	agranular frontal area (frontal cortex)	planning of complex, coordinated movements; involved in memory (recognition) tasks with BA 32 and BA 46
BA09	(frontal cortex)	involved in working memory tasks (with BA 46)
BA10	frontopolar region (frontal cortex)	play a role in strategic processes; involved in memory retrieval and executive function
BA21	middle temporal area (temporal cortex)	play a part in auditory processing and language
BA22	superior temporal area (temporal cortex)	language processing; left: helps with generation and understanding of individual words; right: helps to tell the difference between melody, pitch, and sound intensity
BA40	supramarginal area (parietal cortex)	sequential processing
BA41	anterior transverse temporal area (temporal cortex)	processing of auditory (sound) information
BA42	posterior transverse temporal area (temporal cortex)	processing of auditory (sound) information
BA44	opercular area (frontal cortex)	involved in speech production; involved in processing of sequential auditory stimuli; keeps information in working memory
BA45	triangular area (frontal cortex)	BROCA's area, speech production
BA46	middle frontal area (frontal cortex)	involved in working memory tasks (with BA09)
BA47	orbital area (frontal cortex)	BROCA's area, speech production

blocks. A block was defined to include 5 images prior to stimulation, the stimulation phase (10 images) and 5 images posterior to stimulation. These 15 response blocks were averaged to give an average signal versus time for each subject, hemisphere, VOI, and session.

The response magnitude in each subject, each VOI and each session was quantified using measures computed from the percent signal change time course [Harms and Melcher, 2002]. The "time-average" percent change measures the overall response strength during the stimulation block. It was computed as the mean percent change from $t_b = 3$ to 10. This range beginning from $t_b = 3$ was chosen because it takes some time until the signal reaches its activation plateau. Changes in the signal amplitude over the sessions were tested by ANOVA with repeated measures.

7.3 Results

7.3.1 Behavioral Results

Since all subjects were naive according to FM tones, i.e. they had never performed a discrimination task before this study, the task was quite difficult for the subjects at the beginning. But all subjects showed strong improvements indicated by their rates which are displayed in Figure 7.5 for all subjects and all sessions. It was tested if these differences were significant using an analysis of variance (ANOVA) for repeated measurements, where the session is a factor with five levels. Before performing the ANOVA it was verified that the sensitivity indices (see Equation 7.2) are gaussian distributed (Kolmogorov-Smirnov test) and have equal variances (Mauchly sphericity test). The ANOVA revealed significant differences of the sensitivity indices between the repeated sessions ($F_{(1,4)} = 46.560, p < 0.001$). Additionally regarding the response times of the subjects, they also improved over time. Testing the response times with an ANOVA for repeated measurements revealed significant differences for the response time as well ($F_{(1,4)} = 33.828, p < 0.001$). The sensitivity indices and response time for all subjects and all sessions are shown in Figure 7.6.

7.3.2 ICA-Results and Time Series Analysis

An ICA with 30 components for each subject and each session was performed. The independent components consisted of activation maps and associated time courses. According to the time courses, the components can be classified as oscillatory functions, trend functions,

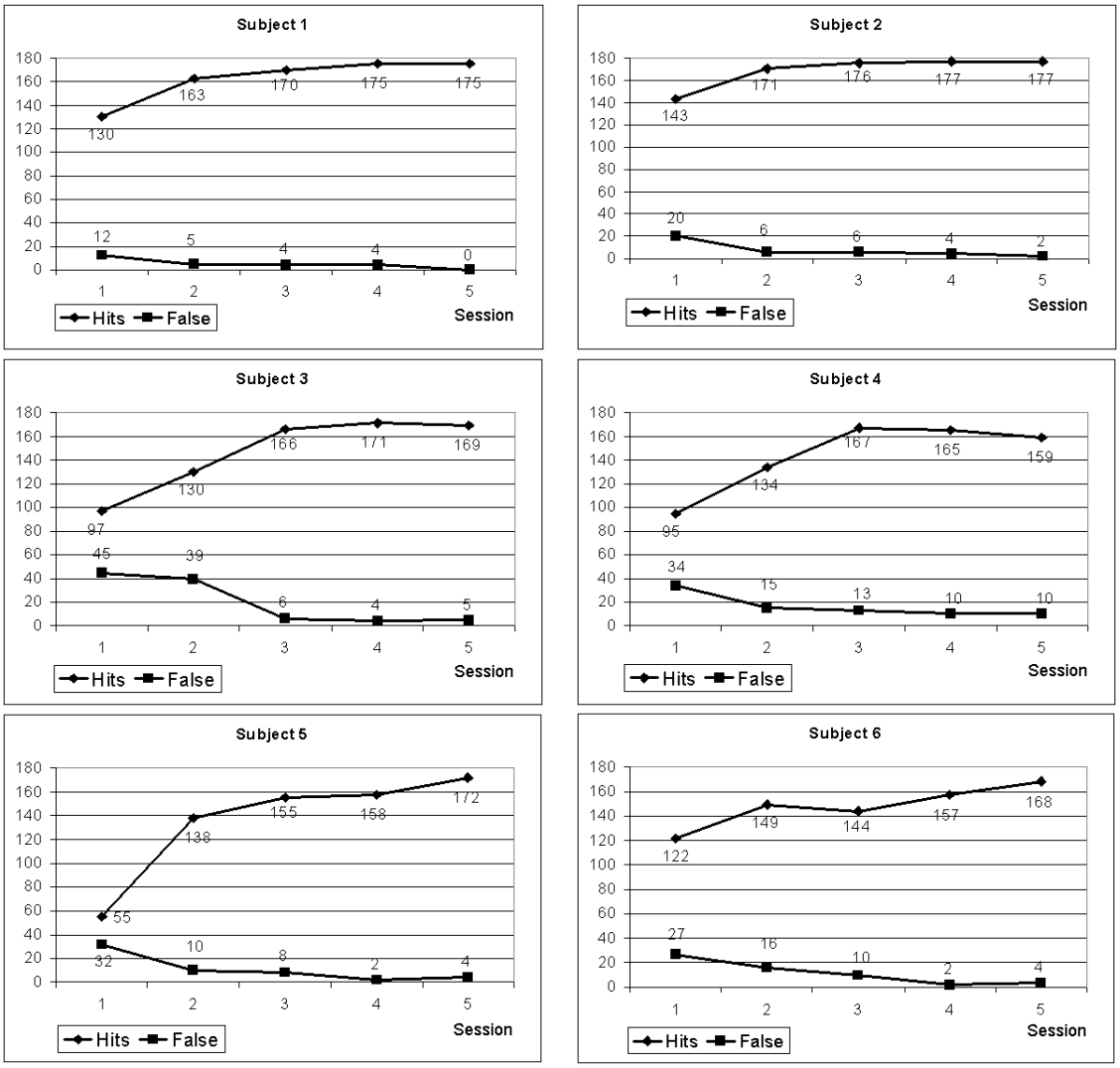

Figure 7.5: Hits and false responses of subjects

noise functions, and some time courses possibly indicating neuronal processes. Figure 7.7 displays the 30 independent time courses of subject 3 for the first session. After selecting the meaningful component by cross-correlating the 30 independent component time courses to the experimental paradigm, the components with correlation coefficient $\rho \geq 0.7$ were further considered. In this example, component 20 had a correlation coefficient of $\rho = 0.7527$ to the stimulation protocol. This component was the only one with a correlation coefficient $\rho \geq 0.7$ and activation clusters in auditory cortex. Even though in the example of subject 3, component 8 seems to be highly correlated to the stimulation protocol, its correlation coefficient was $\rho = 0.4498$ because of time delays to the stimulation protocol that might

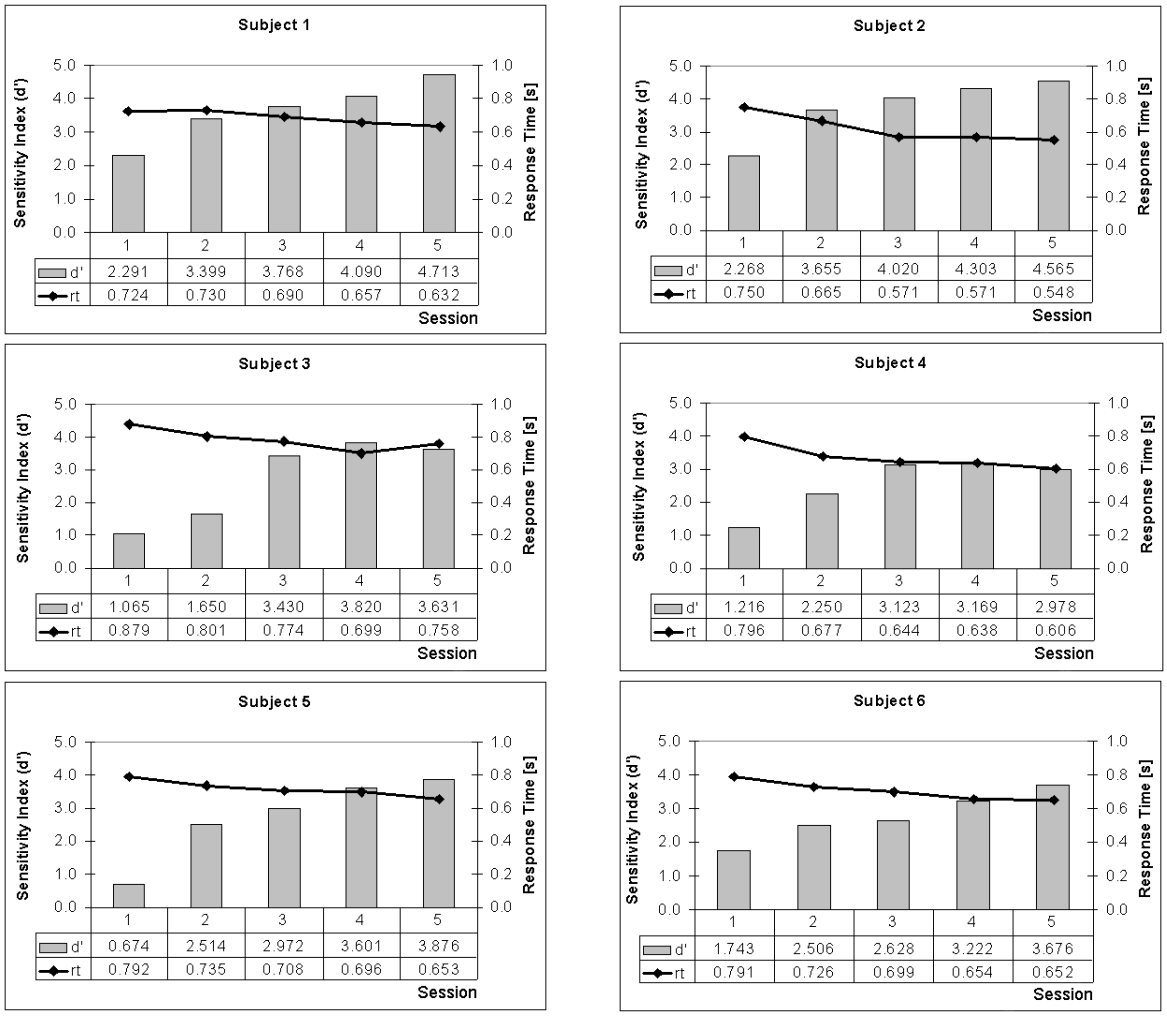

Figure 7.6: Sensitivity indices and response times of subjects

be a sign for later processing of the stimuli or the task, but this component showed no activation in the AC or frontal cortex that would confirm the assumption. In general for all subjects there was at least one component with a correlation coefficient $\rho \geq 0.7$. In cases were more components had a $\rho \geq 0.7$ only one of these components showed activations in AC. Consequently, the selection of one component with correlation coefficient $\rho \geq 0.7$ and location in AC was unique for each subject and each session. After selecting the component with activation clusters in the AC, clusters in other cerebral regions of that component were inspected. Besides activation in auditory cortex we found additional clusters in areas which are supposed to be involved in maintenance and attention processes and in areas which are involved in somato-sensory processes, or motor processes caused by pressing a button to indicate targets. These clusters were defined as VOIs according to Brodmann areas (see

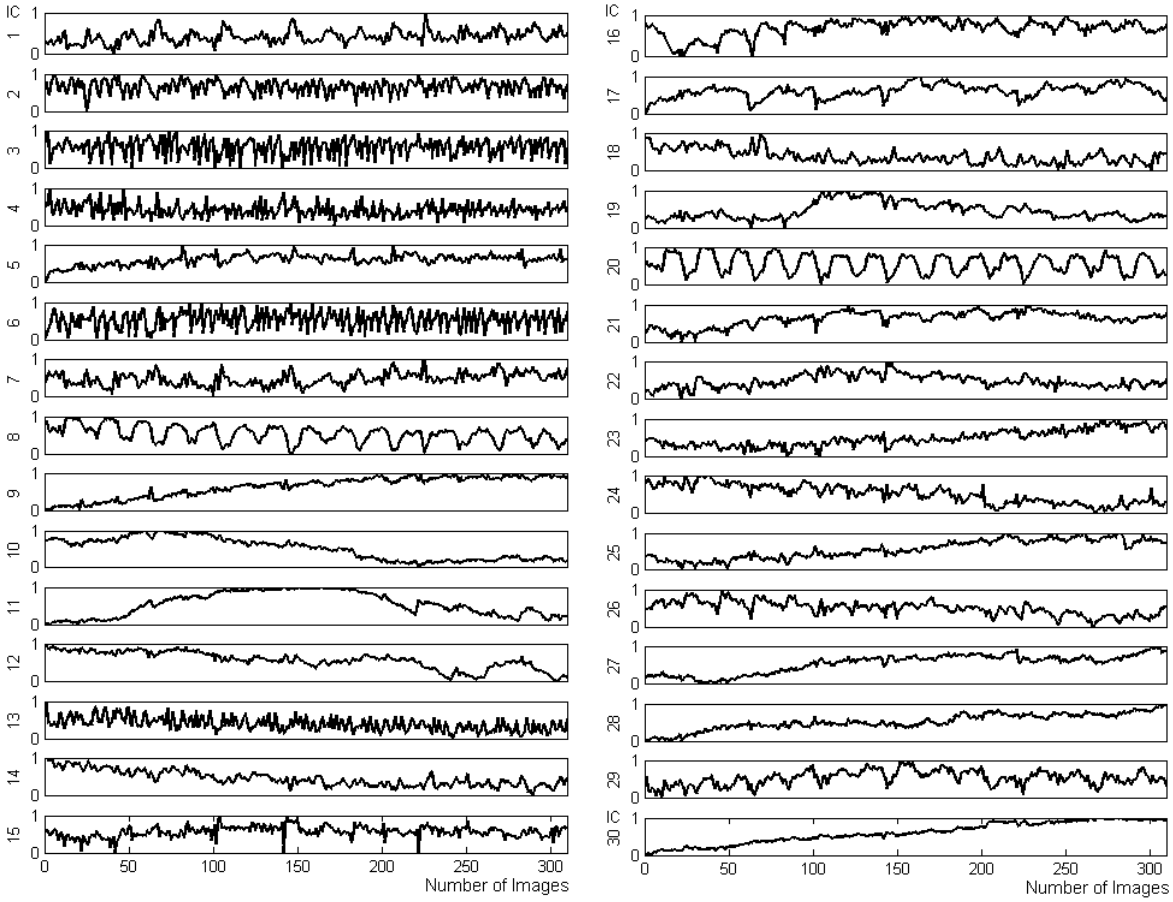

Figure 7.7: 30 independent component time courses of subj. 3 (1. session)

Table 7.1). The AC includes BA 41, BA 42, BA21, and BA 22, which is the Wernicke's area involved in speech comprehension and cognition. The somato-sensory area include BA 40. Motor-related areas include BA 06. Moreover, cognition-related areas include BA 09, BA 10, and BA 46. BA 44, BA 45, and BA 47 are about the Broca's area which is the motor speech centre and keeps information of working memory. All BAs were defined for the left and right hemisphere for all subjects and each session. Additionally, two clusters for the whole left and right AC were defined for each subject and each session.

In a following step the changes in the time course of these VOIs over the five repeated sessions were investigated. In general, there are changes of the signals over the five sessions, i.e. the signals are dynamic. These changes may depend on the subjects and on the areas and are described through different parameters. Figures 7.8 - 7.10 show exemplarily the time courses of three subjects of different VOIs/ACs of all five sessions. The time courses

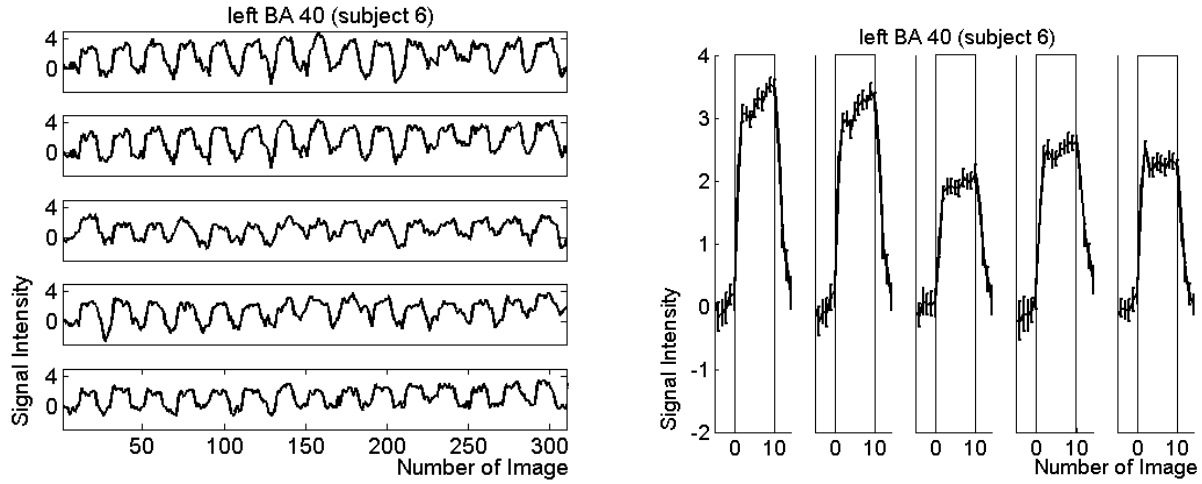

Figure 7.8: Time courses and event-related averages of left BA40 (subj. 6)

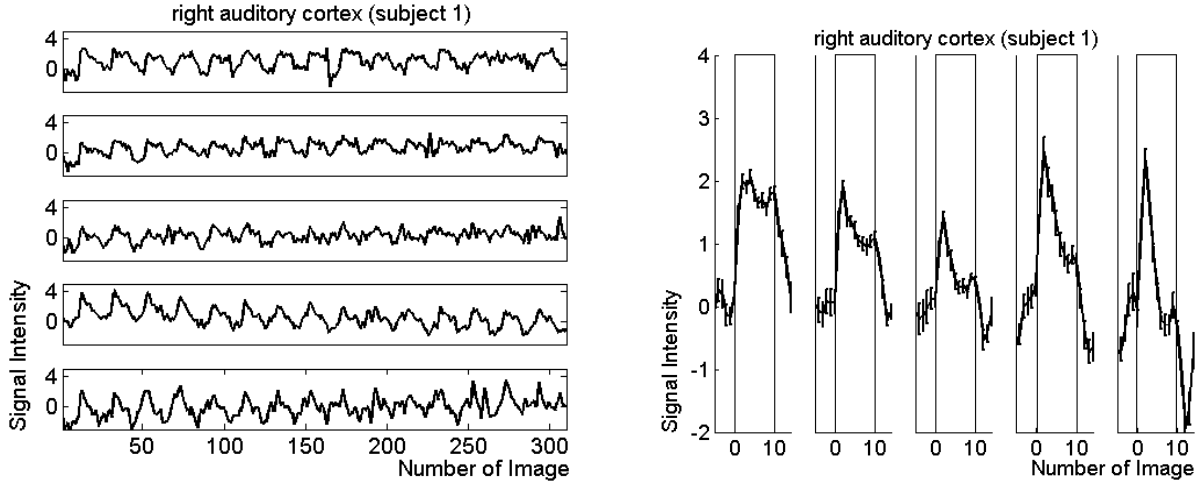

Figure 7.9: Time courses and event-related averages of the right auditory cortex (subj. 1)

are averaged across all voxels of the VOI (left side, sessions displayed from top to bottom). The event-related average (right side, sessions displayed from left to right) displays the percentage signal change averaged over all measured response blocks (the rectangle indicates the stimulation period). Each subplot display the time course for one of the five repeated sessions.

The time courses of the five sessions of subject 6 in the left BA 40 as well as in other areas were characterized by trend increases within the stimulation blocks, see Figure 7.8. Moreover, the time courses are characterized by different signal amplitudes, namely a decrease in signal amplitudes over the repeated sessions. A very interesting finding of subject 1

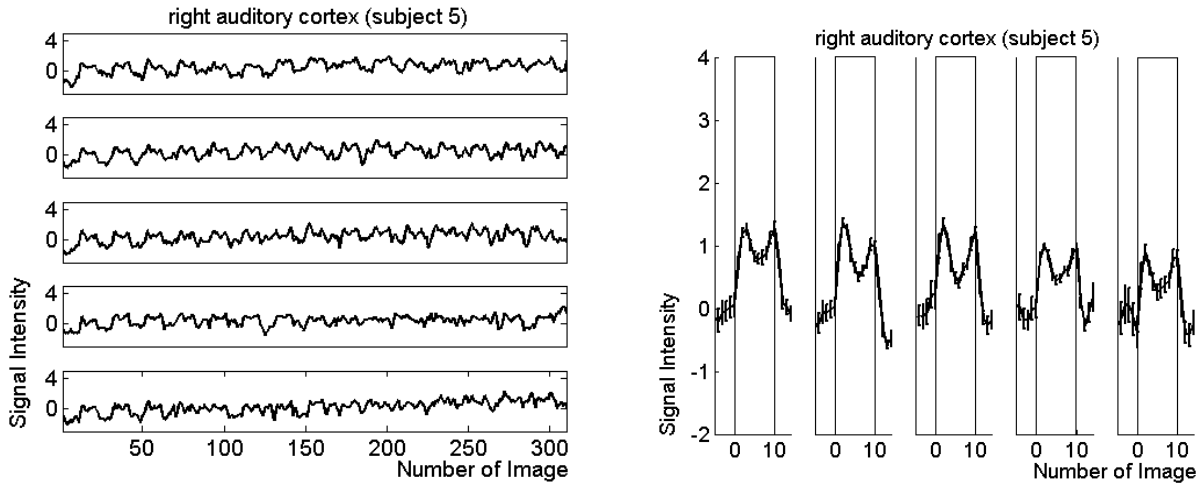

Figure 7.10: Time courses and event-related averages of left auditory cortex (subj. 5)

(Figure 7.9) was that the time courses of the clusters of the first session showed almost the typical/expected hemodynamic response function, i.e. the signal increased after stimulus onset, reached a plateau, and decreased slowly after stimulus offset. But the time courses of the last sessions showed a different behavior. The signal also increased after stimulus onset, mostly on a higher level than the signals in the first session, but the signal did not stay on the plateau, it decreased immediately, so that at the end of the stimulation the signal was already at the baseline. This was often found for the auditory regions and areas involved in maintenance and attention processing and might be explained by adaptation, habituation or learning effects. The time courses of subject 5 (Figure 7.10) do not differ a lot between the repeated sessions, but it was interesting that the time courses do not represent the typical course of the HRF. These time courses are characterized by a decrease within the stimulation block but compared to subject 1 the signal once again increases at the end of stimulation. These examples indicate very individual responses for the WM study.

These temporal changes of signal intensities in the left right AC *between the sessions* should be tested *for all subjects* with an ANOVA for repeated measurements, with session as repetition factor. Thereby for each session, a time average of all stimulation blocks for images 3-10 (considering the delayed onset of the signal) was computed for each subject for left and right AC. The results of ANOVA revealed no significant changes of the average signal intensity between repeated sessions over all subjects ($F_{1,4} = 2.183, p = 0.108$ for the left AC and $F_{1,4} = 1.218, p = 0.334$ for the right AC). This might be explained that the subjects show very different signals, see therefore Figures 7.8 - 7.10. But it is much more interesting if there are significant differences *for every single subject between repeated sessions.* There-

fore, in a second step an ANOVA was performed for each subject and the stimulation blocks served as repetition factor. The *time average for each stimulation block* for images 3-10 for each subject was computed. Table 7.2 summarizes the results of an ANOVA with repeated sessions for each subject for left and right AC. In contrary to the first ANOVA result, for

Table 7.2: Testing temporal signal changes between sessions for each subject.

subject	**left auditory cortex**	**right auditory cortex**
1	$F_{1,4} = 22.116, p < 0.001$	$F_{1,4} = 23.553, p < 0.001$
2	$F_{1,4} = 24.596, p < 0.001$	$F_{1,4} = 23.831, p < 0.001$
3	$F_{1,4} = 22.152, p < 0.001$	$F_{1,4} = 8.892, p < 0.001$
4	$F_{1,4} = 4.581, p = 0.023$	$F_{1,4} = 8.573, p < 0.001$
5	$F_{1,4} = 46.869, p < 0.001$	$F_{1,4} = 8.689, p < 0.001$
6	$F_{1,4} = 1.448, p = 0.252$	$F_{1,4} = 5.089, p = 0.001$

almost every subject for the left AC and the right AC there were significant difference of the average signal change between repeated sessions (except subject 6, left AC), indicating signal changes over repeated sessions.

7.4 Comparing ICA Time Courses to HRF Time Courses in Correlation Analysis

Since we found out that the fMRI signal was very dynamic over the five repeated sessions, the results of the classical correlation analysis should be compared to ICA. The correlation analysis always assumes the same hypothetical time course of neuronal response (i.e. stimulation protocol) and neglects dynamics of the signals. When we used the time courses of the fifth session of all subjects for the left AC and the right AC, which is often very different from the signal of the stimulation protocol, it should be controlled whether the assumption of the HRF was still valid in the fifth session. Two correlation analysis were performed for the fifth session of each subject: 1. correlating the data set to the hypothetical HRF signal and 2. correlating the data set to the time course obtained of AC obtained by ICA. The method of correlation analysis was used in order to have the same p-value for the activation maps, i.e. that it is possible to compare ICA to classical correlation analysis. Figure 7.11 shows activation of the left AC and the right AC of subject 1 of correlation analysis with HRF (light gray) and independent component time course (dark gray). The statistical

maps were drawn at same p-values for both analyses ($p \leq 8 * 10^{-6}$). As could be seen, with the ICA approach much more activation can be detected compared with correlation with HRF. This analysis was performed for each subject with its time course of the independent component of the fifth session. The number of activated voxels for both analyses was counted for each subject and each hemisphere. This result is summarized for all subjects in Table 7.3.

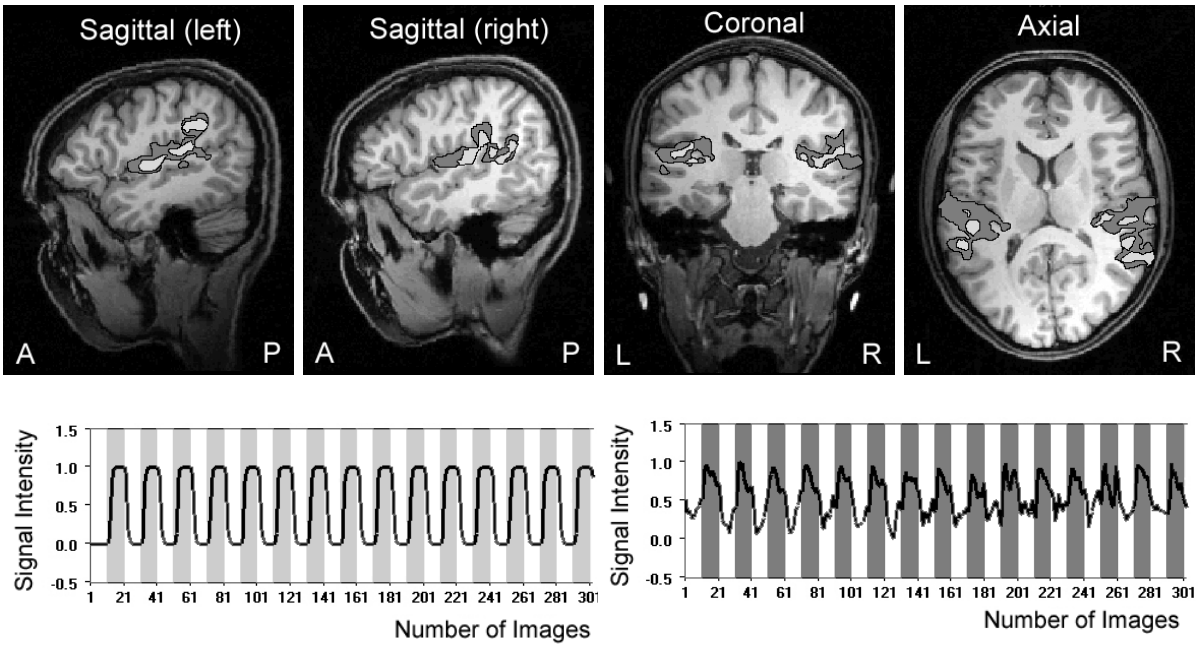

Figure 7.11: Comparing results of correlation analysis with HRF (light gray) and independent component time course (dark gray) (subj.1)

Table 7.3 shows great variability between the subjects. But it is obvious that more voxels are detected with ICA results than with HRF results (see especially subject 1 and subject 5). Since the ICA signal was already adapted to the data set, the correlation analysis with ICA signal is quantitatively better than correlation with HRF signal. The difference between the number of voxels detected by ICA and the number of voxels detected by HRF was further tested by t-test for mean=0. In the left AC the t-test revealed $p = 0.0248$ and in the right AC the t-test revealed $p = 0.0206$. Consequently, there where significant differences between the number of voxels detected by ICA and the number of voxels detected by HRF.

Table 7.3: Comparing the number of voxels of correlation analysis with hemodynamic response function and independent component time course of fifth session

subj.	**left AC**				**right AC**			
	HRF	ICA	% HRF of ICA	Difference ICA-HRF	HRF	ICA	% HRF of ICA	Difference ICA-HRF
1	2638	14157	18.6	11519	3915	13987	28.0	10072
2	19494	21772	89.5	2278	17375	19338	89.8	1963
3	19943	21218	94.0	1275	17558	19418	90.4	1860
4	14512	20314	71.4	5802	13595	20156	67.4	6561
5	5167	15092	34.2	9925	3283	11595	28.3	8312
6	18908	21506	87.9	2598	20138	21800	92.4	1662

7.5 Discussing the Shape of BOLD Responses

In this last section the shape of the BOLD responses should be discussed to support the findings in our fMRI study. There are several publications discussing the shape of BOLD responses. Seifritz et al., 2002 found out that neuronal responses in the AC can be decomposed temporally into independent transient and sustained activity with ICA in different parts of AC. Transient activity is characterized by a peak in the signal, whereas, sustained activity remains on a constant level during the stimulation period. Transient responses typically occur at the onset of a stimulus whereas sustained responses follow the stimulus. Seifritz et al., 2003 discussed that the blood oxygen level-dependent (BOLD) signal time course in AC is characterized by two components, an initial transient peak and a subsequent sustained plateau with smaller amplitude. (It is unclear whether the reduction of amplitude during the sustained period represents underlying neuronal activity-related changes in oxygenation or results from other hemodynamic mechanisms. In studies of Harms and Melcher, 2002, Harms and Melcher, 2003, Harms et al., 2005 they found out that in human auditory cortex, prolonged sound stimuli (30 *sec*) can evoke responses ranging from sustained to highly phasic (i.e. characterized by prominent peaks just after sound onset and offset.) Prolonged (30 *sec*) low-rate (2/*sec*; each stimuli is perceptually distinct) noise elicit sustained responses whereas high-rate (35/*sec*; individual stimuli are not distinguishable) elicit phasic responses with peaks just after train onset and offset. They used the general linear model using a set of basis functions chosen to reflect temporal features of cortical fMRI responses. (Five basis functions were chosen: onset,

sustained ramp, offset, and undershoot for the two response waveshapes, sustained and phasic.) And finally D'Avossa et al., 2003 classified regional BOLD responses with PCA in sustained, transient, and negative time courses.

There are temporal changes in the shape of BOLD responses but in all the studies mentioned above there was no learning task as in our study. In our study we found dynamic changes between and within repeated sessions that might be due to learning related processes. Theses dynamic changes are indicated through signal amplitude changes or trends within stimulation blocks. The study showed that almost the same areas are involved in solving the auditory task for all subjects, namely auditory areas, motion-related areas and cognition-related areas. But these areas show differences in their time courses.

In our study we investigated as small group of subjects, namely 6 subjects with five repeated sessions, these subjects showed very different behaviors in their time courses which could not be predicted in advance. Therefore, a correlation analysis or GLM approach with a static HRF function is not always the best approach to get task-related activations since the GLM needs previous information of contributing signals. In learning-related studies theses information is in general not known in advance. The ICA is a method that do not need previous information of contributing signals to find activation cluster related to the task.

8 Conclusions

Blind source separation by independent component analysis has received attention because of its potential applications in signal processing such as is speech recognition systems, telecommunications and medical signal processing.

In fMRI the signal or signals of interest, in our case the neuronal response of the subjects is seldom recorded in isolation, and is generally mixed with other ongoing background activity and sensor noise, and is almost certainly contaminated by artifacts of either physiological or environmental origins. Furthermore, the signal-to-noise ratio of the desired signal is generally quite poor.

ICA is a technique to estimate statistically independent components from their linear mixtures. Most ICA algorithms are derived by forming a linear demixing model, defining a measure of statistical independence and performing numerical optimization of the independence measure based on the given observations. In this framework, each component is treated as a random variable and the independence measure used by ICA algorithm is a statistical measure such as the higher order statistics, negentropy and information.

As the focus is in particular on fMRI time series additional temporal statistics can be used to identify the unknown sources. One can exploit classical time series decompositions, non-stationarity, temporal correlations and frequency analysis of the time series.

In simulation studies the performance of different ICA algorithms was tested by modelling, mixing, and demixing different signals. ICA was applied to an auditory fMRI study with repeated sessions to investigate if there are dynamic changes in the measured fMRI signals. The analysis revealed that the subjects activated almost the same brain regions but showing different time courses and with different dynamic changes over the five repeated sessions.

The advantage of ICA is that no assumptions of contributing signals have to be made in advance. ICA can separate different signals into independent components and can detect dynamic changes in the signals that might be due to learning performances of the subjects.

Bibliography

[Amari et al., 1996] Amari, S., Cichocki, A., and Yang, H. (1996). *A New Learning Algorithm for Blind Signal Separation.* Advances in Neural Information Processing Systems, volume 8: pp. 757–763.

[Andersen et al., 1999] Andersen, A., Gash, D. M., and Avison, M. J. (1999). *Principal Component Analysis of the Dynamic Response Measured by fMRI: A Generalized Linear Systems Framework.* Magnetic Resonance Imaging, volume 17 (6): pp. 795–815.

[Baddeley, 1992] Baddeley, A. (1992). *Working Memory.* Science, volume 255 (5044): pp. 556–559.

[Bandettini et al., 1993] Bandettini, P., Jesmanowicz, A., Wong, E. C., and Hyde, J. S. (1993). *Processing Strategies for Time-Course Data Sets in Functional MRI of the Human Brain.* Magnetic Resonance in Medicine, volume 30 (2): pp. 161–173.

[Barch et al., 1997] Barch, D., Braver, T., Nystrom, L., Forman, S., Noll, D., and Cohen, J. (1997). *Dissociating working memory from task difficulty in human prefrontal cortex.* Neuropsychologia, volume 35 (10): pp. 1373–1380.

[Baumgartner et al., 2000] Baumgartner, R., Ryner, L., Richter, W., Summers, R., Jarmasz, M., and Somorjai, R. (2000). *Comparison of Two Exploratory Data Analysis Methods for fMRI: Fuzzy Clustering vs. Principal Component Analysis.* Magnetic Resonance Imaging, volume 18 (1): pp. 89–94.

[Beckmann and Smith, 2004] Beckmann, C. and Smith, S. (2004). *Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging.* IEEE Transactions on Medical Imaging, volume 23 (2): pp. 137–152.

[Bell and Sejnowski, 1995] Bell, A. and Sejnowski, T. (1995). *An Information-Maximization Approach to Blind Separation and Blind Deconvolution.* Neural Computation, volume 7 (6): pp. 1129–1159.

[Beyer et al., 1988] Beyer, O., Girlich, H.-J., and Zschiesche, H.-U. (1988). *Stochastische Prozesse und Modelle.* Teubner, Leipzig.

[Biswal and Ulmer, 1999] Biswal, B. and Ulmer, J. (1999). *Blind Source Separation of Multiple Signal Sources of fMRI Data Sets using Independent Component Analysis.* Journal of Computer Assisted Tomography, volume 23 (2): pp. 265–271.

[Bortz et al., 2000] Bortz, J., Lienert, G. A., and Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik.* Springer, Berlin.

[Boscolo et al., 2004] Boscolo, R., Pan, H., and Roychowdhury, V. (2004). *Independent Component Analysis based on Nonparamertic Density Estimation.* IEEE Transactions on Neural Networks, volume 15 (1): pp. 55–65.

[Bradley, 1968] Bradley (1968). *Distribution-Free Statistical Tests.* Prentice Hall, Englewood Cliffs.

[Braver et al., 1997] Braver, T., Cohen, J., Nystrom, L., Jonides, J., Smith, E., and Noll, D. (1997). *A parametric study of prefrontal cortex involvement in human working memory.* Neuroimage, volume 5 (1): pp. 49–62.

[Brechmann et al., 2007] Brechmann, A., Gaschler-Markefski, B., Sohr, M., Yoneda, K., Kaulisch, T., and Scheich, H. (2007). *Working Memory Specific Activity in Auditory Cortex: Potential Correlates of Sequential Processing and Maintenance.* Cerebral Cortex, published online 04.01.2007.

[Brechmann and Scheich, 2005] Brechmann, A. and Scheich, H. (2005). *Hemispheric Shifts of Sound Representation in Auditory Cortex with Conceptual Listening.* Cerebral Cortex, volume 15 (5): pp. 578–587.

[Brodmann, 1909] Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde.* Johann Ambrosius Barth, Leipzig.

[Brown et al., 2001] Brown, G., Yamada, S., and Sejnowski, T. J. (2001). *Independent Component Analysis at the Neural Cocktail Party.* Trends in Neurosciences, volume 24 (1): pp. 54–63.

[Bullmore et al., 1996] Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. (1996). *Statistical Methods of Estimation and Inference for Functional MR Image Analysis*. Magnetic Resonance in Medicine, volume 35 (2): pp. 261–277.

[Calhoun et al., 2003] Calhoun, V., Adali, T., Hansen, L. K., Larsson, J., and Pekar, J. J. (2003). *ICA of Functional MRI Data: An Overview*. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*. pp. 281 – 288.

[Calhoun et al., 2001] Calhoun, V., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). *Spatial and Temporal Independent Component Analysis of Functional MRI Data Containing a Pair of Task-Related Waveforms*. Human Brain Mapping, volume 13 (1): pp. 43–53.

[Calhoun et al., 2005] Calhoun, V. D., Adali, T., Stevens, M. C., Kiehl, K. A., and Pekar, J. J. (2005). *Semi-Blind ICA of fMRI: A Method for Utilizing Hypothesis-Derived Time Courses in a Spatial ICA Analysis*. Neuroimage, volume 25 (2): pp. 527–538.

[Cardoso, 1997] Cardoso, J. F. (1997). *Infomax and Maximum Likelihood for Blind Separation*. IEEE Signal Processing Letters, volume 4 (4): pp. 112–114.

[Cattell, 1966] Cattell, R. B. (1966). *The scree test for the number of factors*. Multivariate Behavioral Research, volume 1: pp. 629–637.

[Chakravarti et al., 1967] Chakravarti, I., Laha, R., and Roy, J. (1967). *Handbook of Methods of Applied Statistics*. John Wiley and Sons, Chichester.

[Cichocki, 2003] Cichocki, A. (2003). *Blind Source Separation Algorithms with Matrix Constraints*. IEICE Trans. Fundamentals.

[Comon, 1994] Comon, P. (1994). *Independent Component Analysis. A New Concept?* Signal Processing, volume 36: pp. 287–314.

[Cover and Thomas, 1991] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications, New York.

[d'Avossa et al., 2003] d'Avossa, G., Shulman, G. L., and Corbetta, M. (2003). *Identification of Cerebral Networks by Classification of the Shape of BOLD Responses*. Journal of Neurophysiology, volume 90 (1): pp. 360–371.

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

[Diggle, 1995] Diggle, P. J. (1995). *Time Series: A Biostatistical Introduction.* Clarendon Press, Oxford.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

[Esposito et al., 2002] Esposito, F., Formisano, E., Seifritz, E., Goebel, R., Morrone, R., Tedeschi, G., and Di Salle, F. (2002). *Spatial Independent Component Analysis of Functional MRI Time-Series: To What Extent do Results Depend on the Algorithm used?* Human Brain Mapping, volume 16 (3): pp. 146–157.

[Esposito et al., 2003] Esposito, F., Seifritz, E., Formisano, E., Morrone, R., Scarabino, T., Tedeschi, G., Cirillo, S., Goebel, R., and Di Salle, F. (2003). *Real-Time Independent Component Analysis of fMRI Time-Series.* Neuroimage, volume 20 (4): pp. 2209–2224.

[Filzmoser et al., 1999] Filzmoser, P., Baumgartner, R., and Moser, E. (1999). *A Hierarchical Clustering Method for Analyzing Functional MR Images.* Magnetic Resonance Imaging, volume 17 (6): pp. 817–826.

[Fisz, 1980] Fisz, M. (1980). *Wahrscheinlichkeitsrechnung und Mathematische Statistik.* VEB Deutscher Verlag der Wissenschaften, Berlin.

[Fletcher et al., 1996] Fletcher, P. C., Dolan, R. J., Shallice, T., Frith, C. D., Frackowiak, R. S., and Friston, K. J. (1996). *Is Multivariate Analysis of PET Data more Revealing than the Univariate Approach? Evidence from a Study of Episodic Memory Retrieval.* Neuroimage, volume 3 (3 Pt 1): pp. 209–215.

[Formi sano et al., 2004] Formi sano, E., Esposito, F., Di Salle, F., and Goebel, R. (2004). *Cortex-based Independent Component Analysis of fMRI Time-Series.* Magnetic Resonance Imaging, volume 22: pp. 1493–1504.

[Formisano et al., 2002] Formisano, E., Esposito, F., Kriegeskorte, N., Tedeschi, G., Di Salle, F., and Goebel, R. (2002). *Spatial Independent Component Analysis of Functional Magnetic Resonance Imaging Time-Series: Characterization of the Cortical Components.* Neurocomputing, volume 49: pp. 241–254.

[Friston, 1996] Friston, K. (1996). *Statistical Parametric Mapping and other Analyses of Functional Imaging Data, in: Brain Mapping: The Methods.* Academic Press, San Diego.

[Friston et al., 1993] Friston, K., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J. (1993). *Functional Connectivity: the Principal Component Analysis of Large (PET) Data Sets.* Journal of Cerebral Blood Flow and Metabolism, volume 13 (1): pp. 5–14.

[Friston et al., 1995] Friston, K., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S. J., and Turner, R. (1995). *Analysis of fMRI Time-Series Revisited.* Neuroimage, volume 2 (1): pp. 45–53.

[Friston et al., 1991] Friston, K. J., Jezzard, P., Frackowiak, R. S. J., and Turner, R. (1991). *Comparing Functional (PET) Images: The Assessment of Significant Change.* Journal of Cerebral Blood Flow Metabolism, volume 13: pp. 5–14.

[Gaschler-Markefski et al., 2003] Gaschler-Markefski, B., Yoneda, K., Kaulisch, T., Brechmann, A., and Scheich, H. (2003). *FMRI Activation of Left Planum Temporale Predicts Task Performsnce in an Auditory Working Memory Task.* In *Proceedings of the International Conference on Auditory Cortex 2003 - Towards a Synthesis of Human and Animal Research.* Shaker Verlage, Aachen, Magdeburg, Germany, p. 22.

[Gössl et al., 2001] Gössl, C., Fahrmeir, L., and Auer, D. P. (2001). *Bayesian Modeling of the Hemodynamic Response Function in BOLD fMRI.* Neuroimage, volume 14 (1): pp. 140–148.

[Goutte et al., 1999] Goutte, C., Toft, P., Rostrup, E., Nielsen, F., and Hansen, L. K. (1999). *On Clustering fMRI Time Series.* Neuroimage, volume 9 (3): pp. 298–310.

[Green and Swets, 1966] Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* John Wiley and Sons, New York.

[Harms et al., 2005] Harms, M. P., Guinan, J., J. J., Sigalovsky, I. S., and Melcher, J. R. (2005). *Short-term Sound Temporal Envelope Characteristics Determine Multisecond Time Patterns of Activity in Human Auditory Cortex as Shown by fMRI.* Journal of Neurophysiology, volume 93 (1): pp. 210–222.

[Harms and Melcher, 2002] Harms, M. P. and Melcher, J. R. (2002). *Sound Repetition Rate in the Human Auditory Pathway: Representations in the Waveshape and Amplitude of fMRI Activation.* Journal of Neurophysiology, volume 88 (3): pp. 1433–1450.

[Harms and Melcher, 2003] Harms, M. P. and Melcher, J. R. (2003). *Detection and Quantification of a Wide Range of fMRI Temporal Responses using a Physiologically-Motivated Basis Set.* Human Brain Mapping, volume 20 (3): pp. 168–183.

[Hartung and Elpelt, 1995] Hartung, J. and Elpelt, B. (1995). *Multivariate Statistik.* Oldenbourg Verlag, München.

[Heeger and Ress, 2002] Heeger, D. J. and Ress, D. (2002). *What Does fMRI Tell us about Neuronal Activity?* Nature Review Neuroscience, volume 3 (2): pp. 142–151.

[Himberg et al., 2004] Himberg, J., Hyvärinen, A., and Esposito, F. (2004). *Validating the Independent Components of Neuroimaging Time Series via Clustering and Visualization.* Neuroimage, volume 22: pp. 1214–1222.

[Hotelling, 1936] Hotelling, H. (1936). *Relations between two sets of variates.* Biometrika, volume 28: pp. 321–377.

[Hu et al., 2005] Hu, D., Yan, L., Yadong, L., Zhou, Z., Friston, K. J., Tan, C., and Wu, D. (2005). *Unified SPM-ICA for fMRI Analysis.* Neuroimage, volume 25: pp. 746–755.

[Hyvärinen, 1999*a*] Hyvärinen, A. (1999*a*). *Fast and Robust Fixed-Point Algorithms for Independent Component Analysis.* IEEE Transactions on Neural Networks, volume 10: pp. 626–634.

[Hyvärinen, 1999*b*] Hyvärinen, A. (1999*b*). *Independent Component Analysis: A Tutorial.*

[Hyvärinen, 1999*c*] Hyvärinen, A. (1999*c*). *Survey on Independent Component Analysis.* Neural Computing Surveys, volume 2: pp. 94 –128.

[Hyvärinen, 2001] Hyvärinen, A. (2001). *Complexity Pursuit: Separating Interesting Components from Time Series.* Neural Computation, volume 13 (4): pp. 883–898.

[Hyvärinen et al., 2001*a*] Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001*a*). *Topographic Independent Component Analysis.* Neural Computation, volume 13: pp. 1527–1558.

[Hyvärinen et al., 2001*b*] Hyvärinen, A., Karhunen, J., and Oja, E. (2001*b*). *Independent Component Analysis*. A Volume in the Wiley Series on Adaptive Learning Systems for Signal Processing, Communications, and Control.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). *Independent Component Analysis: Algorithms and Applications*. Neural Networks, volume 13 (4-5): pp. 411–430.

[Jackson, 1991] Jackson, J. (1991). *A User's Guide to Principal Components*. N.Y. Wiley and Sons, New York.

[Jezzard, 2001] Jezzard, P. (2001). *Functional MRI: An Introduction to Methods*. Oxford University Press.

[Jonides, 2004] Jonides, J. (2004). *How does practice makes perfect?* Nature Neuroscience, volume 7: pp. 10 – 11.

[Jung et al., 2000] Jung, T. P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., and Sejnowski, T. J. (2000). *Removing Electroencephalographic Artifacts by Blind Source Separation*. Psychophysiology, volume 37 (2): pp. 163–178.

[Jung et al., 2001] Jung, T. P., Makeig, S., McKeown, M. J., Bell, T. L., and Sejnowski, T. J. (2001). *Imaging Brain Dynamics using Independent Component Analysis*. Proceedings of the IEEE, volume 89 (7): pp. 1107–1122.

[Jutten and Hérault, 1991] Jutten, C. and Hérault, J. (1991). *Blind Separation of Sources, Part I: An Adaptive Algorithm based on Neuromimetic Architecture*. Signal Processing, volume 24: pp. 1–10.

[Kaiser, 1960] Kaiser, H. F. (1960). *The application of electronic computers to factor analysis*. Education and Psychological Measurement, volume 20: pp. 141–151.

[Kiviniemi et al., 2003] Kiviniemi, V., Kantola, J. H., Jauhiainen, J., A., H., and Tervonen, O. (2003). *Independent Component Analysis of Nondeterministic fMRI Signal Sources*. Neuroimage, volume 19: pp. 253–260.

[Kullback, 1959] Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[Lange, 1996] Lange, N. (1996). *Statistical Approaches to Human Brain Mapping by Functional Magnetic Resonance Imaging.* Statistics in Medicine, volume 15 (4): pp. 389–428.

[Lange et al., 1999] Lange, N., Strother, S. C., Anderson, J. R., Nielsen, F. A., Holmes, A. P., Kolenda, T., Savoy, R., and Hansen, L. K. (1999). *Plurality and Resemblance in fMRI Data Analysis.* Neuroimage, volume 10 (3): pp. 282–303.

[Lee et al., 1999*a*] Lee, T., Girolami, M., and Sejnowski, T. J. (1999*a*). *Independent Component Analysis using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources.* Neural Computation, volume 11 (2): pp. 417–441.

[Lee et al., 1999*b*] Lee, T., Lewicki, M. S., Girolami, M., and Sejnowski, T. J. (1999*b*). *Blind Source Separation of More Sources than Mixtures using Overcomplete Representations.* IEEE Signal Processing Letters, volume 6 (4): pp. 87–90.

[Liao et al., 2006] Liao, R., McKeown, M., and J.L., K. (2006). *Isolation and Minimization of Head Motion-induced Signal Variations in fMRI Data using Independent Component Analysis.* Magnetic Resonance in Medicine, volume 55 (6): pp. 1396–1413.

[Lin et al., 2003] Lin, F., McIntosh, A. R., Agnew, J. A., Eden, G. F., Zeffiro, T. A., and Belliveau, J. W. (2003). *Multivariate Analysis of Neuronal Interactions in the Generalized Partial Least Squares Framework: Simulations and Empirical Studies.* Neuroimage, volume 20 (2): pp. 625–642.

[MacKay, 1996] MacKay, D. (1996). *Maximum Likelihood and Covariant Algorithms for Independent Component Analysis.* Technical Report, Cambridge University, Cavendish Laboratory.

[Makeig et al., 1997] Makeig, S., Jung, T. P., Bell, A. J., Ghahremani, D., and Sejnowski, T. J. (1997). *Blind Separation of Auditory Event-Related Brain Responses into Independent Components.* Proceedings of the National Academy of Sciences USA, volume 94 (20): pp. 10979–10984.

[McKeown, 2000] McKeown, M. (2000). *Detection of Consistently Task-Related Activations in fMRI Data with Hybrid Independent Component Analysis.* Neuroimage, volume 11 (1): pp. 24–35.

[McKeown et al., 2003] McKeown, M., Hansen, L. K., and Sejnowsk, T. J. (2003). *Independent Component Analysis of Functional MRI: What is Signal and What is Noise?* Current Opinion In Neurobiology, volume 13 (5): pp. 620–629.

[McKeown et al., 1998*a*] McKeown, M., Jung, T. P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T. W., and Sejnowski, T. J. (1998*a*). *Spatially Independent Activity Patterns in Functional MRI Data during the Stroop Color-Naming Task*. Proceedings of the National Academy of Sciences USA, volume 95 (3): pp. 803–810.

[McKeown et al., 1998*b*] McKeown, M., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. (1998*b*). *Analysis of fMRI Data by Blind Separation into Spatial Independent Components*. Human Brain Mapping, volume 6 (3): pp. 160–188.

[McKeown and Sejnowski, 1998] McKeown, M. and Sejnowski, T. (1998). *Independent Component Analysis of fMRI Data: Examining the Assumptions*. Human Brain Mapping, volume 6 (5-6): pp. 368–372.

[Meinecke et al., 2002] Meinecke, F., Ziehe, A., Kawanabe, M., and Muller, K. R. (2002). *A Resampling Approach to Estimate the Stability of One-dimensional or Multidimensional Independent Ccomponents*. IEEE Transactions on Biomedical Engeneering, volume 49 (12.2): pp. 1514–1525.

[Molgedey and Schuster, 1994] Molgedey, L. and Schuster, H. (1994). *Separation of a Mixture of Independent Signals using Time Delayed Correlations*. Physical Review Letters, volume 72 (23): pp. 3634–3637.

[Moritz et al., 2003] Moritz, C. H., Rogers, B. P., and Meyerand, M. E. (2003). *Power Spectrum Ranked Independent Component Analysis of a Periodic fMRI Complex Motor Paradigm*. Human Brain Mapping, volume 18 (2): pp. 111–22.

[Mutihac and van Hulle, 2004] Mutihac, R. and van Hulle, M. M. (2004). *Comparison of Principal Component Analysis and Independent Component Analysis for Blind Source Separation*. Romanian Reports in Physics, volume 56 (1): pp. 20–32.

[Newbold, 1995] Newbold, P. (1995). *Statistics for Business and Economics*. Prentice-Hall International, Inc., University of Illinois.

[Ogawa et al., 1990] Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. (1990). *Brain Magnetic Resonance Imaging with Contrast Dependent on Blood Oxygenation*. Proceedings of the National Academy of Sciences of the United States of America, volume 87 (24): pp. 9868–9872.

[Ohl and Scheich, 2005] Ohl, F. and Scheich, H. (2005). *Learning-induced Plasticity in Animal and Human Auditory Cortex*. Current Opinion in Neurobiology, volume 15: pp. 470–477.

[Oja, 1998] Oja, E. (1998). *From Neural Learning to Independent Components*. Neurocomputing, volume 22: pp. 187–199.

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

[Pham et al., 1992] Pham, D. T., Garat, P., and Jutten, C. (1992). *Separation of a Mixture of Independent Sources through a Maximum Likeihood Approach*. presented at the Proc. EUSIPCO.

[Quigley et al., 2002] Quigley, M., Haughton, V. M., Carew, J., Cordes, D., Moritz, C. H., and Meyerand, M. E. (2002). *Comparison of Independent Component Analysis and Conventional Hypothesis-Driven Analysis for Clinical Functional MR Image Processing*. AJNR American Journal of Neuroradiology, volume 23 (1): pp. 49–58.

[Rao, 1964] Rao, C. (1964). *The Use and Interpretation of Principal Component Analysis in Applied Research*. Sankya, volume 26: pp. 329–358.

[Sachs, 1999] Sachs, L. (1999). *Angewandte Statistik*. Springer Verlag, Berlin.

[Samarov and Tsybakov, 2004] Samarov, A. and Tsybakov, A. (2004). *Nonparametric Independent Component Analysis*. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability, volume 10 (4): pp. 565–582.

[Schlittgen and Streitberg, 1995] Schlittgen, R. and Streitberg, B. (1995). *Zeitreihenanalyse*. Oldenbourg, München.

[Seifritz et al., 2003] Seifritz, E., Di Salle, F., Esposito, F., Bilecen, D., Neuhoff, J. G., and Scheffler, K. (2003). *Sustained blood oxygenation and volume response to repetition rate-modulated sound in human auditory cortex*. Neuroimage, volume 20 (2): pp. 1365–1370.

[Seifritz et al., 2002] Seifritz, E., Esposito, F., Hennel, F., Mustovic, H., Neuhoff, J. G., Bilecen, D., Tedeschi, G., Scheffler, K., and Di Salle, F. (2002). *Spatiotemporal Pattern of Neural Processing in the Human Auditory Cortex*. Science, volume 297 (5587): pp. 1706–1708.

[Shapiro and Wilks, 1983] Shapiro, S. and Wilks, M. (1983). *An Analysis of Variance for Normality (Complete Samples).* Biometrika, volume 52: pp. 591–611.

[Silverman, 1986] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York.

[Sohr et al., 2003] Sohr, M., Gaschler-Markefski, B., and Scheich, H. (2003). *Success-Related Effects of an Auditory Working Memory Task on Single-Subject fMRI Activation of Auditory Cortex.* In *Proceedings of the International Conference on Auditory Cortex 2003 - Towards a Synthesis of Human and Animal Research.* Shaker Verlage, Aachen, Magdeburg, Germany, p. 46.

[Stone, 1999*a*] Stone, J. (1999*a*). *Spatial, Temporal, and Spatiotemporal Independent Component Analysis of fMRI Data.* In *Conference Proceedings of Spatial-temporal Modelling and its applications.*

[Stone, 1999*b*] Stone, J. V. (1999*b*). *Independent Component Analysis: An Introduction.* Trends in Cognitive Sciences, volume 6 (2): pp. 59–64.

[Stone et al., 2002] Stone, J. V., Porrill, J., Porter, N. R., and Wilkinson, I. D. (2002). *Spatiotemporal Independent Component Analysis of Event-Related fMRI Data using Skewed Probability Density Functions.* Neuroimage, volume 15 (2): pp. 407–421.

[Stuart and Ord, 1994] Stuart, A. and Ord, J. (1994). *Kendall's Advanced Theory of Statistics*, volume 1, Distribution Theory. John Wiley and Sons Inc., New York.

[Talairach and Tournoux, 1988] Talairach, J. and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Bain.* Georg Thieme Verlag, Stuttgart, New York.

[Thomas et al., 2002] Thomas, C., Harshman, R. A., and Menon, R. S. (2002). *Noise Reduction in BOLD-Based fMRI using Component Analysis.* Neuroimage, volume 17 (3): pp. 1521–1537.

[Turner and Donald, 2005] Turner, G. and Donald, D. (2005). *Study of Temporal Stationarity and Spatial Consistency of fMRI Noise using Independent Component Analysis.* IEEE Transactions on Medical Imaging, volume 24 (6): pp. 712–718.

[van de Ven et al., 2004] van de Ven, V. G., Formisano, E., Prvulovic, D., Roeder, C. H., and Linden, D. E. (2004). *Functional Connectivity as Revealed by*

Spatial Independent Component Analysis of fMRI Measurements during Rest. Human Brain Mapping, volume 22 (3): pp. 165–178.

[Vigario et al., 2000] Vigario, R., Sarela, J., Jousmaki, V., Hamalainen, M., and Oja, E. (2000). *Independent Component Approach to the Analysis of EEG and MEG Recordings*. IEEE Transactions in Biomedical Engeneering, volume 47 (5): pp. 589–593.

[Worsley and Friston, 1995] Worsley, K. and Friston, K. (1995). *Analysis of fMRI Time-Series Revisited – Again*. Neuroimage, volume 2 (3): pp. 173–181.

A Properties of Information-Theoretic Functions

This Appendix summarizes some properties of information-theoretic functions like information, differential entropy, and negentropy as introduced in Section 2.3.1.

Remark, the notation $\int f(x)\mathrm{d}x$ is the integral over the sample space $\mathbf{R}$, and $\int f(\mathbf{x})\mathrm{d}\mathbf{x}$ is the integral over $\mathbf{R}^N$ for the multivariate case.

A.1 Information

Theorem A.1 *The information* (see Equation 2.21) *is nonnegative,*

$$I(1:2) = I(f_1(\mathbf{x}) : f_2(\mathbf{x})) \geq 0, \tag{A.1}$$

with equality if and only if $f_1(\mathbf{x}) \equiv f_2(\mathbf{x})$*, where* $f_1(\mathbf{x})$ *is the joint probability density and* $f_2(\mathbf{x})$ *is the product of the marginal densities (i.e. the joint density under the assumption of independence).*

Proof: [Kullback, 1959 (p. 14-15)] Let $g(\mathbf{x}) = f_1(\mathbf{x})/f_2(\mathbf{x})$. Then

$$\begin{aligned} I(1:2) &= \int f_2(\mathbf{x})g(\mathbf{x})\log g(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int g(\mathbf{x})\log g(\mathbf{x})\mathrm{d}_{f_2}\mathbf{x}, \end{aligned}$$

with $\mathrm{d}_{f_2}(\mathbf{x}) = f_2(\mathbf{x})\mathrm{d}\mathbf{x}$. Substituting $t = g(\mathbf{x})$ and setting $q(t) = t\log t$, $t > 0$; since $0 < g(\mathbf{x}) < \infty$, one may write the Taylor expansion

$$q(g(\mathbf{x})) = q(1) + [g(\mathbf{x}) - 1]q'(1) + \frac{1}{2}[g(\mathbf{x}) - 1]^2 q''(h(\mathbf{x})), \tag{A.2}$$

where $h(\mathbf{x})$ lies between $g(\mathbf{x})$ and 1, so that $0 < h(\mathbf{x}) < \infty$, see Kullback, 1959. Since $q(1) = 0, q'(1) = 1.$ and

$$\int g(\mathbf{x})\mathrm{d}_{f_2}\mathbf{x} = \int f_1(\mathbf{x})\mathrm{d}\mathbf{x} = 1. \tag{A.3}$$

We find

$$\int q(g(\mathbf{x}))\mathrm{d}_{f_2}\mathbf{x} = \frac{1}{2}\int [g(\mathbf{x})-1]^2 q''(h(\mathbf{x}))\mathrm{d}_{f_2}\mathbf{x}, \tag{A.4}$$

where $q''(t) = 1/t > 0$ for $t > 0$. We see from (A.4) that

$$\int g(\mathbf{x})\log g(\mathbf{x})\mathrm{d}_{f_2}\mathbf{x} = \int f_1(\mathbf{x})\log\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\mathrm{d}\mathbf{x} \geq 0, \tag{A.5}$$

with equality if and only if $g(\mathbf{x}) \equiv f_1(\mathbf{x})/f_2(\mathbf{x}) = 1$. ■

Theorem A.2 *The information $I(1:2)$ is additive for independent random vectors $\mathbf{X}$ and $\mathbf{Y}$; that is, for the case that $\mathbf{X}$ and $\mathbf{Y}$ are independent from each other both under H_1 and H_2. The information $I(1:2)$ of two combined vectors $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $I(1:2,\mathbf{X},\mathbf{Y})$.*

$$I(1:2,\mathbf{X},\mathbf{Y}) = I(1:2,\mathbf{X}) + I(1:2,\mathbf{Y}). \tag{A.6}$$

Proof: [Kullback, 1959 p. 12-13]

$$\begin{aligned}
I(1:2,\mathbf{X},\mathbf{Y}) &= \int\int f_1(\mathbf{x},\mathbf{y})\log\frac{f_1(\mathbf{x},\mathbf{y})}{f_2(\mathbf{x},\mathbf{y})}\mathrm{d}(\mathbf{x},\mathbf{y}) \\
&\quad \text{(because of the independence,} \\
&\quad f_i(\mathbf{x},\mathbf{y}) = g_i(\mathbf{x})h_i(\mathbf{y}), \text{ and } \mathrm{d}(\mathbf{x},\mathbf{y}) = \mathrm{d}(\mathbf{x})\,\mathrm{d}(\mathbf{y})) \\
&= \int\int g_1(\mathbf{x})h_1(\mathbf{y})\log\frac{g_1(\mathbf{x})h_1(\mathbf{y})}{g_2(\mathbf{x})h_2(\mathbf{y})}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= \int\left(g_1(\mathbf{x})\log\frac{g_1(\mathbf{x})}{g_2(\mathbf{x})}\int h_1(\mathbf{y})\mathrm{d}\mathbf{y}\right)\mathrm{d}\mathbf{x} + \int\left(h_1(\mathbf{y})\log\frac{h_1(\mathbf{y})}{h_2(\mathbf{y})}\int g_1(\mathbf{x})\mathrm{d}\mathbf{x}\right)\mathrm{d}\mathbf{y}, \\
&\quad \text{where } \int g_1(\mathbf{x})\mathrm{d}\mathbf{x} = 1, \text{and } \int h_1(\mathbf{y})\mathrm{d}\mathbf{y} = 1 \\
&= \int g_1(\mathbf{x})\log\frac{g_1(\mathbf{x})}{g_2(\mathbf{x})}\mathrm{d}\mathbf{x} + \int h_1(\mathbf{y})\log\frac{h_1(\mathbf{y})}{h_2(\mathbf{y})}\mathrm{d}\mathbf{y} \\
&= I(1:2,\mathbf{X}) + I(1:2,\mathbf{Y}).
\end{aligned} \tag{A.7}$$

■

Additivity is the basis for the logarithmic form of information. A sample of N independent observations from the same population provides N times the information in a single observation.

A.2 Differential Entropy

Theorem A.3 *Translation does not change the differential entropy* (Equation 2.27) *of a random variable X with density function $f(x)$* [Cover and Thomas, 1991, p. 233.]

$$H(f(x+c)) = H(f(x)). \tag{A.8}$$

Proof: The proof follows directly from the definition of differential entropy. ■

Theorem A.4 *Multiplying the random variable X (with density function $f(x)$) with a constant factor a ($a \in \mathbf{R}$) gives the following differential entropy:*

$$H(f(ax)) = H(f(x)) + \log|a|, \tag{A.9}$$

where $|a|$ is the absolute value of a.

Proof: [Cover and Thomas, 1991, p. 233] Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and

$$\begin{aligned} H(f(ax)) &= -\int f_Y(y) \log f_Y(y)\, \mathrm{d}y \\ &= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) \mathrm{d}y \\ &= -\int f_X(x)(\log f_X(x) + \log|a|)\, \mathrm{d}x \\ &= H(f(x)) + \log|a|. \end{aligned} \tag{A.10}$$

■

Two further transformations of differential entropy will only be mentioned. The proof can be found in literature, Hyvärinen et al., 2001*b*.

For the multivariate case $\mathbf{y} = \mathbf{A}\mathbf{x}$, the differential entropy $H(f(\mathbf{y}))$ is given by

$$H(f(\mathbf{y})) = H(f(\mathbf{x})) + \log|\det \mathbf{A}|. \tag{A.11}$$

Refer to Hyvärinen et al., 2001*b*, for the proof.

Furthermore, according to Hyvärinen et al., 2001*b*, the differential entropy of the transformation $\mathbf{y} = g(\mathbf{x})$, where g($\mathbf{x}$) is a monotone increasing function for which the inverse mapping $\mathbf{x} = g^{-1}(\mathbf{y})$ exists and is unique, is given by

$$H(f(g(\mathbf{x}))) = H(f(\mathbf{x})) + E\{\log|\mathbf{J}g(\mathbf{x})|\}, \tag{A.12}$$

where $\mathbf{J}g(\mathbf{x}) = \mathbf{J}g(g^{-1}(\mathbf{y}))$ is the Jacobian matrix given by

$$\mathbf{J}g(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g_N(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial g_N(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x})}{\partial x_N} & \frac{\partial g_2(\mathbf{x})}{\partial x_N} & \dots & \frac{\partial g_N(\mathbf{x})}{\partial x_N} \end{bmatrix}. \tag{A.13}$$

A.3 Negentropy

Theorem A.5 *The negentropy $J(f(x))$ of the random variable X with the density function $f(x)$ is always positive*

$$J(f(x)) \geq 0. \tag{A.14}$$

Proof: [Comon, 1994]
The negentropy is defined by

$$J(f(x)) = H(\phi(x)) - H(f(x)), \tag{A.15}$$

using the entropy defined as $H(f(x)) = -\int f(x) \log f(x)\, \mathrm{d}x$, the negentropy can be written as

$$J(f(x)) = -\int \phi(x) \log \phi(x) \mathrm{d}x + \int f(x) \log f(x) \mathrm{d}x. \tag{A.16}$$

Adding and subtracting a term in the definition of negentropy gives

$$\begin{aligned} J(f(x)) &= \int f(x) \log f(x) \mathrm{d}x - \int f(x) \log \phi(x) \mathrm{d}x \\ &+ \int f(x) \log \phi(x) \mathrm{d}x - \int \phi(x) \log \phi(x) \mathrm{d}x, \end{aligned} \tag{A.17}$$

which can be written as

$$J(f(x)) = \int f(x) \log \frac{f(x)}{\phi(x)} \mathrm{d}x + \int (f(x) - \phi(x)) \log \phi(x) \mathrm{d}x. \tag{A.18}$$

Now, by assuming that $\phi(x)$ and $f(x)$ have the same first- and second-order moments and since

$$\log(\phi(x)) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2},$$

and

$$\begin{aligned} \int f(x) \log \phi(x) \mathrm{d}x &= \int -\frac{1}{2} \log(2\pi\sigma^2) f(x) \mathrm{d}x - \int \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} f(x) \mathrm{d}x \\ &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2} \qquad \forall\, f \text{ or } \phi, \end{aligned}$$

it follows that

$$\int \phi(x) \log \phi(x) \mathrm{d}x = \int f(x) \log \phi(x) \mathrm{d}x, \tag{A.19}$$

and the negentropy of (A.18) may be written as an information-function:

$$J(f(x)) = \int f(x) \log \frac{f(x)}{\phi(x)} \mathrm{d}x. \tag{A.20}$$

This proves, referring to (A.1), that the negentropy as well the information is positive

$$J(f(x)) \geq 0, \tag{A.21}$$

with equality if and only if $\phi(x) \equiv f(x)$. ■

Additionally, the negentropy is invariant for invertible linear transformations.

Theorem A.6 *The negentropy $J(f(\mathbf{x}))$ of a multivariate random vector $\mathbf{x} = (x_1, \ldots, x_N)$ is invariant for linear transformations.*

Proof: Consider the invertible linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{A.22}$$

where $\mathbf{y}$ is an N-dimensional random vector. The covariance of $\mathbf{y}$ is given by

$$E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{A}\mathbf{C}\mathbf{A}^T, \tag{A.23}$$

where $\mathbf{C}$ is the covariance matrix of $\mathbf{x}$. Using the result of Example 2.5, where the differential entropy of a multivariate gaussian distribution can also be written as

$$H(\phi(\mathbf{x})) = \frac{1}{2} \log |\det \mathbf{C}| + \frac{N}{2}[1 + \log(2\pi)], \tag{A.24}$$

and using the result of Equation A.11, where the differential entropy of a transformation is given as $H(f(\mathbf{y})) = H(f(\mathbf{x})) + \log |\det \mathbf{A}|$, the negentropy of the invertible linear transformation can be computed as

$$\begin{aligned} J(f(\mathbf{A}\mathbf{x})) &= \frac{1}{2} \log |\det(\mathbf{A}\mathbf{C}\mathbf{A}^T)| + \frac{N}{2}[1 + \log(2\pi)] - (H(f(\mathbf{x})) + \log |\det \mathbf{A}|) \\ &= \frac{1}{2} \log |\det \mathbf{C}| + 2\frac{1}{2} \log |\det \mathbf{A}| + \frac{N}{2}[1 + \log(2\pi)] - H(f(\mathbf{x})) - \log |\det \mathbf{A}| \\ &= \frac{1}{2} \log |\det \mathbf{C}| + \frac{N}{2}[1 + \log(2\pi)] - H(f(\mathbf{x})) \\ &= H(\phi(\mathbf{x})) - H(f(\mathbf{x})). \end{aligned} \tag{A.25}$$

■

In particular negentropy is scale-invariant, i.e. multiplication of a random variable by a constant does not change its negentropy. This was not true for differential entropy [Hyvärinen et al., 2001*b*, p. 113], as shown in Theorem A.4.

A.4 Approximation of Information-Theoretic Functions

Using the entropy or negentropy in practice would be computationally difficult, because the integral of these functions involves the probability density function, which is often not known. Therefore the entropy or negentropy are more theoretical functions. In practice some estimates are used. One classical method of approximating these functions is based on higher-order cumulates such as kurtosis. This idea is based on using an expansion like a Taylor expansion [Hyvärinen, 2001]. This expansion is taken for the probability density function of a continuous random variable assuming that the density function $f_X(\xi)$ is near the standard gaussian density where X has zero-mean and unit variance

$$\varphi(\xi) = \exp(-\xi^2/2)\sqrt{2\pi}. \tag{A.26}$$

Usually the Gram-Charlier expansion is used which is a special series expansion where the target function is approximately composed from Hermite polynomials H_0, H_1, ... as basis functions. The polynomials are defined by the derivatives of the standardized gaussian density as

$$\frac{\partial^i \varphi(\xi)}{\partial \xi^i} = (-1)^i H_i(\xi)\varphi(\xi). \tag{A.27}$$

H_i is a polynomial of order i, to become more familiar with the Hermite polynomials, they are computed as,

$$H_i(\xi) = \xi^i - \frac{i^{[2]}}{2 \cdot 1!}\xi^{i-2} + \frac{i^{[4]}}{2^2 \cdot 2!}\xi^{i-4} - \frac{i^{[6]}}{2^3 \cdot 3!}\xi^{i-6} + \ldots, \tag{A.28}$$

where

$$i^{[a]} = i(i-1)(i-2)\cdots(i-(a-1)) = \frac{i!}{(i-a)!}. \tag{A.29}$$

Following this, the first Hermite polynomials are given by [Stuart and Ord, 1994]:

$$\begin{aligned} H_0(\xi) &= 1 \\ H_1(\xi) &= \xi \\ H_2(\xi) &= \xi^2 - 1 \\ H_3(\xi) &= \xi^3 - 3 \\ H_4(\xi) &= \xi^4 - 6\xi^2 + 3. \end{aligned}$$

These polynomials have the characteristic, that they form an orthogonal system

$$\int \varphi(\xi) H_i(\xi) H_j(\xi) d\xi = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases} . \tag{A.30}$$

The Gram-Charlier expansion of the density of X including the two first nonconstant terms, is given by [Stuart and Ord, 1994]

$$f_X(\xi) \approx \hat{f}_X(\xi) = \varphi(\xi) \left(1 + \frac{\kappa_3}{3!} H_3(\xi) + \frac{\kappa_4}{4!} H_4(\xi)\right). \tag{A.31}$$

As said, the expansion is based on the idea that the density function of X is close to a gaussian one, which allows a Taylor-like approximation and the nongaussian part is directly given by the higher-order cumulates (the third- and fourth-order cumulant, κ_3 and κ_4 named skewness and kurtosis). Using this density approximation the differential entropy can be approximated as

$$H(f(x)) \approx -\int \hat{f}_X(\xi) \log \hat{f}_X(\xi) d\xi. \tag{A.32}$$

Using again the idea that the density is close to a gaussian one, the cumulants in (A.31) are very small and their logarithms can be approximated through

$$\log(1 + \varepsilon) \approx \varepsilon - \varepsilon^2/2, \tag{A.33}$$

which gives then

$$\begin{aligned} H(f(x)) \quad \approx \quad & -\int \varphi(\xi) \left(1 + \frac{\kappa_3}{3!} H_3(\xi) + \frac{\kappa_4}{4!} H_4(\xi)\right) \cdot \\ & \left[\log \varphi(\xi) + \frac{\kappa_3}{3!} H_3(\xi) + \frac{\kappa_4}{4!} H_4(\xi) - \left(\frac{\kappa_3}{3!} H_3(\xi) + \frac{\kappa_4}{4!} H_4(\xi)\right)^2 /2\right] d\xi. \end{aligned} \tag{A.34}$$

This expression can be simplified to

$$H(f(x)) \approx -\int \varphi(\xi) \log \varphi(\xi) d\xi - \frac{\kappa_3^2}{2 \cdot 3!} - \frac{\kappa_4^2}{2 \cdot 4!}. \tag{A.35}$$

Consequently, the approximated negentropy of a standardized random variable is given by

$$J(f(x)) \approx -\frac{1}{12} E\{X^3\}^2 - \frac{1}{48} \text{kurt}(X)^2, \tag{A.36}$$

a computationally simple approximation of the measure of nongaussianity of a random variable.

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit mit dem Thema:

"Analysis of Functional Magnetic Resonance Imaging Time Series by Independent Component Analysis"

ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Insbesondere habe ich nicht die Hilfe einer kommerziellen Promotionsberatung in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation, Diplom- oder ähnliche Prüfungsarbeit eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, 26.02.2007

Danksagung

Mein Dank gilt an dieser Stelle besonders Frau Prof. Dr. Waltraud Kahle und Herrn PD Dr. Siegfried Kropf, die diese Arbeit betreut haben. Für ihre Anregungen, Hinweise und Ideen bei der Erstellung dieser Arbeit, möchte ich ihnen herzlich danken.

Des Weiteren möchte ich herzlich dem Leibniz-Institut für Neurobiologie (Direktor: Prof. Dr. Henning Scheich) und besonders dem Speziallabor "Nicht-Invasive Bildgebung" (Laborleiter: Dr. André Brechmann) sowie allen Kollegen und Kolleginnen für ihre Unterstützung danken. Besonderer Dank gilt Frau Dr. Birgit Gaschler-Markefski, für die Ermutigung zur Promotion und ihre unzähligen hilfreichen Hinweise bei der Erstellung dieser Arbeit.

Ausserdem gilt mein Dank Prof. Dr. Georg Reiser, der es mir ermöglichte durch das Graduiertenkolleg "Neurobiologische Grundlagen von Erkrankungen des Nervensystems" umfassende Kenntnisse auf dem Gebiet der Neurobiologie zu erwerben.

Natürlich möchte ich auch meiner Familie und meinen Freunden danken, die mich die ganze Zeit bei meinem Vorhaben unterstützt und ermutigt haben.

Lebenslauf

Name:	Mandy Sohr
Anschrift:	Grundstr. 40 55218 Ingelheim
Geburtsdatum:	20. Februar 1980
Geburtsort:	Lutherstadt Wittenberg
Staatsangehörigkeit:	deutsch

Schulausbildung

09/1986 - 08/1991	Polytechnische Oberschule, Lutherstadt Wittenberg
09/1991 - 08/1992	Sekundarschule, Lutherstadt Wittenberg
09/1992 - 08/1998	Lucas-Cranach-Gymnasium, Lutherstadt Wittenberg

Hochschulausbildung

10/1998 - 09/2002	Hochschule Magdeburg-Stendal Studiengang Statistik
10/2002 - 08/2006	Leibniz-Institut für Neurobiologie in Magdeburg, Doktorandin im Speziallabor "Nicht-Invasive Bildgebung" in Zusammenarbeit mit dem Institut für Mathematische Stochastik der Otto-von-Guericke-Universität Magdeburg

Stipendien

10/2002 - 09/2005	Stipendiatin im Graduiertenkolleg "Neurobiologische Grundlagen von Erkrankungen des Nervensystems" der Otto-von-Guericke-Universität Magdeburg gefördert durch die Deutsche Forschungsgemeinschaft

Berufliche Tätigkeit

seit 09/2006	Biostatistikerin bei Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein

Publikationen

Sohr, M., Gaschler-Markefski, B., and Kahle, W. (2004) Neural Success-Related Correlates of Human Auditory Working Memory: An fMRI Study Analyzed with Independent Component Analysis. *Proceedings of Longevity, Aging and Degradation Models in Reliabibility, Public Health, Medicine and Biology*, p. 274-283, St. Petersburg, Russia, 2004

Sohr, M., Brechmann, A., and Kahle, W. (2006) Using Independent Component Analysis of fMRI Time Series to Investigate Task Related Activation. *Proceedings of the International Conference: Statistical Methods for Biomedical and Technical Systems*, p. 455-460, Limassol, Cyprus 2006

Brechmann, A., Gaschler-Markefski, B., Sohr, M., Yoneda, K., Kaulisch, T., Scheich, H. (2007) Working memory specific activity in auditory cortex: Potential correlates of sequential processing and maintenance *Cerebral Cortex*, published online 04.01.2007

Präsentationen

Sohr, M., Gaschler-Markefski, B., and Scheich, H. (2003) Success-related effects of an auditory working memory task on single-subject fMRI activation of auditory cortex. *International Conference on Auditory Cortex 2003 - Towards a Synthesis of Human and Animal Research* Magdeburg 2003 (Poster)

Sohr, M. (2004) Independent Component Analysis: Eine Methode zur Analyse von funktionellen Magnetresonanztomographie-Zeitreihen. *Pfingsttagung der Deutschen Statistischen Gesellschaft*, Leipzig, 03.-04. Juni 2004 (Vortrag)

Sohr, M. (2005) Analyzing biomedical time-series. *Pfingsttagung der Deutschen Statistischen Gesellschaft*, Münster,19.-20. Mai 2005 (Vortrag)

Sohr, M., Brechmann, A., Gaschler-Markefski, B. and Scheich, H. (2005) Working Memory Effects in Human Auditory Cortex: fMRI amplitude predicts task performance. *11th Magdeburg International Neurobiological Symposium: "Learning and Memory"* May 28 - June 01, 2005 (Poster)

Sohr, M., Brechmann, A., and Kahle, W. (2006) Using Independent Component Analysis of fMRI Time Series to Investigate Task Related Activation. *International Conference: Statistical Methods for Biomedical and Technical Systems*, Limassol, Cyprus 2006 (Vortrag)